Ethics, Politics & Society
A Journal in Moral and Political Philosophy

Trust, trustworthiness and the moral dimension in human-AI interactions narrative

ORCID: 0000-0001-8018-291
Received: 11/01/2025
Accepted: 16/01/2025

Donatella Donati
University of L'Aquila
donatella.donati@univaq.it

ABSTRACT The growing use of Autonomous Agents (AAs) in both private and public sectors raises crucial questions about trust. As AI systems take on increasingly complex tasks and decisions, their interactions with human agents (HAs) raise questions about the relevance and applicability of traditional philosophical concepts of trust and trustworthiness (sections 1 and 2). In this paper, I will explore different accounts of trust in AAs, arguing against both the complete dismissal of trust as misplaced (section 4) and the application of "genuine" trust frameworks (section 5). My aim is to lay the groundwork for the understanding that the moral complexity of interactions with AAs goes beyond the mere reliance we place on inanimate objects (section 6).
KEYWORDS Trust; trustworthiness; AI; trustworthy AI; ethics.

RESUMO A utilização crescente de agentes autónomos (AAs), tanto no sector privado como no público, levanta questões cruciais sobre a confiança. À medida que os sistemas de IA assumem tarefas e decisões cada vez mais complexas, as suas interacções com agentes humanos (AHs) levantam questões sobre a relevância e a aplicabilidade dos conceitos filosóficos tradicionais de confiança e fiabilidade (secções 1 e 2). No presente documento, explorarei diferentes relatos de confiança em AAs, argumentando contra a rejeição total da confiança como descabida (secção 4) e a aplicação de quadros de confiança "genuínos" (secção 5). O meu objectivo é lançar as bases para a compreensão de que a complexidade moral das interacções com os AAs ultrapassa a mera confiança que depositamos em objectos inanimados (secção 6).
PALAVRAS-CHAVE Confiança; fiabilidade; IA; IA fiável; ética.

1 Introduction

Artificial Intelligence (AI) systems – those systems implemented with some AI techniques - involve the simulation of human intelligence: these technologies are programmed to learn, make decisions, and perform tasks autonomously. By handling functions that have traditionally required human intelligence and actions—such as speech recognition,

problem-solving, and decision-making—these systems are increasingly able to carry out complex tasks with impressive efficiency and sophistication. These technologies are becoming an essential part of our everyday life: the adoption of AI-based systems across diverse sectors (mainly driven by their efficiency) is significantly increasing their interaction with humans, reshaping traditional notions of trust and the moral dimension involved in these relationships.

AI systems now play a central role in shaping how we work, communicate, and make decisions. These technologies, often classified as artificial agents (henceforth, AAs)[1], are capable of executing complex functions with minimal or no human supervision. As human agents (HAs), we are increasingly relying on these autonomous systems to carry out critical tasks and decisions across various domains. For example, virtual assistants help manage tasks and retrieve information, recommendation systems personalize entertainment, and AI-powered translation tools bridge language gaps. In the private sector, businesses use AI to optimize supply chains, enhance customer service through chatbots, and analyze data for decision-making. In the public sphere, AI supports functions like traffic management, healthcare diagnostics, and law enforcement.

As AI systems grow in complexity, so too does the nature of our interactions with them. Certainly, the relationship between (HAs) and AAs can be often seen as one of delegation—where tasks and decisions traditionally handled by humans are increasingly performed by AI systems. And wherever delegation takes place, it implies a certain form of trust from those delegating towards the entity or system entrusted with the task. As Fossa (2020, p. 66) nicely puts it:

> since AAs take an active part in the social organization of work, as humans do, it is easy to see the reason why trust may seem to be required.

This growing need for trust in AI has sparked a substantial body of literature and numerous initiatives aimed at building *trustworthy* AI systems, enhancing human trust in these technologies, and exploring the dynamics of the trust relationship that humans develop with AI. For example, the AI Act, a document that aims at establishing a legal framework for the development, market placement, and the use of AI systems

1 I will henceforth use the notions of AAs, AI and AI systems interchangeably.

in the European Union, is built around the notion of "trustworthy AI". As stated in the AI Act:

> The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human-centric and *trustworthy artificial intelligence (AI)* while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation." (AI Act, 2021, 1) (*emphasis mine*).

In addition to regulatory frameworks like the AI Act, major tech companies such as Google also use the concept of trustworthiness to describe their AI systems. For example, in the section on responsibility and safety within Google Cloud, the company emphasizes its commitment to trust and transparency in AI:

> The challenge is to do so in a way that is proportionately tailored to mitigate risks and promote reliable, robust, and *trustworthy* generative AI applications, while still enabling innovation and the promise of AI for societal benefit." (Google Cloud, 2024) (*emphasis mine*).

Google has devoted an entire document to the concept of trust in AI, titled *Google Cloud's Approach to Trust in Artificial Intelligence* (2023). This document further explores how trustworthiness underpins their approach to AI.

Also in the academic literature on computer science human-computer interaction (HCI) the concepts of trust and trustworthiness have become central. Many scholars have proposed definitions of trust, suggesting that these concepts can be defined as:

> the behavioral integrity of a system, which behaves as expected for all transactions" (Agca et al., 2022)

> As digital technologies evolve and become more refined and effective, our expectation has shifted toward trusting them—by delegating and refraining from supervision—in the execution of important tasks." (Taddeo, 2017, p. 565)

These are just a few examples that reflect the growing significance of the concepts of trust and trustworthiness in regulatory, industry and academic contexts. However, the emphasis on trust in and trustworthiness of AI is not limited to these examples: across various sectors and organizations, there is a widespread use of these concepts.

It should now be clear that recent advancements in AI have highlighted a growing focus on trust and trustworthiness in the context of artificial systems and their interactions with humans. And it is also crucial, also for the aim of this paper, that when considering trustworthy AI systems, it is essential to understand that this concept encompasses both trust in the outcomes they produce—such as their beneficence and technical robustness—and trust in the processes underlying those outcomes (including fairness, avoidance of bias, transparency, among other ethical values). Recognizing this dual focus is critical for addressing the full spectrum of moral and practical concerns related to trust. Traditionally, philosophical discussions on trust have been confined to "human-to-human relationships" (the complex interactions we now have with AAs are a relatively recent phenomenon). In recent years, however, some philosophers have begun to explore these new forms of interaction focusing on trust, and within the philosophical literature, the treatment of trust in HA-AA relationships tends to follow two opposing approaches:

i. One approach argues that the concept of trust does not apply to interactions with artificial agents. Instead, it posits that reliance is the appropriate concept, since trust can only exist between human agents—beings capable of moral reasoning (as discussed in section 4).

ii. A second, less common and more recent perspective attempts to adapt classical trust theories to account for relationships involving AI, extending them to include the specific type of trust that humans place in AAs (as explored in section 5).

Both of these approaches aim to make sense of trust and to bridge the gap between traditional views of trust in the context of human-machine interactions. However, they share a key characteristic: both tend to eliminate the moral dimension from these relationships (as discussed sections 4, 5, and 6). In the last section (section 6), I will attempt to sketch an alternative perspective—one that acknowledges and incorporates the moral dimension of our interactions with AAs, as I believe that this aspect should not be overlooked. While I will leave this question open for further exploration, I believe this direction holds promise for better understanding the ethical and epistemological complexities at play.

2 Philosophical orthodoxy on trust and trustworthiness relativism

MacIntyre In this section, I will outline various conceptions of trust as discussed in the philosophical literature, emphasizing how these analyses predominantly center on HAs[2]. Within the broader philosophical tradition, three primary concepts related to trust can be identified.

First, we encounter the general notion of trust, which can be represented by a two-place predicate: *x trusts y* (cf. Faulkner, 2015, p. 16). For example, Amina (x) trusts Fatima (y).

The second notion is contractual trust, where an agent, x, trusts another agent, y, to perform a specific action, *z : x trusts y to z*. Contractual trust is expressed as a three-place predicate, linking two agents to an action. For example, Amina (*x*) trusts Fatima (*y*) to drive her kids to school (*z*).

The third key concept related to trust is trustworthiness. Trustworthiness is a one-place predicate, describing a property that an individual instantiates. For instance, Amina is trustworthy. Trustworthiness is considered an "epistemic virtue" that Amina exemplifies. A definition of trustworthiness is given by Jones (2012, pp. 70-2):

> B is trustworthy with respect to A in a domain of interaction D, if and only if she competent with respect to that domain, and she would take the fact that A is counting on her, were A to do so in this domain, to be a compelling reason for acting as counted on. (…) To

2 As is also described in Tallant and Donati 2020.

> be trustworthy with respect to A in D thus requires that B be capable of recognizing that A is counting on her and, roughly, what they are counting on her for.

These notions—trust, contractual trust, and trustworthiness—are often distinct from one another, a distinction that is widely shared in the classical philosophical literature[3]. To illustrate this, consider the following example: we may not trust a particular individual or company in general (lack of "general" trust), yet we might still trust them to accomplish a specific task (contractual trust). For example, you may not trust a restaurant because of previous bad experiences, but you still trust them to prepare your favorite dish correctly when you order it for a special occasion.

Another important distinction in the philosophical literature is between mere reliance and a deeper, morally loaded concept of reliance, which is indeed understood as trust. Building on Hawley's analysis, this distinction highlights the difference between simply depending on someone or something to act in a certain way and genuinely trusting them, which involves normative expectations and ethical implications. Here's Hawley (2014, p. 2) on the distinction:

> we often rely upon inanimate objects but we do not grant them the rich trust we sometimes grant one another; inanimate objects can be reliable but not genuinely trustworthy. Moreover our reactions to misplaced trust differ from our reactions to misplaced reliance. Suppose I trust you to look after a precious glass vase, yet you carelessly break it. I may feel betrayed and angry; recriminations will be in order; I may demand an apology. Suppose instead that I rely on a shelf to support the vase, yet the shelf collapses, breaking the vase. I will be disappointed, perhaps upset, but it would be inappropriate to feel betrayed by the shelf, or to demand an apology from it..

The key distinction between trustworthiness and reliance lies in the presence of some moral commitment. We hold an attitude of mere reliance, rather than trust, when dealing with inanimate objects: for example, we rely on a ladder to hold our weight or a shelf to support a stack of books. In contrast, our attitude shifts to a morally richer notion of trust

3 Cf. Faulkner, 2015.

when it involves human agents.[4] For instance, Maria trusts Amina to look after her dog—a scenario that entails expectations of care, responsibility, ecc.

To conclude, classical analyses of trust have traditionally focused on human-to-human relationships. However, with the evolving nature of our societal interactions, it has become increasingly important to explore how these concepts might apply to or need to be adapted for human-AI relationships. As mentioned earlier in section 1, these notions are already being used in various contexts, and philosophers, too, have begun investigating the nature of trust in the relationships between HAs and AAs.

3 Defining artificial agents

Before turning the discussion to the concepts of trust in AAs and trustworthy AI, it is necessary to clarify the specific type of artificial agents I will focus on. For the purposes of this paper, I will restrict the domain to those agents discussed in Section 1, so to autonomous systems that are implemented with some AI techniques, and that exhibit intelligent behavior in decision-making and task execution without the need of human supervision and intervention. These are the kinds of agents that typically perform tasks traditionally undertaken by HAs, but where the consequences of these actions or decisions have significant implications for human lives (i.e. harms and benefits) and so can be morally loaded. In other words, I am concerned with agents exhibiting "intelligent autonomy" at a sophisticated level. To better illustrate this, I will provide a few examples.

Many technologies can be described as autonomous, including thermostats, landmines, and autonomous vehicles. While all of these devices are capable of activating themselves, the type and degree of autonomy they exhibit differ significantly. Indeed, autonomy is not a binary property but a gradable one (cf. Wheeler, 2019, p. 345). For instance, the autonomy of a thermostat is different from that of an autonomous vehicle. A thermostat can be considered a "rudimentary" technology in com-

4 Even though mere reliance can still exist in interactions with human agents, the philosophical canon generally holds that morally loaded interactions are typically seen as possible only between human agents.

parison to the complexity of an autonomous vehicle and its autonomy too. Although a thermostat activates and performs tasks autonomously, its operation remains relatively simple: when the temperature drops, a sensor triggers the mechanism to switch the heating on. In contrast, an autonomous vehicle performs far more complex and critical tasks, such as safely driving your children to school, determining the most efficient route to take, and making split-second decisions to navigate dynamic road conditions. The scope and stakes of these tasks highlight the profound differences in the levels of autonomy between these two technologies—and the impacts they have on human lives: a malfunctioning thermostat may result in an uncomfortable temperature in the house, but the consequences are usually limited to discomfort and are unlikely to have significant moral effects. In contrast, a malfunctioning autonomous vehicle could cause significant harm: potentially leading to accidents, injuries, or even loss of life, due to the vehicle's role in navigating complex and dynamic environments.

I focus on this second type of autonomous systems, often referred to as artificial agents (AAs)- as anticipated in section 1. These systems are particularly interesting because they invite us to explore the relationships humans establish with them, and particularly the notion of trust. Unlike simpler autonomous systems, AAs are capable of making decisions and acting independently, without human supervision or intervention, by evaluating different situations and in nuanced manner. Indeed, AAs interact with HAs in complex ways, raising key ethical and practical concerns. More examples of such systems include recommender algorithms, facial recognition technologies, autonomous vehicles, assistive robots, robot surgeons, and autonomous weapons. The reason why these systems are far more interesting than the simpler ones is because they have the potential to generate both benefits and harms to humans that, clearly, that carry significant moral implications. For instance, autonomous weapons may have the power to decide whether to take a human life (Sharkey 2019), while robot surgeons could determine whether to proceed with a critical surgery (Formosa et al. 2022), systems like facial recognition can perpetuate bias, especially against minority groups, (Gebru 2020, Buolamwini and Gebru 2018) all raising serious ethical concerns. The literature across ethics, computer science, and human-computer interaction (HCI) is filled with discussions about such examples.

As mentioned above, the level of autonomy and sophistication of the systems just mentioned clearly varies. While recommender systems are highly complex, they do not match the sophistication of an assistive robot, which performs a wide range of tasks and interacts with people in ways that closely resemble human-to-human interaction. For instance, an assistive robot, not only provides recommendations but also helps, makes decisions, and supports an elderly person in their daily activities, and clearly, its moral impact on the person it assists is far greater than the impact a thermostat, for example, could have.[5] The importance of trust and trustworthiness our interactions with these systems is undeniably central. Building, maintaining, and analyzing trust in such technologies presents unique challenges. Clearly, the interactions with these types of AAs are significantly more complex than those with simpler systems, as the potential implications of their behaviors on our lives are far more significant. Again, these AAs operate in contexts where the implications of trust touch on critical areas like safety and human well-being more in general.

4 What are the implications of philosophical orthodoxy for trust in AAs?

Although the philosophical literature has traditionally concentrated on trust and trustworthiness within the realm of human-to-human interactions, as outlined in the classical analysis discussed in Section 2, recent years have seen some philosophers turn their attention to these concepts in the context of artificial intelligence. In this section, I will examine accounts that argue that trust and trustworthiness are not the appropriate concepts to apply to our interactions with AAs. These accounts constitute the prevailing view in this area of philosophy, and are closely aligned with classical philosophical accounts of trust and trustworthiness.

As outlined in section 2, trust and trustworthiness go beyond mere reliance. While reliance may lead to disappointment if expectations are not met (I might feel disappointed if the shelf fails to hold the weight of my pile of books), trust introduces a deeper vulnerability—it carries the potential for betrayal or a profound sense of being let down (I would

5 Cf., e.g., Formosa, 2021.

feel betrayed by my best friend if I trusted her to look after my dog while I'm on vacation and she doesn't).

Building on this distinction, the primary argument of those who reject the application of trust and trustworthiness to human-AI (or HAs-AAs) interactions is that these concepts are fundamentally inapplicable in this context. They contend that trust is not an attitude humans can genuinely adopt toward artificial agents, nor is trustworthiness a property that an artificial agent can truly possess. Trust can only exist between HAs (i.e. moral agents), and trustworthiness is a quality that is exclusive to HAs. Unlike "genuine" trust, mere reliance is a pragmatic stance—an action based on the expectation or probability that a given technology will perform as intended. As Freiman (2023, p. 1351) puts it:

> Overall, the attitude of trust entails an expectation for the trustee to fulfill their commitments and be aware that they are trusted. (…) a trustworthy agent, therefore, has the power to betray the trustor. Can (trustworthy) AIs betray humans? The field of social epistemology is infused with anthropocentric concepts (…) and in everyday language we associate these concepts with non-humans.

Indeed, most philosophers critique the concept of 'Trustworthy AI,' arguing that we should avoid anthropomorphizing such systems. Several criticisms of 'Trustworthy AI' include, but are not limited to the following (I will quote at length, to better clarify what I have just stated):

> Trust is a relationship between peers in which the trusting party, while not knowing for certain what the trusted party will do, believes any promises being made. Artificial Intelligence (AI) is a set of system development techniques that allow machines to compute actions or knowledge from a set of data. Only other software development techniques can be peers with AI, and since these do not "trust", no one actually can trust AI. (Bryson, 2020, p. 1)

> AI is simply a set of system development techniques and therefore does not qualify as a 'peer', only other software development techniques can be peers with AI, and since these do not have the capacity to trust, no one actually can trust AI. (…) talking about trust in AI instead of reliability, and about trustworthy AI instead of reliable

> or accountable AI may have serious consequences. (Sutrop, 2019, pp. 511-2)

> (…) one needs to either change 'trustworthy AI' to 'reliable AI' or remove it altogether. The rational account of reliability does not require AI to have emotion towards the trustor (affective account) or be responsible for its actions (normative account). (Ryan, 2020, p. 17)

Also, Fossa's (2020) account conveys this prevailing idea in the philosophical literature. He argues that trust and trustworthiness are not the appropriate concepts to apply in the context of AI. Fossa argues that there are significant differences between delegating tasks to AAs and delegating tasks to HAs. He suggests that the behavior of an AA merely imitates that of an HA. As Fossa puts it: "Deciding to frame HA→AA task delegation by reference to trust implies that, in such a relationship, something more is at stake that cannot be accounted for solely by the functional notion of reliance. What is this additional element?" (Fossa, 2020, p. 70). He argues that trust is specifically needed in human-to-human task delegation because humans are believed to have the autonomy to make decisions and choose their own goals. This autonomy brings with it moral expectations, allowing the person who trusts to hold the other accountable for any breaches, demand justifications, and even feel offended by a betrayal of trust. Fossa then argues that it is impossible to consider AAs as entities capable of independently choosing between competing objectives and betraying trust, since they lack an intrinsic connection to purposes. In contrast to humans, who are beings capable of setting their own purposes, AAs are "purpose-built artifacts" (Bryson & Kime, 2011), created for specific tasks. Their objectives are always defined by their designers, and any failure of an AA should not be perceived as a betrayal but rather as a disappointment in terms of functional performance. Therefore, Fossa argues that the relationship between humans and artifacts is generally one of reliance, not trust.

In conclusion, many of the philosophers discussed above argue that anthropomorphizing AI systems is problematic, defending the idea that the relationship between HAs and AAs should be based on reliance rather than trust. They warn against attributing human-like qualities to AI, particularly in terms of moral status or responsibility, as doing so could result in the misattribution of accountability (cf. Fossa, 2020). AAs, on these views, are akin to mere inanimate objects. To conclude

this section, I would like to highlight this insightful quote by Freiman, who, in exploring the theoretical foundations of trust, states:

> The field's roots are found within traditional Anglo-American analytic philosophy. Within social epistemology, the standard view on trust is that trust relations are based on human quality such as goodwill. Therefore, trust relations are only possible between individual persons, however, on a generous interpretation, they involve groups. This view rests upon a commonly acknowledged distinction between a genuine trust and mere reliance. (…) Unlike genuine trust, mere reliance is a way of acting in light of the probability that technology will perform successfully. Genuine trust entails a moral aspect, that mere reliance does not. (2023, p. 1353)

5 An alternative perspective: making sense of trust in AAs

Few philosophers, while drawing on classical accounts of trust, argue that it can be meaningful to apply the concepts of trust and trustworthiness to AI systems. In this section, I will outline two perspectives on trust in Artificial Agents (AAs) put forward by Simion and Kelp (2023) and Zanotti et al. (2023).

First, Simion and Kelp argue against the anthropocentric distinctions between trust and reliance in existing literature and proposes a new account of trustworthy AI that offers a unified rationale for generating context-specific, objective frameworks.

Their work represents an attempt to make sense of trust in AI from a philosophical perspective, offering a systematic account of AI trustworthiness that departs from traditional anthropocentric approaches. They propose that trustworthiness should be understood as a disposition to fulfill functionally sourced obligations. The central claim is that trustworthiness is a matter of how closely an AI approximates maximal trustworthiness, which is defined as having a maximally strong disposition to meet its obligations. Degrees of trustworthiness, in turn, depend on the AI's proximity to this ideal.

Unlike accounts of trust that rely on human-like traits, such as will or character, this framework anchors trustworthiness in the norms governing proper AI functioning, which is determined by adherence

to either *d-functional norms* (aligned with the designer's intentions) or *e-functional norms* (aligned with reliable performance that sustains the artifact's utility).

This approach also provides a philosophical background to address the limitations of "list-based" proposals for trustworthy AI[6]. Simion and Kelp argue that trustworthy-making properties are those that correspond to an AI's disposition to fulfill its functionally sourced obligations: whether a specific property is salient in a given context depends on the type of AI and the practical demands of its application. For example, explainability may be crucial for a credit-scoring AI, as it enables users to understand why a mortgage was rejected, but it might not be essential for a diagnostic AI when patients cannot meaningfully interpret complex medical explanations.

By emphasizing contextual thresholds for trustworthiness, the framework accommodates two key dimensions: breadth (the range of obligations the AI fulfills) and depth (the strength of its disposition to fulfill those obligations). It also distinguishes between attributive ascriptions of trustworthiness (e.g., "Ann is a trustworthy physician") and predicative ascriptions (e.g., "Ann is trustworthy"), with context shaping the obligations relevant to each case.

Crucially, this account avoids anthropocentric assumptions and does not require AIs to have human-like psychological traits to qualify as trustworthy. Instead, it generalizes the concept of trustworthiness to include artificial systems by grounding it in functional norms. This non-anthropocentric view aims at explaining variations in trustworthiness requirements across different AI systems and contexts and provides a robust framework for making sense of trust in AI. On Simion and Kelp's view, by grounding trustworthiness in the fulfillment of functionally sourced obligations is possible to offer a unified, context-sensitive framework that clarifies the normative foundations of trustworthy AI.

The second approach is put forward by Zanotti et al. (2023). They argue that, when designing and regulating AI systems, the focus should not be limited to their technical performance (such as robustness and accuracy) but should also encompass their ethical behavior, including, for example, properties such as fairness and transparency. On their view,

6 On their view, the several proposals on trustworthy AI list features that supposedly make AIs trustworthy (such as safety, fairness, and transparency) face two main issues: they lack explanatory adequacy, as they fail to clarify why specific properties are included, while the second has to do with the distinction between trustworthiness and mere reliability (cf. Simion and Kelp, 2023, p. 2).

relying solely on a concept like "reliance" is insufficient to capture the full nature of trust in AI systems. Zanotti et al. propose that the concept of Trustworthy AI (TAI) is better suited to address these concerns, as it integrates both reliability and ethical considerations. Indeed, Zanotti et al. argue against existing perspectives on TAI, particularly motivational and purely epistemic approaches. Motivational accounts, which focus only on reliability, overlook the need for a more comprehensive understanding of TAI, while epistemic views fail to incorporate crucial ethical aspects, such as fairness and respect for human autonomy. As they put it:

> To sum up, we have identified three elements of trust and trustworthiness that are common to H – H [human-human] and H – AI [human-AI] interactions: (i) reliability is the basis for trust; however, (ii) reliability is not enough, for the notion of trust is also grounded in an ethical dimension; finally, (iii) trust and trustworthiness provides us with a nonepistemic guarantee in contexts of vulnerability and risk. Identifying these elements allows us to maintain that H – H and H – AI trust are two distinct notions that nonetheless share a conceptual core and motivates our use of *trust* – and not some other notion – in applications to AI systems. (2024, p. 2699).

I believe this approach is headed in the right direction, as it not only recognizes the importance of proper functioning in ensuring trustworthiness but also underscores the need for consideration of the moral dimension.

6 Is this enough?

While I find the latter account outlined in the previous section promising and heading in the right direction, I believe there is an even stronger reason to make sense of trust in AI. I will illustrate this with an example.

Autonomous vehicles (AVs) offer a powerful example of how trust in AAs can raise complex *moral* questions. While people are used to trusting machines like elevators or washing machines, which have predictable, routine functions, AVs are designed to make decisions that directly affect human lives. The key moral distinction is that AVs, like other simi-

lar AAs, can make decisions that involve moral judgments under uncertainty, unlike inanimate objects which follow pre-set instructions or mechanisms without "subjective" decision-making. Although extreme, one classic example of a moral dilemma that AI systems (like AVs) may face is a revised version the trolley problem[7]: very roughly, if the autonomous vehicle faces an unavoidable collision, should it prioritize the safety of its passengers or the pedestrians in its path? What if the AV must choose between killing a child or an elderly person? The trust that humans place in AVs goes beyond expecting a safe trip to the store—it involves trusting that the AI will make morally sound decisions that reflect societal values about life, fairness, and harm reduction. And this matters because autonomous vehicles are tasked with making moral decisions that, traditionally, have been made by humans. This means that developers must embed ethical principles in AI decision-making algorithms. When humans trust AVs, they are implicitly trusting these systems to perform ethically, even in complex and life-and-death situations. But it is not only about the ethical principles implemented in the AAs—an important aspect to consider when it comes to our interactions with such systems is the psychology behind the trustee (i.e. the HA).

The process of building, maintaining, and analyzing trust in such systems makes the human relationship with these technologies particularly challenging. And following Zanotti et al., I believe the moral dimension must not be excluded—whether by dismissing AAs from the moral sphere or by attempting to make this moral aspect unnecessary through a unifying account of trust, such as those proposed by Simion and Kelp.

A decision made by a sophisticated AA—such as a driverless car that "chooses" to swerve and run over pedestrian A rather than pedestrian B, or a drone deciding whether or not to attack a target—has a significant moral impact[8] that cannot be equated with the failure of a bookshelf or the malfunction of a thermostat. Relying on a driverless car to safely

7 Awad et al. (2018) and their Moral Machine Experiment (an online platform to explore public opinions on ethical dilemmas faced by autonomous vehicles, particularly in life-and-death scenarios. It collects large-scale data on how people prioritize the lives of different individuals or groups in hypothetical accident situations). https://www.moralmachine.net

8 One might argue that the activation of a landmine, which is akin to an ordinary inanimate object, carries a huge moral impact. This is true. However, unlike a landmine, an autonomous weapon ***actively assesses*** a situation and ***makes a decision*** about whether to engage a target—an inherently more complex process that also involves uncertainty about the behavior of the AA.

transport you is fundamentally different from placing a book on a shelf and trusting that the shelf will hold it.

The *expectations* and *potential disappointment* we experience with AI systems are fundamentally different from those we have toward inanimate objects.[9] Returning to the examples mentioned above, a malfunctioning thermostat may only cause minor discomfort, while an autonomous vehicle that fails to safely navigate traffic could lead to accidents and serious harm, raising significant moral concerns. Similarly, if a bookshelf fails to hold the weight of a pile of books, the disappointment comes from the expectation that the shelf, as an inanimate object, would perform its basic function. However, the moral dimension is minimal, as the bookshelf is not considered capable of intentional action, whereas an autonomous weapon system assessing a situation and making the wrong decision about a target could result in catastrophic consequences. These examples illustrate that our relationship with AI system entails a moral dimension that goes beyond mere reliance on inanimate objects and their proper functioning in different contexts, and where the consequences of failure are far more significant and morally charged.

What I intend to convey is that while it may be a mistake to treat autonomous systems as moral agents, it is equally flawed to consider them as mere inanimate objects. Although difficult to define precisely, trust appears to operate on a spectrum, with varying degrees depending on the system involved. In this regard, AAs likely occupy a space between human agents and inanimate objects. Both positions—the view that trust in AAs is misguided and should be understood only as reliance, and the view that seeks to justify trust in AAs—tend to conceptualize AAs as objects that function without acknowledging the moral dimension of the interaction, which appears to be a critical aspect of the relationship. The first kind of views "declassify" human-AI interactions by equating them to interactions with inanimate objects, suggesting that only the concept of reliance, not of trust, is applicable. In contrast, the second view risks creating an overly broad definition of trust, one that includes relationships where the moral dimension is absent, thereby neglecting the ethical significance present in certain human-AA interactions.

In real-world contexts such as the roads we travel on, healthcare, and military applications, trusting autonomous agents goes beyond

9 Cf., among others, Grimes et al,. 2021, Viik, 2020.

expecting reliability and mere functionality—it involves trusting them with the power to make morally complex decisions. This task is further complicated by the significant variations among AAs in terms of their sophistication and the contexts in which they operate. For example, healthcare contexts present unique challenges, as they require AAs to navigate ethical dilemmas involving patient autonomy, human dignity, beneficence, and so on. Additionally, the dynamic interplay between the identity and intentions of designers, the utility and context of an AI systems' development or application, and the cultural models of users all profoundly influence how trust is established and maintained. These factors highlight the multifaceted nature of our relationships with AAs and the importance of a nuanced, context-sensitive approach to understanding trust in these systems. Trusting AAs, therefore, involves not only ensuring their technical competence but also aligning their decision-making with the moral values of society and the individual expectations of users— tasks that demand a greater effort than is typical for traditional machines.

References

Agca, M. A., Faye, S., & Khadraoui, D. (2022). A survey on trusted distributed artificial intelligence. *IEEE Access*. Retrieved from https://www.nist.gov/system/files/documents/2022/11/16/Muhammed%20Akif%20AĞCA%202.pdf

Artificial Intelligence Act. (2024). European Parliament legislative resolution of 13 March 2024.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A. J.-F., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bryson, J., & Kime, J. J. (2011). Just an artifact: Why machines are perceived as moral agents. Retrieved from https://www.cs.bath.ac.uk/~jjb/ftp/BrysonKime-IJCAI11.pdf

Bryson, J. (2018). AI & global governance: No one should trust AI. United Nations.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 77–91. Retrieved from https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Faulkner, P. (2015). The attitude of trust is basic. *Analysis, 75*(4), 424–9.

Formosa, P. (2021). Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. *Minds & Machines, 31*, 595–610. https://doi.org/10.1007/s11023-021-09579-2

Formosa, P., Rogers, W., Griep, Y., Bankins, S., & Richards, D. (2022). Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. *Computers in Human Behaviour, 133*. https://doi.org/10.1016/j.chb.2022.107296

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

Freiman, O. (2023). Making sense of the conceptual nonsense 'trustworthy AI'. *AI and Ethics, 3*, 1351–60.

Gebru, T. (2020). Race and gender. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI. Oxford University Press.* https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Google Cloud. (2023). Google Cloud's approach to trust in artificial intelligence (Kaganovich, M., Kanungo, R., & Hanssen, H.).
Retrieved from https://services.google.com/fh/files/misc/ociso_securing_ai_governance.pdf

Grimes, M. G., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems, 144.*
https://doi.org/10.1016/j.dss.2021.113515

Hawley, K. (2014a). Trust, distrust and commitment. *Noûs, 48*(1), 1–20.

Hawley, K. (2014b). Partiality and prejudice in trusting. *Synthese, 191*, 2029–45.
https://doi.org/10.1007/s11948-020-00228-y

Jones, J. (2012). Trustworthiness. *Ethics, 123*, 61–85.

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

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

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*, 2749–67.

Sharkey, A. (2019). Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology, 21*, 75–87. https://doi.org/10.1007/s10676-018-9494-0

Simion, M., & Kelp, C. (2023). Trustworthy artificial intelligence. *AJPH, 2*, 8.
https://doi.org/10.1007/s44204-023-00063-5

Sutrop, M. (2019). Should we trust artificial intelligence? *Trames Journal of the Humanities and Social Sciences, 23*(4), 49.

Taddeo, M. (2009). Defining trust and E-trust: From old theories to new problems. *International Journal of Technology and Human Interaction, 5*(2), 23–35.
https://doi.org/10.4018/jthi.2009040102

Taddeo, M. (2017). Trusting digital technologies correctly. *Minds & Machines, 27*, 565–568.
https://doi.org/10.1007/s11023-017-9450-5

Tallant, J. (2019). You can trust the ladder, but you shouldn't. *Theoria, 85*, 102–18.

Tallant, J. (2022). Trusting what ought to happen. *Erkenntnis.*
https://doi.org/10.1007/s10670-022-00608-9

Tallant, J., & Donati, D. (2020). Trust: From the philosophical to the commercial. *Philosophy of Management, 19*, 3–19. https://doi.org/10.1007/s40926-ok019-00107-y

Viik, T. (2020). Falling in love with robots: A phenomenological study of experiencing technological alterities. *Paladyn, Journal of Behavioral Robotics, 11*(1), 52–65.
https://doi.org/10.1515/pjbr-2020-0005

Wheeler, M. (2019). Autonomy. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI.* Oxford University Press.

Zanotti, G., Petrolo, M., & Chiffi, D. (2023). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society.* https://doi.org/10.1007/s00146-023-01789-9