Calogero Caltagirone[*], *Lucy Conover*[**], *Livio Fenga*[***], *Federica Russo*[**], *Dolores Sanchez*[****], *Angelo Tumminelli*[*]

On the Relationship Between Interpersonal Trust and Technological Reliability: Some Reflections on Trustworthy AI in the AI Act[*****]

Abstract

Trust and trustworthiness have become central concepts in the ethics and governance of artificial intelligence (AI). The issue arises because of the inherent uncertainty of AI system outputs, which are often described as opaque systems or black boxes. The need of a legal/regulatory framework, as a response and remediation to AI opaqueness, has been rapidly recognised by the European Union, which has put in place, in a time record, a powerful piece of legislation: the AI Act. Valuable efforts of the EU notwithstanding, regulation shows tension with private/public law, and mostly tension in understanding of trust/trustworthiness from jurisprudence. The tension arises because, from a legal perspective, an instrument cannot possess trust, which, in some philosophical traditions, is defined as a strictly human-to-human relation. This tension is substantiated, for instance, in philosophical anthropology. Yet, as the terminology of trust has stuck, in this paper we offer a different understanding of this notion. Building on philosophical and anthropological contributions, we extend the meaning of trust to technical artefacts using an argument by analogy. We propose a semantics for trust/trustworthiness based on a relational framework that places AI in a network of relations, and so

* LUMSA University, Rome.
** Utrecht University.
*** University of Exeter.
**** University Carlos III, Madrid.
***** This research was funded by the European Union HORIZON-RIA SOLARIS project, grant number 101094665. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Authors are listed in alphabetic order. Angelo Tumminelli and Calogero Caltagirone conceptualized the paper, initially coordinated the writing of the text, and drafted section 4 and 5. Livio Fenga drafted section 2, Dolores Sanchez drafted section 3. Federica Russo and Lucy Conover drafted introduction and conclusion, and contributed to contents and argument structure of all the sections, harmonizing the text and coordinating the revisions. All authors reviewed and approved the final version. We thank Antonio Estella, the editors of the journal, and the reviewers for helpful comments on earlier versions of this article.

SIFM

FM/MP *Filosofia morale/Moral Philosophy*, vol. 2, 2025 • ISBN: 9791222328041 • DOI: 10.7413/fmmp0076
© 2025 – MIM EDIZIONI SRL

trust/trustworthiness are not properties of the artefact alone, but of a network in which human agents still play a fundamental role.

Keywords
Artificial Intelligence, AI Act, Trust, Interpersonal Trust, Technological Reliability.

1. Introduction

Can/should we trust machines? In particular, can we achieve an attitude of trust towards digital devices, and in particular, Artificial Intelligence (AI)? The recent and rapid developments in the field of computer science and AI place these questions centre-stage of ethico-legal debates. The question whether we humans can trust machines is not new, but the debate features a new dimension. The latest generation of AI systems, also called generative AI, has extraordinary capabilities in terms of computational power, with the ability to generate outputs from a given prompt. The focus of this paper is not on whether these AI systems genuinely generate and understand new contents, but on whether human agents can trust them (and their outputs). As the discussion unfolds, we will notice that a reassessment of the notion of trust and the corresponding characteristics of "trustworthiness" (that is required by an increasing need for legal acts and norms, not least the EU AI Act) is needed. The question arises because, as is already well-documented, these AI systems generate outputs that are used to inform humans' decisions and actions. In this sense, it is correct to call these systems epistemic or cognitive technologies (Alvarado 2002; Babushkina, 2013), and to raise the question whether, to what extent, or under what conditions, we can or should trust them.

From a methodological point of view, the paper utilizes the distinction between three fundamental categories related to the concept of "trust" (Fabris 2020):

a. Trust as an interpersonal event: In this sense "trust" indicates the *interpersonal* and anthropological relationship by which relationships between people are built.

b. Trustworthiness: reliability as an extension of trust: When, instead, we speak of trustworthiness, we mean that interpersonal trust can also be extended to the relationship between human beings and technical artefacts.

c. Reliability as procedural safety: From a legal point of view, the concept of "reliability" is preferred; reliability indicates the procedural safety of a device, i.e. its performative ability to perform the task for which it was designed. Technical artefacts are reliable because they offer no surprises, if they properly work and follow certain procedures.

This paper will explore the possibility of applying the *anthropological* category of "trust" (first definition above) to human-machine relationships (second definition) to clarify whether it is possible to extend the experience of trust to interactions between individual subjects and artificial agents; the question arises because of the reliability of these technologies (third definition). Categories a-c are not mutually exclusive. In developing our argument, we will show how all three are needed to justify talking about trust in AI/trustworthy AI. On this basis, the paper re-considers the possibility of including the concept of trust in legal regulation. The first part of the article will investigate trust as a fundamental anthropological structure, rooted in the experience of interpersonal relations, and then move on to the study of the interaction between humans and AIs, focusing on "technological trust" from a normative and statistical point of view. A complex epistemological framework emerges from the article, bringing together perspectives that may differ from one another but are harmonised in the view that trust can be attributed to an artefact such as an AI system not directly, but in an analogical way, and because the artefact is part of a socio-technical system in which actors stand in various relations with each other.

2. Predicting Complexity and Technological Reliability

With the renewed interest in artificial intelligence, and the rapid spread of AI systems in various sectors, we have witnessed increasing attention to the reliability of outputs of such systems. In particular, the question whether or not we can trust outputs we cannot fully understand, often called "opaque" or "black-box" systems. This section examines the various techniques to increase the technological *reliability* of AI and mathematical models (category c mentioned in section 1).

One approach is the development of explainable AI (XAI). XAI aims to make black box models more interpretable without sacrificing their predictive power. Techniques such as feature importance scoring, partial dependence plots, and local interpretable model-agnostic explanations (LIME) provide insights into the decision-making processes of "black box" models (Ribeiro *et al.*, 2016; Lundberg & Lee, 2017; Linardatos *et al.*, 2021). These tools help users understand why a model made a particular decision and identify potential biases. At the core of the rapid advancement of artificial intelligence (AI) lies a critical distinction between "black box" models and "white box" models. The differences between these models have implications for the technological reliability of AI and mathematical models, and the measures needed to ensure trust and accountability in these systems.

White box models are defined by their transparency and interpretability. These models rely on explicit mathematical equations derived from theoretical principles or empirical observations (Hastie *et al.*, 2013; Salih & Wang, 2024). The transparency of white box models allows developers to fully understand and interpret the internal workings of the model. The primary advantage of white box models lies in their ability to provide insights into the system's behaviour and causal relationships. This is particularly important in fields such as engineering, physics, and economics, where understanding the underlying mechanisms is crucial for both advancing knowledge and making informed decisions.

The internal workings of black box models are opaque, making it difficult to discern how inputs are transformed into outputs. Despite this opacity, black box models are renowned for their significant predictive power, especially when handling large and complex datasets. Examples of black box models include neural networks, support vector machines, and ensemble methods like random forests and gradient boosting. The primary strength of black box models is their flexibility and ability to adapt to new data, however, the lack of interpretability in black box models poses significant challenges. Without a clear understanding of how decisions are made, it becomes difficult to diagnose errors, ensure fairness, and maintain trust in the model's predictions. This can be particularly problematic in critical applications such as healthcare and finance, where the consequences of incorrect predictions can be severe (Obermeyer *et al.*, 2019; Rudin, 2019; Hasanzadeh *et al.*, 2025).

Choosing between white box and black box models depends largely on the specific requirements of the problem at hand. In practice, a hybrid approach that combines the strengths of both white box and black box models can be highly effective. Interdisciplinary collaboration is another key factor in ensuring the reliability of AI and mathematical models. By fostering collaboration between AI developers, domain experts, ethicists, and regulators, we can ensure that these models are not only technically sound but also aligned with ethical and social norms. For instance, in healthcare, collaboration between AI researchers and medical professionals can help develop AI systems that are both accurate and trustworthy. In conclusion, there is a significant body of literature on the technical side of explainability when it comes to AI. To supplement this, we need a greater focus on how regulation comes into play. Balancing regulation, explainability, and trustworthiness poses challenges, which we explain in the next section.

3. Trust and Legal Regulation

When it comes to regulating the use of artificial intelligence, how do we define trust as a part of legislation? Since early EU engagement on AI regulation, there has been consensus and understanding that the proposed regulatory framework for AI should foster trustworthy AI.[1] There is an outline, the "Trustworthy AI Guidelines," proposing "lawful, responsible, and robust" use of AI technology, as designated by the 52 members of the High-Level Expert Group (HLEG) (Floridi, 2019, p. 1). Under these lines, can the HLEG Trustworthy AI guidelines serve as a case study to investigate the question: *can trust be implemented into regulations?* The HLEG Trustworthy AI guidelines consist of seven essential qualities for trustworthy AI (European Commission):

1. human agency and oversight,
2. technical robustness and safety,
3. privacy and data governance,
4. transparency,
5. diversity,
6. non-discrimination and fairness,
7. social environment and well-being,

Comparing these seven essential qualities with the three definitions of trust from Fabris (2020), we see there is more attention on tangible focus, such as technical robustness and data governance for example, than on more distinctively philosophical aspects. Pushback for such a framework mainly stems from a charge of anthropomorphism. According to this argument, trust can only be attributed to a human whose qualities are falsely projected onto the technology in the form of regulation (Ryan 2020). Along these lines, in the cautionary words of HLEG Member Thomas Metzinger: "Machines are not trustworthy; only humans can be trustworthy (or untrustworthy)" (Metzinger 2019). Is this critique of anthropomorphism misguided? We need more intentional research on what modes of trust are being placed in the technology, bearing in mind that the question is not just theoretical but has important consequences at the jurisprudential level.

In the debates in Philosophy of Technology, some have argued that there are "types" of trust that can be attributed to a human, and those that cannot. For example, Philip Nickel, makes a distinction between rational-choice and motivation-attributing accounts of trust, with the former being like reliability, and the latter applying to the motivation

[1] Within this act, are the 2019 Trustworthy AI Guidelines (Recital 27 2019)

of the trusted themselves. Given the more nuanced nature of the second form of trust, it is not so clear how motivation-attributing trust would apply to a non-human technological device (Nickel 2019, p. 430). Furthermore, the anthropomorphism discussion connects to the directionality of trust. Political scientist Russell Hardin investigated the difference between mutual trust versus one-way trust in his book *Trust and Trustworthiness*. With mutual trust, the desire to be perceived as trusted by your trustor is an incentive to trust, and vice versa. This does not necessarily apply to one-way trust, as the trusted is not the actor who benefits from the trust (Hardin 2006, p. 46). Using this definition, mutual trust would be impossible to achieve with AI, unless you were to somehow prove it consciously benefitted from being trusted. With these perspectives in mind, certain types of trust are possible with AI (as we will discuss in the next section), and trust should not be completely taken off the table for regulations. Rather, conceptualisation of "trust" should be made more specific, keeping in mind that technological progress receives its ethical orientation from humans.

Yet, there are numerous issues with this argument. For one, a large portion of the sources investigating trust are written before the "AI boom" that started in late 2022, or about technology in a broader sense, and within the context of cybersecurity, or yet again about trust in a non-technological context, or all the above. We can certainly use this literature, such as Nickel and Hardin, to help understand trust, but the concern is that trusting AI is something completely new and serves a strategic purpose: induce people to place trust in AI so that they will use it more and, hence, unlock the technology's economic and social potential (Laux et. alt. 2024). Thus, pre-existing literature on trust must be used cautiously, especially if we take into consideration the Private Law sphere where the interplay with current liability schemes and damage law may trigger unexpected consequences.

A second consideration is that the AI Act comes to fit into a very well established architecture of other regulatory instruments (GDPR, Product Liability, etc.). In the Public Law sphere, the focus is on the construction of a robust set of legal principles that will be used in relation to AI and touch upon a myriad of well-established principles that range from access to justice, equality, legal security, to the very essence of Democracy providing the right governance framework. There has been an increased use of data algorithms and AI instruments in public administration institutional bodies to govern our societies (Janssen and Kuk 2016), but the legal notion of trust within these regulatory contexts and the uncertainties surrounding the future development of AI technologies has no antecedents and needs a new doctrinal outline in a sector that is predominantly monopolized by private companies.

Furthermore, the AI Act (AIA), enacted on Feb 2, 2025, presents a series of rules and guidelines for the ethical use of AI, including risk categories, prohibited uses for AI (e.g. social scoring systems), responsible management of biometrics etc. The High-Level Summary of the AI Act introduces the normative notion of trustworthiness. Thus, since early EU engagement on AI regulation the notion of trust has been assumed to be central in the regulatory process, as a way to address principles such as safety, security, transparency and fundamental rights, all at once. It is important to underline that in the ambitious regulatory coverage of the current digital transformation in Europe, there exists an oversimplification of a highly complex and heterogeneous set of closely related concepts around the normative notion of trust in its current approach, and well captured in the categories laid down by Fabris. Yet, there is a real need to adopt the AI Act, and Recital 148 of the AI Act explicitly states:

> […] the effective implementation of the AI Act (AIA) throughout the EU depends on uniform, coordinated and well-funded governance setting. The normative framework of AI generally speaking after the approval of the Act by the EU parliament the 3rd of March, even though not immediately enforceable, has been widely and positively accepted in the middle of a reality where uncertainties surround the future development of AI technologies and their social impacts and still the dominant setting.

The normative perspective grounds the attribution of trust in legal theory as well as in assumptions at the level of philosophical anthropology. Nedelsky (2011), for instance, provides a good introduction to the theoretical assumptions at work in the field of law. The approach is based on the relational dimension of law as part of a human experience central to the concept and institutions by which we organize our collective lives. Still, the fact that public institutions and governments are also implementing AI at a rapid pace is an important drive to further consider the complexity of the current regulatory context and the need for a rightly-founded normative account of trust and AI reliability at stake without departing from the relational dimension of law but understanding that some assumptions could be disruptive within the legal ontology. Yet, as we aim to show in the rest of the paper, attribution of trust to AI systems requires an exercise in conceptual design that we undertake in Sections 4 and 5.

Recent analyses of legal nature introducing variables that could influence trust with the intention to map the different positive values that help building trust in regulatory contexts (i.e. public transparency, public accountability, participatory models for AI governance, consent, etc) are a valid illustration of sounded arguments supporting the AI Act's notion of trust. There is already some relevant scholarship effort

addressing this issue in the highly controversial field of cybersecurity (Taddeo, Floridi 2019). Floridi, one of the experts of the HLEG on this work "Establishing the rules for Building Trustworthy AI" makes a conceptual distinction to address the "reliability" of AI systems and the "human trust". He asserts: 'since AI will become increasingly important and pervasive, it must work reliably, in ways that anyone can trust will be for the benefit of humanity and the whole environment". (Floridi 2019, p. 61).

This analysis, far from exhaustive, intends to shed light on the understanding of the notion of trustworthiness introduced by the AI Act, the current model of risk regulation and the perceived acceptability of risks based on a trust relationship with EU institutional setting and regulators. The idea behind it is that the effort to develop "trustworthy AI" through regulatory laws such as the AI Act acknowledges a need for AI to be trusted if it is to be widely adopted. This has been recently highlighted by Laux et alt. (2024) affirming: "The emphasis on trustworthiness serves a strategic purpose: induce people to place trust in AI so that they will use it more and, hence, unlock the technology's economic and social potential."

In the following, we re-analyse trust from the perspective of philosophical anthropology (section 4), and we supplement this anthropological understanding with a "network approach" (section 5). A network approach does not limit trust to the technology itself but rather distributes trust across the entire network of actors and information connected to it. In such an approach, regulations could and should define trustworthy AI not only within the scope of AI itself, but across all the actors involved in creating AI, as well as producing the information the AI uses to operate.

4. The Anthropological Status of Trust in Interpersonal Relations

Before developing on the idea of "trust as a network" (section 5), we mobilise literature from philosophical anthropology to problematize the notion of "trust", rooted in human-human relations (i.e., category a of Fabris 2020).

According to some philosophers, human beings are "from" and they live "for" trust, and their being and actions are structured by the experience of trust (Buber 1923, Caltagirone 2020, Fabris 2020). As a relational act, trust is rooted in the mutual recognition between persons, and constitutes the key to human identity: in fact, those who live without a trusting disposition in the direction of the Other/their Others are unable to achieve human fulfilment (Alici 2012, p. 64). According to Alici, by living trust, humans say a fundamental "yes" to life and the relationships that

establish it; the human being opens themself to themself, to others, and to the world.

Trust, in its original anthropological structure, conditions all human experience and behaviour, establishes the human quality of each person's relationship with existence in the forms of the relationship with oneself, others, the things of the world, with transcendence (Giddens 1994). As a constitutive dimension of humanity, and a condition of possibility for existential openness, trust expresses both an attitude of reliance on existence and existence, perceived as good and worthy of being experienced, and the relational dimension that makes human existence possible in its inception and development. Because it is constitutive, trust is at the origin of the human experience, as it expresses the promise of a good life on the part of otherness. In this sense, existential trust, which marks the common and universal human experience of the shared human, always allows new progress and conquests, configuring itself as a fundamental hope, which resists all disappointments. It is an anthropological "figure" whose connotation lies in the possibility of giving and restoring credit, having originally received it, to the quality of human relations and relations between humans, due to the fact that the humanity of man, the subject of relations in relation, has in relying on someone, considered and experienced as trustworthy, its original and originating core (Buber 1923, Caltagirone 2019).

Trust in otherness, being a condition of possibility for a truly human life, develops in conjunction with the formation of an original intimate sense of trust that, later, in the existential unfolding, lays the foundation for a stable identity of the self. In fact, existential trust, which stems from the direct experience that every human realises in their own life, is the initial movement of entrustment to the original otherness. This, in the very act of trusting, indicates its specific way of existing and being in the world, characterised by the weaving of relational bonds that establish and constitute it in its dignity of being and acting. This is because trust, in a reciprocal process of openness to otherness, is the basis of human relationships. As the central "figure" of human experience, trust makes one truly and fully human, since the act of trusting an original and originating otherness, being structurally relational, is rooted in the heartfelt recognition of the trustworthiness of others. This act is given in the reciprocity of the relationship that exists between the moment of passivity, as a condition of the pathos of human experience, and the moment of activity, as a practical situation understood as a decisive and peculiar characteristic of human action.

The human is not thinkable and identifiable without the experience of co-constitutive reciprocity, which in its originality is an experience of wonder and surprise. In this sense, trust implies from the outset that reci-

procity of relational experiences establishes reliable bonds. By discovering as their own what they have not placed, inasmuch as they are never at the origin of their own beginning, it is in "feeling" and being "affected" that the human, the subject of relations in relationship (Totaro 1992).

While trust is addressed to a concrete "Thou" who stands before (Buber 1923), trustworthiness is the attitude of one who trusts in an undefined and not immediately recognisable otherness such as that of transcendence. For example, in the Jewish context the experience of faith is an experience of reliance, in Hebrew "emunà" (Buber 1950). Similarly, in the case of the relationship between the human being and technology (category b of Fabris 2020), we can also speak of reliance because when the subject trusts a technical artefact, he/she is directing his/her trust not so much towards an immediately recognisable Thou, but towards a technological apparatus at the origin of which the ethics of the designer must be placed. Therefore, it is now appropriate to return to the concept of *trustworthiness*.

The idea of the trustworthiness of the other is indeed fundamental to the sense of continuity of human identity and is based on the mutual and universal recognition between humans who reciprocally entrust themselves, honouring and loving one another. A reciprocity, which being based on the correlation of call, response and involvement, speaks of the original structure of the shared human that is actualised in receptivity, particularly in corporeality/spatiality. In this way, reliance indicates a universal trust that extends to the totality of humans, even those not immediately recognisable.

Indeed, in its nature, every human being is called upon to trust otherness. Being the condition of access to the meaning of reality for every human being, who chooses to entrust themselves to what is proposed to them as credible, thus arriving at a real awareness of himself, trust represents the horizon within which every human life open to meaning is nourished in all its forms of experience. Trusting "of" Other/others in a full and direct reciprocity accompanies trusting in otherness. For example, it is expressed in an exemplary manner in the reciprocity of the child's affective recognition through the maternal and paternal affective experience and, as such, comes to constitute itself as a paradigmatic form of the trust relationship. This relationship of reciprocity, however, should not be understood only between two individuals, but also as a network of relations. In other words, reciprocity is inherently relational, including a socio-technical network (and all sorts of normative preconditions for and implications for trust), as well as material conditions.

Therefore, if trust is a condition of possibility of the human being, it pervades every relationship that the individual lives, including that with technology. Since technical artefacts are not human, but the artificial

extension of the subject's creative intentionality, it is more appropriate to speak of trustworthiness, according to category b of Fabris (2020). A technical artefact, in fact, is reliable because its credibility can be traced back to the act of the programmer(s) and developer(s) – this simple example evokes the idea of trust as a network, to be further discussed in section 5.

Just as "human" trust is not simply or solely addressed to a person, but is reciprocal and relational, the same applies to "technological trustworthiness". Our argument is to show that "technological trustworthiness" can be used to denote an *analogical extension* of interpersonal trust: we do not simply or solely trust an artefact (or its results), but we trust the process that leads to those results (Russo, Schliesser, Wagemans 2022) and the designer who organised the process. This shift in focus from the output to the process and the designer has important implications.

Firstly, if we trust the "output" of an algorithmic procedure, we are implicitly saying that we trust the process that brought it about. In this sense, trustworthiness in a techno-scientific context is not blind faith. Thus, it would be more accurate to speak of reliability (category c of Fabris 2020), rather than trust, because what we really trust are aspects of the modelling and implementation process. Secondly, although it has become common in everyday scientific and philosophical language to attribute certain properties and actions to artefacts or models, these are convenient shortcuts, but they obscure the fundamental role that humans continue to play in the whole process (Russo 2022). In other words, machine reliability is also an analogue outcome of interpersonal trust, e.g. that the developers or programmers have done the job correctly and that the machine is therefore reliable. It should then be pointed out that, from an anthropological perspective, it is not so inappropriate to speak of trustworthiness in machines as an analogical extension of trust, although from other perspectives (such as the legal one, for example) this does not make sense, as discussed in section 3.

5. Trust as a Network Dimension of the Human Person and its Openness to Technology

Having clarified that trust is a structural dimension of human beings in their intrinsic relation in the last section, we will now try to understand how to analogically extend trust to the relationship between humans and AI. The dimension of trust, extending to technological artefacts, takes the form of trustworthiness (O'Neill 2020). Starting from an anthropological point of view, trustworthiness can be understood as an extension of trust by means of the category of *analogy*. This discourse, valid from

an anthropological and ethical point of view, but more complicated from a purely legal perspective, is nonetheless useful for further clarification of the relationship between designer, device and user.

It is therefore necessary to distinguish trust as an interpersonal experience from technological trustworthiness, as an experience relating to the relationship between humans and technologies. Indeed, as a form of openness to otherness, trust can extend to machines, which are, in fact, a human production and, as such, an expression of personal creative freedom. Therefore, when we say we trust a technical artefact, we trust the people who designed it; in turn designers are also part of a network that includes developers, producers, regulators, etc. In this sense, AI systems are trustworthy if they keep alive the relationship of trust between the person who uses them and the programmer who designed them. Technological reliability then stems from the disposition of trust to extend to otherness by always referring to an interpersonal human experience.

As anticipated, the theme "trustworthy technology" is not new and it has been problematized in the literature before the advent of AI (see, e.g., Nickel 2010). It should be emphasised that trustworthiness is closely linked to the issue of credibility. A system may appear credible because it produces convincing texts or images, but still be unreliable if the information it generates is inaccurate or unverifiable. Credibility, therefore, must be based on interpersonal trust and verifiability. The fiduciary expectation may concern characteristics of the trust recipient that affect their behaviour, role, personality or entire identity. A summary list may include: ability, intelligence, courage, discretion, sensitivity, responsible behaviour with respect to a mandate, authority, consistency, generosity, honesty, adherence to certain values and moral principles, friendship and love. Credibility, likewise, can be understood as a challenge to identity: to be credible and, therefore, worthy of trust, the subject repository of trust places in a position of continuous analysis that leads them to be responsive to the idealised image that is made of them and, therefore, to be better. By way of example, one can think of the educational institution which, in order to respond to the image presented publicly, undertakes to improve its organisational reality. It can be inferred that the relationship of trust is not something external to the identity of the social agents involved but the relationship concretely plays out and constructs their identity. Can a technological artefact and an artificial intelligence device be credible and, therefore, reliable?

Returning to the topic of anthropomorphism, one of the greatest difficulties in assessing AI and analysing its ethical impact is the tendency of people to anthropomorphise it. Moreover, this becomes particularly problematic when we attribute human *moral* characteristics to it. The media offer us images of humanoid machines with extraordinary capa-

bilities daily. Movies, novels, and TV series depict sentient robots, so it is almost natural for us to associate, categorise and define these machines in human terms. In addition, while we associate human activities and abilities with machines, an important problem arises when this "anthropomorphisation" is linked to human moral activities, such as trust. To propose that AI should be regarded as reliable is a very serious statement. One must consider the fact that the AI, not possessing emotional states and thus not being able to be responsible for its actions (a requirement inherent in both the affective and normative value of trust), cannot be considered reliable on such a basis; yet it can be assessed for its reliability.

While AI fulfils the requirements of the relational meaning of trust, we show that this is not a type of existential trust (category a) but is instead a form of reliability (category c). Ultimately, even sophisticated machines such as AI should not be considered trustworthy in themselves, as this undermines the value of interpersonal trust. This anthropomorphisation of AI (affective value of trust), we argue, shifts responsibility away from those who develop and use it (normative value of trust). Tech companies (especially BigTech) work under the assumption that AI is something we can, and should, trust. However, as we have been arguing through this section, we cannot place trust in the technical artefact alone. All actors and processes that are part of the socio-technical system (including AI technology itself) participate in the process of deeming an AI trustworthy. This network perspective is essential because, as noticed above, we never trust simply or solely an artefact or its output, but trust emerges from a network of relations. Bisconti *et al.* (2024) explain that nowadays, with artificial agents gaining the foreground of sociality and communication abilities, we need to rethink the concept of socio-technical systems as synthetic, precisely to signal that many relations take place across human and artificial agents. Reliability is normally attributed to machines but, as mentioned above, it should also refer to those who design them.

Engineers and computer scientists may claim that AI is reliable because the process is reliable. This reliability concerns some *technical* specifications of the artefact, for instance pertaining to its material characteristics but also pertaining to the design and development process that leads to it. Such processes do not happen in a vacuum, but are instead constituted by human and institutional decisions and actions. This may seem obvious, but there has been an enduring discourse in the sciences as well as philosophy of science that has neglected the role of human agents in aspects such as "reliability". Thus, we routinely talk about a model being reliable, neglecting that the reliability of a model in large part depends on how human agents have designed and developed the model. While this way of talking about scientific outputs (models, explanations, theories) is parsimonious, it has had the side effect of obfuscating responsibilities of

humans in techno-scientific processes. Making human actors more visible in these processes also serves to show that reliability are not "absolute" properties, and instead very much depend on who is using the machine, for what purpose, and under what conditions. So too are the designers and organisations that develop, support and disseminate it.

The definition of the ontic trust from a meta-ethical perspective can be a significant horizon of reference to resolve the moral, legal and political issues that arise from the thickening of the multiple and varied relationships in the network or, as Luciano Floridi (2017) writes, in the infosphere. Directly applying trust to technology risks overshadowing the anthropological characterisation of the trust form of relying on others/others as a good to be promoted, protected and valued, and implicitly promotes a blind faith in technical artefacts, obscuring the important responsibilities that humans still have. Since an ethical perspective unfolds through the entire span of human life, the anthropological structure of trust highlights the central role of responsibility of human beings in the course of their actions. A character that allows a shift of attention from the simple concentration on the deontological, moral and juridical dimensions of acting, to the valorisation and promotion of the care that the subject, natural and/or artificial, entertains with information, which entails understanding how information, precisely because it is not produced exclusively by the natural agent subject, implies the reliability. Yet, even if such a meta-ethical perspective is adopted, the legal question of who holds responsibility and accountability for the outputs of AI systems or their potential mis-use is not automatically solved.

This means considering the (human or artificial) subject that produces information as an ordered set of events or states of the world, and shaped by the informational relationships that constitute information societies. The interactions between agents, human and/or artificial, cannot take place within the horizon of normativity alone; such interactions, in turn, generate reliable information through the network of relations which forms of trust (a-c identified in section 1) take. Informational agents, human and artificial, are both producers and interlocutors of it, directly or mediate, they are involved in the trust in them and in the trust in "others" agents, stabilising plots of relations and interaction for the smooth functioning of the network according to a trust-logic such as to qualify widen the social space. This is an important point because, as artificial agents acquire increasingly more agency, autonomy, and also power in processing information, we should not forget that this does not put them automatically on a par with human agents. Arguments have been made that the modes of information processing, although the outputs may be quite similar, are very different (Floridi 2017), but our point here is normative: human agents still hold more responsibility than artificial agents (Russo 2022).

Therefore, given that it is a two-way relationship, trust is a meta-ethical concept that takes the normative form of placing constraints on the types of behaviour that both citizens and scientific experts are legitimised to enact in their informational interactions. Interpersonal trust, constituting the foundation of the social bond, as well as its capacity to generate trustworthiness of the trustee, then *analogically* extends to technology within a complex society in which humans and machines are in constant interaction. This is the only way the concept of trust can be given *some* sense from a legal perspective. It is then a question of understanding trust as an interpersonal act *and* technological reliability as an impersonal and analogical extension of this act capable of granting credibility to the artefact. As we said earlier, the categories of trust identified by Fabris are not mutually exclusive, and instead help us understand how, using an argument from analogy, and in the context of a network of relations, we can trust (the outputs of) AI systems. We agree here with arguments made in the debate requiring more presence of human beings in these processes (see e.g. argument made about trusting self-driving cars), and this holds too at the level of institutions that are supposed to provide standards and their enforcements. As much in the sphere of interpersonal trust as in that of technological reliance, free personal choice comes into play, the subject's capacity to live their relationships in view of an ethical end and their willingness to allow themselves to be transfigured by the transcendent, that is, by what lies beyond them.

6. Conclusion

We have examined the legal, technical, anthropological, and network dimensions of the debate on trust, humans and AI machines. There is a rich philosophical debate as to whether AI has the capacity of being a genuine object of trust, one of the main instances for this stance is based on the lack of human qualities such as intentionality, if this is missing, we should be prevented from considering such attributions. The extension to institutions (like in institutional trust) cannot be adduced as an argument that would make it possible to automatically extend the notion of trust to machines; after all, institutions are formed by humans, in a way that machines are not. Among relevant differences: humans are behind machines while in institutions, humans are the very essence of them.

Furthermore, these arguments about the meaning of trust are not purely theoretical. There is a real set of practical reasons to understand trust and AI. We can say that in the public law sphere, efforts should be focusing on the construction of a robust set of legal principles that could be used in the field of AI. More attention should be given to citizen awareness and

data-research on citizen acceptance of AI in Public Institutions. In a second phase and depending on how this principle-based framework would work, we could think of a more rule-based approach. Further, in the private law sphere, efforts should concentrate on analysing the extent to which the current responsibility, liability and damage law schemes in the EU and in the member states work to solve AI models' issues. New regulation on AI should focus on the problems posed by AI's reliability, rather than on how to enhance trust in AI models. There is awareness within the EU that certain challenges and concerns accompany AI deployment in general, particularly regarding safety, security and fundamental rights, and the relevance of this for our current discussion is that only humans could be accountable for.

Overall, by providing an interdisciplinary deep-dive on the subject of AI and trust, we hope to provide a more nuanced view on the matter. The outcome of this research has shown how it is possible to establish a transversal and fruitful dialogue between different points of view in order to offer a useful roadmap for reflection for our time. The theme of trust in AI, in fact, lends itself to an interdisciplinary analysis involving different spheres of humanities and social science research: only starting from the encounter and epistemological dialogue between knowledge can we inhabit and interpret the phenomenon of trust in AI within the era of change in which we live.

References

Afroogh, S.
2022 "A Probabilistic Theory of Trust Concerning Artificial Intelligence: Can Intelligent Robots Trust Humans?" in AI & ETHICS, vol. 2, 13-14.
Alici, L.
2012 *Fidarsi. All'origine del legame sociale*. Edizioni Meudon.
Alvarado V. *et al.*
2002 *Selection of EOR/IOR Opportunities based on Machine Learning*, conference paper SPE 78332 presented at European Petroleum Conference, Aberdeen, United Kingdom 29 – 31 October 2002.
Babushkina E. V. *et al.*
2013 *Forecasting IOE/EOR Potential Based on Reservoir Parameters*, presented at IOR 2013 – 17th European Symposium on Improved Oil Recovery held in St. Petesburg, Russia 16 – 18 April 2013.
Boissière, U.
2020 *Restaurer la confiance aujourd'hui*. Hermann.
Breiman, L.
2001 *Statistical modeling: The two cultures. Statistical Science*, 16(3), 199-231.
Buber M.
1923 *Ich und Du*, in *Werke, I, Schriften zur Philosophie*. Kösel e Lambert Schneider.

Buber M.,
1950 *Zwei Glaubensweisen. Mit einem Nachwort von David Flusser*. Lambert Schneider.
Caltagirone, C.
2020 (a cura di) *La fiducia generatrice di legami*, in 'studium Ricerca (Sezione on-line di Filosofia) 116 (4).
Cohen, M.
2023 *The Nature and Practice of Trust*. Routledge.
Doshi-Velez, F., & Kim, B.
2017 *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.
Estella de Noriega, A.
2023 "Trust in Artificial Intelligence Analysis of the European Commission proposal for a Regulation of Artificial Intelligence", in Indiana Journal of Global Studies, vol. 30, Issue 1, 2023.
Fabris, A.
2020 *Trust: a Philosophical Approach*. Springer.
Faulkner, P. & Sipmson, T.
2017 (a cura di). *The Philosophy of Trust*. OUP.
Floridi L.
2019 "Trusting artificial Intelligence in cybersecurity is a double-edge sword" in «Nature Machine Intelligence».
Floridi, L.
2017 *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*. Raffaello Cortina Editore.
Floridi, L.
2020 *Pensare l'infosfera. La filosofia come design concettuale*. Raffaello Cortina.
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … & Vayena, E.
2018 AI4 People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28, 689-707.
Gaillet, A., Perlo, N. & Schmitz, J.
2019 *La confiance: un dialogue interdisciplinaire*. Université Toulouse.
Gambetta, D.
1989 (ed.). *Le strategie della fiducia. Indagini sulla razionalità della cooperazione*. Einaudi.
Giddens, A.
1994 *Le conseguenze della modernità. Fiducia e rischio, sicurezza e pericolo*. Il Mulino.
Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L.
2018 *Explaining explanations: An overview of interpretability of machine learning*. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 80-89). IEEE.
Hardin, R.
2002 *Trust and Trustworthiness.*

Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A.
2025 *Bias recognition and mitigation strategies in artificial intelligence healthcare applications*. NPJ Digital Medicine, 8(1), 154.
Hastie, T., Tibshirani, R., & Friedman, J.
2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
"High-Level Summary of the AI Act." (2024) *EU Artificial Intelligence Act.*
HLEG AI (2019 *Ethics Guidelines for trustworthy AI. Retrieved from High-Level Expert Group on Artificial Intelligence.*
Hunyadi, V.
2020 *Au début est la confiance*. Editions Le Bord de l'eau.
Hwang, I.D.
2017 *Which type of trust matters? Interpersonal vs. Institutional vs. political trust*, Bank of Korea working Paper no. 2017-15. Available at SSRN: https://ssrn.com/abstract=2967051 or http://dx.doi.org/10.2139/ssrn.2976051.
Jagielski M. *et al.*
2021 "Manipulating machine learning: Poisoning Attacks and Countermeasures for Regression Learning", Cryptography and Security, retrieved from https://doi.org/10.48550/arXiv.1804.00308.
Laux, J.. Wachter S and Mittelstadt
2024 "Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk", in Regulation and Governance, 18, 3-32.
LeCun, Y., Bengio, Y., & Hinton, G.
2015 Deep learning. Nature, 521(7553), 436-444.
Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.
Luhmann, N.
2002 *La fiducia.* Il Mulino.
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
Marzano, M.
2014 *Avere fiducia. Perché è necessario credere negli altri*. Mondadori.
Metzinger, T.
2019 "EU Guidelines: Ethics washing made in Europe".
Mittelstadt B., Wachter S., Ch. Russell
2023 "To protect science, we must use LLMs as zero-shot translators". Nature Human Behaviour, 7, 1830-1832, November.
Natoli, S.
2016 *Il rischio di fidarsi*. Il Mulino.
Nedelsky, J.
2011 *Law's Relations. A Relational Theory of Self, Autonomy and Law*. OUP.
Nickel, P.
2010 "Can we Make Sense of the Notion of Trustworthy Technology". 29(3), pp. 429-444.[MOU39].

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
O'Neill O.
2020 *Trust and Accountability in a digital Age*, «Philosophy» 95(1), pp. 3 – 17.
Pelligra, V.
2007 *I paradossi della fiducia, Scelte razionali e dinamiche interpersonali*. Il Mulino.
Resta, E.
2009 *Le regole della fiducia*. Laterza.
Ribeiro, M. T., Singh, S., & Guestrin, C.
2016 "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 (pp. 1135-1144).
Rudin, C.
2019 *Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead*. Nature Machine Intelligence, 1(5), 206-215.
Russo, F.
2022 *Techno-scientific Practices: An Informational Approach*. Rowman & Littlefield.
Ryan M.
2020 *In AI We Trust: Ethics, Artificial Intelligence, and Reliability*, in 'science and Engineering Ethics», Springer.
Salih, A. M., & Wang, Y. (2024). Are Linear Regression Models White Box and Interpretable?. arXiv preprint arXiv:2407.12177.
Selbst, A. D., & Barocas, S.
2018 *The intuitive appeal of explainable machines*. Fordham Law Review, 87, 1085.
Taddeo M.
2011 "Defining trust and e-trust" in International Journal of Technology and Human Interaction, 23-35.
Totaro F.
1999 *Non di solo lavoro. Ontologia della persona ed etica del lavoro nel passaggio di civiltà,* Vita e Pensiero.
Vallier, K. & Weber, M.
2023 (eds.). *Social Trust. Foundational and philosophical issues*. Routledge.