On the Relationship between Transparency, Explainability and Trust in AI systems: a Conceptual Analysis

This paper challenges the idea that transparency and explainability build trust in AI systems. We survey conflicting empirical evidence on the topic and then clarify the main concepts involved in the argument. Based on this conceptual clarification, we argue that transparency and explainability do not convey a complete understanding of how an AI system works, and are not relevant factors for building trust in AI systems. Accordingly, when the objective is to create trust in AI systems, transparency and explainability are neither necessary nor sufficient; therefore it is not rational to pursue them for this reason alone. We conclude that, while the results of Explainable Artificial Intelligence (XAI) may be useful for other reasons, it is both necessary and possible to build trust in AI systems through alternative approaches such as rigorous validation and sound institutional arrangements and practices.

Keywords: *transparency, explainability, explainable AI, trust, artificial intelligence*

Author Information

Alessio Tartaro, University of Sassari, Department of Humanities and Social Sciences, Italy.
https://orcid.org/0000-0002-0382-3083

Mihály Héder, Budapest University of Technology and Economics, Dept. for Philosophy and History of Science / HUN-REN SZTAKI, Hungary.
https://orcid.org/0000-0002-9979-9101

How to cite this article:

Tartaro, Alessio, Mihály Héder. "On the Relationship between Transparency, Explainability and Trust in AI systems: a Conceptual Analysis".
Információs Társadalom XXV, no. 4 (2025): 25–43.
https://dx.doi.org/10.22503/inftars.XXV.2025.4.2

1. Introduction

In artificial intelligence, a black box is an AI system whose internal workings are not visible or understandable to the user. The user is only able to provide input and receive output, without any insight into how the model arrived at its output (e.g., a decision or a prediction). The "black box" nature of these systems makes them inherently *opaque*. The "black box problem" in AI refers to the challenges and issues that arise as a consequence of the opacity of AI systems. A central argument is that opacity negatively affects trust in AI systems. According to Zednik (2021, 266), "end users are less likely to trust and cede control to machines whose workings they do not understand". Similarly, Haque et al. (2023, 1) argue that "the opacity of AI systems can reduce end users' trust and reliance on using AI-based systems while making critical decisions".

A growing number of scholars and policymakers are increasingly calling for transparency and explainability to solve the "black box problem" and build trust in AI systems. For example, Ribeiro et al. (2016, 1135) state that "explaining predictions is an important aspect in getting humans to trust and use machine learning". Similarly, Guidotti et al. (2019, 2) write that "the availability of transparent machine-learning technologies would lead to a gain of trust". According to Floridi et al. (2018, 701), "it is especially important that AI be explicable, as explicability is a critical tool to build public trust in, and understanding of, the technology". This position is echoed in the Ethics Guidelines for Trustworthy AI (AI HLEG 2019, 3), that considers transparency and explainability as "crucial for building and maintaining users' trust in AI systems". This emphasis on transparency was also carried over to the EU AI Act.

Researchers in the field of Explainable Artificial Intelligence (XAI) aim to address the challenges of the "black box problem" by developing techniques that can make the internal workings of AI systems transparent and explainable (Barredo Arrieta et al. 2020). In a comprehensive literature review, Haque et al. (2023) list trust as one of the main effects of explainability in artificial intelligence. Langer et al. (2021) consider trust as one of the desiderata of users and deployers of AI systems, and refer to explainability as a way to achieve this desideratum. Accordingly, there is a general consensus that addressing the "black box" issue in AI systems, primarily through enhanced transparency and explainability, is likely to bolster trust. This represents the main thesis discussed in this paper: *transparency and explainability build trust in AI systems*. For brevity, we will refer to this thesis with the "transparency-trust thesis".

Transparency and explainability in AI systems can encompass various facets, including the model, algorithm, data, and broader aspects of development and usage (Andrada et al. 2023). The focus shifts depending on the subject matter. For instance, transparency regarding data involves details about data collection, provenance, annotation processes, and composition (Bertino et al. 2019). Our research narrows this broad spectrum to cases where transparency or explainability elucidates the rationale behind an AI system's specific outputs in response to given inputs. This functional perspective remains neutral to whether the model, data, or both are the

subjects of transparency or explainability. To address the question of "why an AI system produces certain outputs given certain inputs," the need for explaining or making transparent the model, data, or both can vary case by case.

While transparency and explainability are primarily associated with building trust in AI systems, they also serve additional purposes. For instance, transparency in training data—detailing the origin and composition of datasets—offers insights into data quality and the potential for bias. Moreover, it enables checks on whether the system draws inferences from relevant and representative data, assists in identifying and rectifying bugs, and helps guard against malicious or adversarial data injections (Koene et al. 2019). In this paper, we do not question these varied applications of transparency and explainability. Instead, our focus is narrowly on their instrumental role in fostering trust in AI systems.

Finally, proponents of the transparency-trust thesis aim to foster trust from the perspective of the end user. A regulator or a certifying body might need transparency to perform validation that requires such details about the system that the end-user would not be able to leverage. Therefore, our paper also focuses on trust in the eye of the end-user.

The "transparency-trust thesis" is widely endorsed in the debate around the ethical and social implications of Artificial Intelligence, and it constitutes the rationale supporting significant research efforts, policy initiatives, investments, and funding. However, although many studies empirically investigate the effects of transparency and explainability on trust in particular settings (see section 2), none of them analyse the conceptual consistency of the "transparency-trust thesis". More often than not, some influence of opacity, transparency, and explainability on trust is just taken for granted, and empirical evidence is offered to corroborate this thesis. Nevertheless, the constituent concepts of this thesis are not always clearly defined and adequately discussed, and their mutual interdependencies are not sufficiently elaborated, which makes it hard to conduct meaningful research and make informed decisions about AI. There is a need for more discussion on these concepts in relation to AI, as if the "transparency-trust thesis" is false, this would require rethinking the way trust in AI systems works and how we can build it.

This paper provides a conceptual clarification of the concepts of opacity, transparency, explainability, and trust, in order to show various inconsistencies in the "transparency-trust thesis". We further explore the concept of "understandability," which is central to the discussion on opacity, transparency, and explainability of AI systems. We argue that when the objective is to create trust in AI systems, transparency and explainability are neither necessary nor sufficient and therefore it is not rational to pursue them, or, at least the pursuit needs other justification, perhaps a more epistemic goal. To support this thesis, we proceed as follows. In section two, we present related work on the effect of transparency and explainability on trust. In section three, we elaborate clear definitions of the concepts under investigation. Based on this, in section four, we put forward arguments against the "transparency-trust thesis". In the conclusion, we summarise our findings and outline future lines of research.

2. Related work

A number of empirical studies on the effects of transparency and explainability on trust provide conflicting evidence on the validity of the "transparency-trust thesis".

Based on a review of the literature, Shin (2021) formulates a series of hypotheses on the effects of explainability and causability on perception, trust, and acceptance of AI-based news recommendation systems. Two of these hypotheses, i.e., "explainability positively influences user perception of AI transparency" and "perceived transparency positively influences the user trust in AI", are supported by empirical evidence. Through a series of tests and surveys involving 350 individuals experienced with algorithmic news services, the author claims to validate these hypotheses and concludes favourably in support of the "transparency-trust thesis".

Kartikeya (2022) finds that a high level of transparency contributes to an increase in user trust. In the study, respondents were asked to predict the star rating of a restaurant based on the text of a review and the output of a machine learning model. By varying the amount of information given by the model to the users, the study finds that with increased transparency, i.e., more information, trust also increases as measured by the model's influence on the respondents. Additionally, the study found that any additional insight into the model's decision making will increase trust, regardless of whether the model is correct or not.

Similar results are reported in the medical sector by Liu et al. (2022). The study finds that explainability contributes to increasing trust in medical AI. In addition, transparency measures, e.g., providing information about the source of training data, the algorithm used, and the quality of the model, can help to promote trust and recognition of AI's value among physicians. Finally, the authors also find that trust is a key factor influencing physicians' intentions to use AI. Likewise, explanations work as a trust mechanism in healthcare according to Wysocki et al. (2023). These results support the "transparency-trust thesis".

This supporting empirical evidence is countered by an equal number of studies showing a neutral, or even negative, effect of transparency and explainability on trust. Papenmeier et al. (2019) investigate the effect of explanations on user trust in a machine learning-based text classifier, considering factors such as overall accuracy of the system, the fidelity level of the explanation, and the user's level of consciousness. They find that the accuracy of the system is the most decisive factor for fostering user trust, with higher accuracy leading to higher trust. It should be noted here that accuracy of a system is orthogonal to the transparency of a system, i.e., it refers to outputs that we can measure for black-box and non – black-box systems the same way. The study also finds that the influence of explanation fidelity on user trust is complex and varies depending on the accuracy of the system: for systems with medium accuracy, a high-fidelity explanation does not harm user trust, while a low-fidelity explanation does. In addition, explanation leads to a decrease in trust for systems with high accuracy.

Kizilcec (2016) tested the effect of transparency on user trust in the context of peer assessment in an online course. The study shows that transparency has a variable

effect on trust depending on the initial expectations of users. While the violation of expectations decreases trust, providing some transparency with procedural information helps to rebuild trust, but this effect is nullified when too much information is provided about the system, leading to a decrease in trust.

Schmidt et al. (2020) studied how transparency affects trust when users have to decide whether to accept the results of an AI system in predictions, classifications, and recommendations tasks. The study finds that higher levels of transparency may not necessarily imply higher levels of trust or acceptance when it comes to dealing with AI's output. On the contrary, unintuitive explanations, although faithful, can lead to mistrust. In addition, the study also finds that overly trusting wrong predictions can occur, particularly when the task is difficult.

Finally, Ghassemi et al. (2021, 475) argue that "the desire to engender trust through current explainability approaches represents a false hope". Focusing on the healthcare sector, they critically examine the use and limitations of explainability techniques. The study highlights that explanations provided by AI systems, such as heat maps (also known as saliency maps) in image analysis, are often superficial. These heat maps indicate which areas of an image the AI system considers when making a decision but fail to address the crucial question of the appropriateness of such focus. This conflation can lead users to mistakenly believe these explanations are comprehensive reflections of the AI's decision-making process, potentially creating over-trust and automation bias. Users might perceive the AI as more transparent and understandable, despite the superficial nature of the explanations, leading to an unwarranted confidence in the system's capabilities. Thus, the authors argue that simply providing explanations does not suffice to build genuine trust. They advocate for more rigorous and thorough validation procedures as an alternative to the "transparency-trust thesis", underscoring the need for a more critical approach to AI explainability.

This review of the literature shows that the "transparency-trust thesis" is much more controversial than initially alleged. Beneath the commonly held position according to which "transparency and explainability build trust", there is a reality where the relationship between these elements is far from clear. The effect of transparency and explainability on trust is variable, can be positive, neutral, and negative, and depends on numerous factors such as, among others, the task, the accuracy of the AI system, the context of use, the type of explanation, the amount of information provided, what is made transparent (the model, training data etc.), the type of user, the expectations of users, and the fidelity of explanations. Even experiments conducted within the same sectors, e.g., healthcare, provide conflicting empirical evidence. Such a variable impact of transparency and explainability on trust has been observed in human-AI teams in the field of aviation as well (Lopez et al. 2024).

This shows that these studies investigating the relationship between transparency, explainability and trust, lack ecological validity and therefore cannot be generalised (Zerilli et al. 2022). Finally, as concluded in a recent review of the empirical literature on explainable AI (Kandul et al. 2023, 17), "much of this research does not live up to the rigorous standards of empirical research" and many of the results report inconsistent and contradictory findings on the effect of explainability on trust.

3. Defining key concepts

Since the validity of the "transparency-trust thesis" cannot be determined strictly empirically, we consider it appropriate to shift the analysis to a conceptual level. In the rest of the paper, we aim to advance the discussion on the topic by clarifying the key concepts involved in the "transparency-trust thesis" and examining their complex interrelationships. These concepts are: opacity, transparency, explainability, understandability and trust.

3.1. Opacity

Opacity is associated with a lack of understanding of how an AI system works. Researchers have identified several forms and causes of opacity.

Burrell (2016) distinguishes between three forms of opacity: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully. We refer to these notions as Opacity-1, Opacity-2, and Opacity-3. All these types of opacity contribute to a lack of understanding. Among these, however, opacity-3 has a special status. While opacity-1 stems from intentional corporate secrecy to defend intellectual property and trade secrets and opacity-2 depends on the lack of users' coding skills and competences, opacity-3 emerges from intrinsic characteristics of an AI system. According to the author, the scale and complexity of machine learning algorithms are such that they are opaque even to the experts who develop them. This depends on several factors, e.g., the learning abilities, the quantity and high number of dimensions/features of the data, and the computational resources needed by the AI systems, as also argued in related work (Facchini and Termine 2022)

Grünke (2019) further elaborates on the notion of opacity-3, which they call "epistemic opacity," through a comparative analysis of the Stockfish (rule-based) and AlphaZero (neural network-based) chess engines. According to the author, the two chess engines exhibit different types of epistemic opacity. Stockfish is epistemically opaque due to the sheer amount of calculations that it does. The number of positions that the engine calculates each second far exceeds human capabilities (up to 60 million positions per second). Consequently, it is impossible for a human to reconstruct and understand step-by-step the process Stockfish follows to make a move. Since this opacity depends on the cognitive capacities and limitations of a human agents, the author calls it "contingent epistemic opacity". On top of this, AlphaZero exhibits an additional form of opacity. Not only does AlphaZero calculate millions of positions per second, but it also represents features of the game in a way difficult to understand for humans because we do not have equivalent human concepts for some of the features represented in the neural network. This form of opacity, called "fundamental epistemic opacity," concerns the way in which the neural network models the characteristics of the game. Humans do not understand why AlphaZero makes certain moves because humans and the AI system develop and deploy two different representations of the game.

The distinction between two dimensions of opacity is also supported by Boge (2022). According to the author, deep neural networks (DNNs) are characterised by h-opacity and w-opacity. The former refers to the lack of understanding due to an agent's cognitive limitations and is thus similar to contingent epistemic opacity. The second refers to what is learned by the DNN, i.e., how it models a phenomenon, and thus is similar to fundamental epistemic opacity.

Finally, Héder (2023) identifies additional sources of epistemic opacity. According to the author, AI systems are partially opaque because of the physical complexity of computers. The argument is that the level of investigation of AI systems is their logical model, which, contrary to the commonly held view, is not perfectly embodied. He proposes that any AI that we may encounter is to be investigated as a cyber-physical system in a physical environment. For instance, the thermodynamic processes in computers are fed back into the system as input for randomness, therefore making the comprehension of these complexities necessary for the understanding of the behaviour of the machine. The author points out that, among other factors, self-modification by responding to the environment (via machine learning) plays a pivotal role in raising epistemic opacity. Since most AI systems are characterised by both complexity and self-modification, his arguments apply to most machine learning-based AI systems.

In conclusion, these analyses of the concept of opacity show that it depends both on the purposes and characteristics of its users, and on intrinsic properties of AI systems. The different sources and types of opacity are considered hindering factors in the user's ability to understand a system. Among these factors, the scale and complexity of AI systems play a pivotal role in generating epistemic opacity, as acknowledged in all the analyses considered.

3.2. Transparency and explainability

In the current debate, transparency and explainability often overlap, along with the related concept of interpretability. In this section, we adopt the taxonomy proposed by Lipton (2018) and (Barredo Arrieta et al. 2020) in order to clarify the meaning of these concepts.

According to Lipton (2018), transparency and explainability are aspects of interpretability. Transparency is understood as the opposite of opacity and so "it connotes some sense of understanding the mechanism by which the model works" (Lipton 2018, 12). Transparent models are usually considered understandable by design. These models can be transparent at different levels: at the level of the entire model (simulatability), at the level of individual components (decomposability), and at the level of the training algorithm (algorithmic transparency) (Barredo Arrieta et al. 2020; Lipton 2018). A transparent simulatable model is such that a human should be able, given the input data, to produce the model's output in a reasonable time through the same calculations as performed by the model. A transparent decomposable model is a model whose parts (inputs, parameters, calculations) are such that they can be interpreted and explained. Algorithmic transparency refers to the

learning algorithm rather than the model itself. Algorithmic transparency only requires knowledge of the algorithm and not of the data or learned model, since it just concerns how the algorithm creates the model. According to Barredo Arrieta et al. (2020, 88), "a model is considered to be transparent if by itself it is understandable". Based on this definition, Barredo Arrieta et al. (2020) and Lipton identify a number of models, e.g., linear/logistic regression, decision trees, Bayesian models etc., that are transparent in at least one of the above senses, i.e., simulatability, decomposability, algorithmic transparency, collectively referred to as "transparent models".

As far as the concept of explainability is concerned, Lipton (2018) understands it as a form of post-hoc interpretability. This approach involves generating explanations for the outputs of AI systems after they have been produced. Unlike built-in interpretability, where the decision-making process is transparent from the outset, post-hoc interpretability seeks to retrospectively elucidate how a system arrived at its conclusions. It aims to demystify complex AI models by examining their outputs and deducing the contributing factors, thereby providing insights into the system's operational logic. This approach appears to be equivalent to "local explainability" from the field of explainable AI (Héder 2023). Similarly, according to Barredo Arrieta et al. (2020, 92), explainability techniques "aim at communicating understandable information about how an already developed model produces its predictions for any given input". Explainability techniques are used when the model is not transparent in any of the senses above. In this case, understanding cannot be achieved by directly examining the model because the model itself is opaque enough to prevent any form of understanding. To overcome this problem, an understanding of the model is sought out through an explanation external to the model that provides additional information about it. Explanations can be provided, for example, in the form of textual explanations, visualisations, saliency maps, and examples (Lipton 2018). Within the post-hoc explainability techniques, the primary distinction is between model-agnostic and model-specific techniques. The former can provide explanations regardless of the model considered, while the latter can be applied only to particular kinds of models, e.g., deep neural networks.

From this brief discussion, we agree with (Barredo Arrieta et al. 2020, 88) that "*understandability emerges as the most essential concept in XAI*". Consequently, the concept of understandability is also crucial for clarifying the "transparency-trust thesis". For this reason, instead of dwelling on more in-depth analysis on the concepts of transparency and explainability, we prefer to go straight to the heart of the matter and confront the concept of understandability. In the next section, we try to shed light on this concept in order to clarify what is meant by "understanding an AI system".

3.3. Understandability

The conceptual analysis undertaken so far indicates that the challenge in explaining or making AI systems transparent primarily revolves around the issue of understanding such systems. Consequently, this shifts our focus from transparency and

explainability to understandability, prompting an investigation into the meaning of "understanding an AI system".

Addressing this question, researchers in the field of Explainable AI often draw upon insights from epistemology. For instance, Paez (2019) identifies two distinct categories of understanding applicable to AI: objectual understanding and understanding-why. Objectual understanding involves grasping the relationships within the system, akin to understanding the parts and their interconnections within a whole. This form of understanding parallels that provided by transparent-by-design models (see section 3.2). On the other hand, understanding-why goes beyond this, as it encompasses the ability to engage with counterfactual scenarios and predictions. This level of understanding aligns with the goals of post-hoc explainability techniques.

Additionally, Paez (2019) proposes an alternative dichotomy: mechanical understanding versus functional understanding. Functional understanding focuses on the purposes and functionalities of a system. In contrast, mechanical understanding delves into the specific components, processes, and immediate causal mechanisms within that system. Paez (2019) exemplifies this through the understanding of an alarm clock. One can understand the alarm's function either through the lens of mechanical understanding – a completed circuit activating the buzzer – or through functional understanding – the clock set to awaken its owner at a designated time.

These different kinds of understanding AI of systems do not resolve the question of the desired depth of understanding we want in the context of AI systems. Paez (2019) posits that within the realm of AI, both objectual understanding and understanding-why are intimately connected, and so we need both. He further argues that mere functional understanding is insufficient, advocating for an integration of mechanical understanding as well.

While we agree with Paez (2019) on this point, it is important to note his omission regarding the specific characteristics that make a system understandable. Asserting that understanding an AI system involves engaging with counterfactual scenarios and predictions raises the critical question of our capability to do so and under which circumstances. This consideration is paramount, as it pertains to our fundamental ability to understand certain AI systems.

We propose that the following characteristics contribute to the understandability of AI systems: linearity, monotonicity, few (and simple) interactions among features, rule-based nature (Molnar 2022). By identifying these characteristics, we can better understand the factors that make AI systems more understandable to humans.

A model is linear if it describes the connection between its inputs and outcomes using a linear, or straight-line, function. This linear relationship is easier to comprehend than a non-linear one, as it represents a simpler kind of connection. Similarly, a model is monotonic if any increase or decrease in an input consistently leads to either an increase or a decrease in the output, depending on the nature of the function. These types of relationships are more straightforward to understand compared to those where changes in inputs and outputs do not follow a regular pattern.

It is also beneficial for a model to have a limited number of simple features that interact in straightforward ways. This simplicity makes it easier to track how the

model is working and what it is trying to represent. Additionally, if the model uses simple inputs and is based on clear rules, like IF-THEN statements, it becomes more understandable.

These characteristics are emphasised for AI systems due to their role in enhancing the system's comprehensibility. However, it's important to note that as the size and complexity of these AI systems expand, their understandability can diminish, even with linear or monotonic models. Similarly, decision trees start off as straightforward but can lose clarity as they become more intricate. The same applies to rule-based systems.

The main point here is that simple AI systems are easier to understand. This is in line with the observation that the difficulty in understanding AI systems often stems from their scale and complexity. Thus, a simpler AI system is more likely to be understandable, allowing for deeper insights into its functioning and decision-making processes, unlike complex systems with numerous parameters and intricate interactions.

This conclusion is supported by an additional remark. In section 3.1., we found that the opacity of AI systems is largely due to their complexity. Consequently, it makes sense to argue that a non-opaque, i.e. understandable, AI system is a simple system. In the case of simple AI systems, we can have meaningful insights about its internal states, comprehend specific outcomes, and make accurate inferences and predictions about the system's behaviour. This is not the case for complex, large, and opaque AI systems, which have millions or billions of parameters and non-linear interactions, making it challenging to grasp their inner workings and decision-making processes.

3.4. Trust

The "transparency-trust thesis" implies that transparency and explainability have a positive effect on trust as they counterbalance the negative effect of opacity on trust. However, what is meant by trust in the context of AI is difficult to determine accurately because research in this field uses different definitions of trust. In this section, we provide our understanding of "trust in AI systems" in the context of the "transparency-trust thesis".

Much of the debate on this topic revolves around the question of whether it makes any sense at all to talk about "trust in AI systems". Freiman (2023) offers a clear analysis of why this expression is conceptual nonsense according to numerous philosophical accounts of trust. In a nutshell, talking about trustworthy AI systems does not make sense because trustworthiness entails human qualities, such as responsibility, morality, intentionality. Since AI systems lack these human qualities, humans cannot meaningfully trust an AI system. Accordingly, it makes no sense to talk about a "trustworthy AI". At most, one should talk about "reliable AI" (Freiman 2022, 6). Similarly, Alvarado (2023) argues for a reduced scope of trust in AI. This argument, however, is refuted by other authors, for example Floridi and Sanders (2004), who are comfortable talking about the mind-less morality of artificial agents and therefore find it consistent to talk about "trust in AI systems".

Both arguments hinge on the assumption that a dependable connection exists between the potential for trustworthy AI systems and the presence of human attributes, like morality and agency, as a prerequisite for these systems to be considered trustworthy. Freiman posits that, due to the lack of such human characteristics in AI systems, discussing the concept of Trustworthy AI is nonsensical. Conversely, Floridi and Sanders argue that, since AI systems do have moral traits, this paves the way for the possibility of them being regarded as trustworthy.

Although it is impossible to completely avoid philosophical assumptions on such subjects, we aim to develop an account of "trust in AI systems" that is less dependent on these assumptions. Specifically, we seek to establish an understanding of trust that does not presuppose any direct relationship between the capacity to place trust and the presence of human qualities in the entity being trusted. In light of this perspective, humans can trust AI systems even in the absence of human-like qualities, and AI systems can be considered "trustworthy" without possessing attributes such as morality, responsibility, or other human-like traits. Consequently, we move away from the positions exemplified respectively by Freiman (2023) and Floridi and Sanders (2004), and develop an account of "trust in AI systems" rooted in studies on "trust in technology" by McKnight et al. (2009).

According to the authors, three meanings of "trust in technology" can be distinguished: trust in specific technology, propensity to trust general technology, and institution-based trust in technology.

Trust in a specific technology is defined as "a willingness to depend on the specific technology in a given situation in which negative consequences are possible" (McKnight et al. 2009, 7). This kind of trust is formed by two components: trusting intention, i.e., willingness to depend on the technology, and trusting beliefs, i.e., the judgement that the technology has desirable attributes, particularly reliability, functionality, and helpfulness. Trusting beliefs are positively related to trusting intention, as individuals are more willing to depend on technology that they believe is trustworthy.

Propensity to trust general technology refers to the willingness to "trust technology across situations and persons" (McKnight et al. 2009, 7). It can take two forms. On the one hand, the propensity to trust in general technology may result from the belief that technology is usually consistent, reliable, functional and helpful. On the other hand, it may be reflected in the belief that positive outcomes will result from the use of technology.

Finally, institution-based trust in technology arises from the belief that, when using technology, "success is likely because of supportive situations and structures" (McKnight et al. 2009, 9). This form of trust depends either on the belief that "success with the specific technology is likely because one feels comfortable or favourable when one uses the general type of technology of which this specific technology is an instance" or when "regardless of the characteristics of the specific technology, one believes structural conditions like guarantees, contracts, support, or other safeguards exist in the general type of technology that make success likely" (McKnight et al. 2009, 9).

Building on this account of "trust in technology", we define "trust in AI systems" as *the attitude of a person to accept AI decisions or to rely on an AI system to perform a task, accompanied by the belief that the system will make a decision or perform the task in line with the person's expectations*. This definition is consistent with, and indeed extends, the previous account of trust in technology provided by McKnight et al. (2009). First, it takes into consideration the two components of trust in specific technology, i.e., trusting intention and trusting beliefs. The person's willingness to rely on an AI system to perform a task represents the trusting intention, while the belief that the system will perform the task or make a decision in line with their expectations represents the trusting belief. On top of this, our definition also implies a distinction between people who actively use AI systems and thus can decide to rely on them, and people who are passive subjects in their relationship with AI systems, and thus can only accept AI decisions. This extension reflects the growing role of AI systems in decision-making processes, and the importance of trust in these systems for individuals who are subject to AI-generated decisions. For example, if we consider the use of an AI system for disease diagnosis, a doctor is an active user because they rely on the AI system to make a diagnosis, while the patient is a passive subject because they do not use the AI system but are affected by its decision. When there is trust in AI systems, both doctor and patient show the intention and belief to rely on the system and accept its decisions respectively. Consequently, when we talk about "trust in AI systems", it is important to identify who is trusting, because it helps in understanding the different perspectives and expectations of these two groups. Those who actively use AI systems and those who are affected by its decisions may have different requirements, criteria and thresholds for trust, and it is therefore important to take them into account for designing and developing of AI systems aligned with the expectations and needs of both active users and passive subjects.

Secondly, our definition considers "trust in AI systems" as a particular case of trust in general technology, in that it is described as an attitude and is therefore similar to the propensity to trust in general technology. Consequently, when the objective is to build trust in AI systems, one does not start from scratch, but from a pre-existing level of trust in related technologies. This is important to take into account when developing strategies to build trust in AI systems.

Finally, the concept of institution-based trust is also incorporated in our definition, as we acknowledge that the expectation of positive results from the use of technology can be motivated by the belief that there are institutional arrangements in place that make technology trustworthy. Regulations, standards, certifications, market surveillance, and assurance mechanisms can play a crucial role in building trust in AI systems.

Although our definition still entails certain philosophical assumptions, such as trust as an attitude and trust as a function of technology's properties, we believe it provides a more practical understanding of the concept of trust in the context of the "transparency-trust thesis". What the proponents of this thesis are concerned about is that the opacity of AI systems hinders the uptake of AI and prevents diffusion of its benefits, which is why they propose transparency and explainability as a solution. User trust, and more broadly public trust, is instrumental in fostering the uptake of

AI. With "trustworthy AI", users will use artificial intelligence more, and the general public will be more willing to be subjected to AI-based or AI-assisted decisions relevant to their lives.

4. Transparency and explainability don't build trust in AI systems

After the conceptual clarification of the main terms involved in the "transparency-trust thesis", we can now turn to verifying its conceptual consistency and validity. In this section, we provide some arguments to support the idea that when the objective is to create trust in AI systems, transparency and explainability are neither necessary nor sufficient and therefore it is not rational to pursue them for this reason alone. This amounts to a rejection of the "transparency-trust thesis". This does not mean that there are no other reasons to pursue AI transparency, like some epistemic value.

We first elucidate the relationship between the concepts involved in the "transparency-trust thesis". Proponents of this thesis posit that opacity, transparency, and explainability significantly influence trust in AI systems by affecting their understandability. Opacity, primarily stemming from the scale and complexity of AI systems, diminishes their understandability (Section 3.1). Conversely, transparency and explainability are believed to enhance the understandability of AI systems by diminishing their opacity (Section 3.2). As opacity, transparency, and explainability all pertain to understandability, we concur with Barredo Arrieta et al. (2020, p. 88) that this is the central concept in this discussion. In the "transparency-trust thesis", understandability acts as a pivotal factor connecting transparency and explainability to trust in AI systems. The argument posits that users are more likely to trust AI systems that are more understandable due to increased transparency and explainability. In contrast, when AI systems are opaque and challenging to grasp, users' trust may diminish.

Consequently, we examined the concept of understandability (Section 3.3) in the context of AI systems. We first noticed that understandability depends on the user's background, expertise, cognitive abilities, and familiarity with AI systems. Motivated by this insight, we focused on characteristics of an AI system that promote understandability. We pinpointed qualities such as linearity, monotonicity, limited and straightforward interactions among features, and rule-based nature, ultimately concluding that simplicity is a crucial factor in understanding AI systems. This finding aligns with the notion that the complexity of AI systems, conversely, is a primary source of opacity that hinders understanding.

Finally, we clarified what it means to trust an AI system (Section 3.4). We proposed a working definition based on McKnight et al.'s (2009) account of trust in technology. According to this definition, trust in an AI system is the attitude of a person to accept AI decisions or to rely on an AI system to perform a task, accompanied by the belief that the system will make a decision or perform the task in line with their expectations. In line with some of the empirical studies presented in Section 2, our account of trust in AI systems recognises that accuracy and reliability, as well as

the existence of supportive institutional structures, contribute to building trust in AI systems. Interestingly, understandability plays no role in McKnight et al.'s (2009) account of trust in technology.

Based on these premises, we can now develop a twofold argument against the "transparency-trust thesis". On the one hand, we reject the assumption that transparency and explainability convey a significant understanding of AI systems. On the other hand, we reject the idea that understandability is a necessary and sufficient element for trust.

In order to support the first side of our argument, we need to clarify what the role of explanations is in promoting the understandability of AI systems. In our account, since understandability is linked to simplicity, of the key issue is identifying a link between explanation and simplicity. In this context, we argue that explanations should work to bridge the gap between the inherent complexity of AI systems and the user's ability to understand their inner workings to the extent that this is simple enough. This is exactly what XAI aims to do. For example, LIME (Ribeiro et al. 2016) is an algorithm designed to explain the predictions made by AI black-boxes. It works by creating a simple linear model that locally approximates the complex AI model to be explained. It does so by perturbing the input features and observing the changes in the output. Based on these observations, LIME learns a sparse linear model that captures the underlying patterns in the AI's decision-making process. Using this linear model, LIME can then generate explanations for the AI's predictions in the form of lists that rank the most important factors contributing to the decision. This simple rank amounts to an explanation of the main elements influencing the underlying complex processes of the original model. Similarly, a visualisation like a saliency map can serve as an explanation by highlighting the most important regions of the input data, such as image pixels, that contribute significantly to the model's outcome. In other words, a saliency map shows which parts of the image have the most influence on the model's decision-making process. These visualisations simplify the understanding of the AI system by focusing on key input areas.

Accordingly, the idea of making AI systems transparent and explainable ultimately amounts to creating simple models for complex systems in a way that the simple model still conveys all important characteristics of the complex system. However, this is not always possible. Several challenges limit the effectiveness of transparency and explainability in conveying significant understanding of AI systems.

First of all, large and complex systems considered transparent by design are not always understandable. For example, the size of decision trees may be such that it is difficult to understand them, even though the operational principle is very simple. As Molnar (2022, 84) writes: "the more terminal nodes and the deeper the tree, the more difficult it becomes to understand the decision rules of a tree". Consequently, given an input, it may be difficult for a human to understand why a decision tree model produces a certain output.

Secondly, one of the main issues related to post-hoc explainability techniques is that the explanations are not faithful to what the original model computes (Rudin 2019). This is equivalent to saying that the explanations do not actually explain how the model works. Ghassemi et al. (2021) and Rudin (2019) make this point

when analysing saliency maps as an explainability method. For example, highlighting a region of an image only tells where a neural network looks to classify an image. But it says nothing about why it focuses on that region, nor whether this is correct or whether saliency map allows any inductive reasoning about future prediction.

Thirdly, our analysis of opacity-3 (Burrell 2016), fundamental opacity (Grünke 2019), and w-opacity (Boge 2022) shows that there is a limit to the possibility of making complex AI systems simple and understandable. As these studies show, the complexity, the size, the way certain phenomena are modelled by deep neural networks, constitutes an insurmountable obstacle to human understanding. In fact, the best performing models not only are opaque, but they are usually an ensemble of many complex and opaque models. This scale and complexity allow these systems to match or outperform humans in many relevant tasks, and this is where the benefits and risks of AI lie.

Finally, since explainability methods do not reduce the complexity of an AI system, just create a less complex, thus more understandable model of them, they introduce a gap between the explaining model and reality. Due to the inability to reduce complexity of the real system on the one hand and the upper limit of the complexity of a human-understandable model on the other hand, the gap can be very large indeed. The existence of the gap means that several simplifying decisions need to be made while constructing the model, each of them underdetermined by the real system itself. The result is inevitably a less complex but also less accurate model, and we have already seen that the lack of accuracy can undermine trust. In our case it could mean that the trust in the explaining model can come into question.

As for the other side of the argument, even if in the future the development of explainability techniques were to succeed in finally dispelling the black box problem and make any AI system understandable, this would still not be enough to build trust in AI systems. As shown in our account of "trust in AI systems" in section 3.4, trust is a multidimensional construct involving multiple dimensions. However, understandability is not a factor influencing trust in technology. Trusting intentions and beliefs are reinforced when technology has attributes such as consistency, reliability, and helpfulness. Transparency and explainability do not facilitate the attribution of these properties to an AI system. On the contrary, extensive testing and validation, as well as sound institutional arrangements and practices, do.

5. Conclusion

In this paper we have argued against the "transparency–trust thesis" according to which transparency and explainability build trust in AI systems. We first showed that empirical studies on the topic provide conflicting evidence regarding the effect of transparency and explainability on trust. For this reason, we shifted the analysis to a conceptual level, clarifying the main concepts involved in the "transparency–trust thesis". On this basis, we found that transparency and explainability do not contribute to a complete understanding of an AI system, and that explainability is

not a relevant factor in building trust in AI systems. Accordingly, we concluded by rejecting the "transparency-trust thesis".

Our conclusions do not imply that the results achieved in the field of explainable AI are invalid, but only that they are not relevant for building trust in AI systems. Anyone wishing to challenge our conclusion would therefore have to demonstrate two claims. The first is that understanding is a constitutive factor of trust in AI systems. The second is that transparency and explainability significantly increase our understanding of large, complex and opaque AI systems. If these issues are explicitly addressed, clarified, and, contrary to our expectations, positively solved, this could give a more solid foundation to the field of Explainable AI.

Meanwhile, our findings suggest that we proceed in a different direction when it comes to building trust in AI systems. On the one hand, as also argued by Ghassemi et al. (2021), rigorous validation could prove the trustworthiness of an AI system even in the absence of an understanding of its inner workings. This validation should be as holistic as possible, involving datasets, algorithms, models, and adopt a socio-technical perspective, considering impacts on individuals and society. On the other hand, institutional arrangements such as regulations, standards, certifications, market surveillance, and assurance may contribute to building institution-based trust in AI systems. In a society where AI systems play an increasingly important role in high-stakes decision-making processes, it is essential that such systems are trustworthy and that both active users and passive subjects trust them. And that they have good reasons to do so.

References

AI HLEG. "Ethics Guidelines for Trustworthy AI." Accessed December 31, 2019. https://ec.europa.eu/futurium/en/ ai-alliance-consultation

Alvarado, R. "What kind of trust does AI deserve, if any?" *AI and Ethics* 3, no. 4 (2023): 1169–1183. https://doi.org/10.1007/s43681-022-00224-x

Andrada, G., Clowes, R. W., and Smart, P. R. "Varieties of transparency: Exploring agency within AI systems." *AI & SOCIETY* 38, no. 4 (2023)*:* 1321–1331. https://doi.org/10.1007/s00146-021-01326-6

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020)*:* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bertino, E., Merrill, S., Nesen, A., and Utz, C. "Redefining Data Transparency: A Multidimensional Approach." *Computer* 52, no. 1 (2019)*:* 16–26. https://doi.org/10.1109/MC.2018.2890190

Boge, F. J. "Two Dimensions of Opacity and the Deep Learning Predicament." *Minds and Machines* 32, no. 1(2022): 43–75. https://doi.org/10.1007/s11023-021-09569-4

Burrell, J. "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data & Society 3*, no. 1 (2016). https://doi.org/10.1177/2053951715622512

Facchini, A., and Termine, A. "Towards a Taxonomy for the Opacity of AI Systems." In *Philosophy and Theory of Artificial Intelligence* edited by V. C. Müller, 73–89. Springer International Publishing, 2021. https://doi.org/10.1007/978-3-031-09153-7_7

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28, no. 4 (2018): 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. and Sanders, J. W. "On the Morality of Artificial Agents." *Minds and Machines* 14, no 3 (2004): 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Freiman, O. "Making sense of the conceptual nonsense 'trustworthy AI'." *AI and Ethics* 3 (2023): 1351–1360. https://doi.org/10.1007/s43681-022-00241-w

Ghassemi, M., Oakden-Rayner, L. and Beam, A. L. "The false hope of current approaches to explainable artificial intelligence in health care." *The Lancet Digital Health* 3, no. 11 (2021): 745–750. https://doi.org/10.1016/S2589-7500(21)00208-9

Grünke, P. "Chess, Artificial Intelligence, and Epistemic Opacity." *Információs Társadalom* 19, no. 4 (2019): 7-17. https://doi.org/10.22503/inftars.XIX.2019.4.1

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51, no. 5 (2019): 1–42. https://doi.org/10.1145/3236009

Haque, A. B., Islam, A. K. M. N., and Mikalef, P. "Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research." *Technological Forecasting and Social Change* 186 (2023). https://doi.org/10.1016/j.techfore.2022.122120

Héder, M. "The epistemic opacity of autonomous systems and the ethical consequences." *AI & SOCIETY* 38 (2023): 1819–1827. https://doi.org/10.1007/s00146-020-01024-9

Kandul, S., Micheli, V., Beck, J., Kneer, M., Burri, T., Fleuret, F., and Christen, M. "Explainable AI: A Review of the Empirical Literature." *SSRN Scholarly Paper* (2023) https://doi.org/10.2139/ssrn.4325219

Kartikeya, A. "Examining Correlation Between Trust and Transparency with Explainable Artificial Intelligence." *Lecture Notes in Networks and Systems* 507 (2022): 353–358. https://doi.org/10.1007/978-3-031-10464-0_23

Kizilcec, R. F. "How Much Information? Effects of Transparency on Trust in an Algorithmic Interface." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395. New York, USA: Association for Computing Machinery 2016. https://doi.org/10.1145/2858036.2858402

Koene, A., Clifton, C., Hatada, Y., Webb, H., and Richardson, R. *A governance framework for algorithmic accountability and transparency*. 2019.
https://doi.org/10.2861/59990

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research." *Artificial Intelligence* 296 (2021).
https://doi.org/10.1016/j.artint.2021.103473

Lipton, Z. C. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue* 16, no. 3 (2018): 31–57.
https://doi.org/10.1145/3236386.3241340

Liu, C.-F., Chen, Z.-C., Kuo, S.-C., and Lin, T.-C. "Does AI explainability affect physicians' intention to use AI?" *International Journal of Medical Informatics* 168 (2022).
https://doi.org/10.1016/j.ijmedinf.2022.104884

Lopez, J., Textor, C., Lancaster, C., Schelble, B., Freeman, G., Zhang, R., McNeese, N., and Pak, R. "The complex relationship of AI ethics and trust in human–AI teaming: Insights from advanced real-world subject matter experts." *AI and Ethics* 4 (2024): 1213–1233.
https://doi.org/10.1007/s43681-023-00303-7

McKnight, H., Carter, M., and Clay, P. *TRUST IN TECHNOLOGY: DEVELOPMENT OF A SET OF CONSTRUCTS AND MEASURES* (2009).
https://core.ac.uk/download/pdf/301349124.pdf

Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, 2022.
https://christophm.github.io/interpretable-ml-book/

Páez, A. "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)." *Minds & Machines* 29, (2019): 441–459.

Papenmeier, A., Englebienne, G., and Seifert, C. "How model accuracy and explanation fidelity influence user trust in AI." In *IJCAI Workshop on Explainable Artificial Intelligence* (XAI) 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. New York, USA: Association for Computing Machinery, 2016.
https://doi.org/10.1145/2939672.2939778

Rudin, C. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence*1, no. 5 (2019): Article 5.
https://doi.org/10.1038/s42256-019-0048-x

Schmidt, P., Biessmann, F., and Teubner, T. "Transparency and trust in artificial intelligence systems." *Journal of Decision Systems* 29, no. 4 (2020): 260–278.
https://doi.org/10.1080/12460125.2020.1819094

Shin, D. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI." *International Journal of Human Computer Studies* 146 (2021).
https://doi.org/10.1016/j.ijhcs.2020.102551

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., and Freitas, A. "Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making." *Artificial Intelligence* 316, (2023). https://doi.org/10.1016/j.artint.2022.103839

Zednik, C. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34, no. 2 (2021): 265–288. https://doi.org/10.1007/s13347-019-00382-7

Zerilli, J., Bhatt, U., and Weller, A. "How transparency modulates trust in artificial intelligence." *Patterns* 3, no. 4 (2022). https://doi.org/10.1016/j.patter.2022.100455