A Formal Account of AI Trustworthiness: Connecting Intrinsic and Perceived Trustworthiness

Piercosma Bisconti[1], Letizia Aquilino[2], Antonella Marchetti[2], Daniele Nardi[4]

[1]Consorzio Interuniversitario Nazionale per l'Informatica, AIIS Lab
[2] Università Cattolica del Sacro Cuore, CeRiToM - UniToM - Department of Psychology
[4]Sapienza University of Rome, Dipartimento di Ingegneria Informatica, Automatica e Gestionale "A. Ruberti"
piercosma.bisconti@dexai.eu, letizia.aquilino@unicatt.it, antonella.marchetti@unicatt.it, nardi@diag.uniroma1it

Abstract

This paper proposes a formal account of AI trustworthiness, connecting both intrinsic and perceived trustworthiness in an operational schematization. We argue that trustworthiness extends beyond the inherent capabilities of an AI system to include significant influences from observers' perceptions, such as perceived transparency, agency locus, and human oversight. While the concept of perceived trustworthiness is discussed in the literature, few attempts have been made to connect it with the intrinsic trustworthiness of AI systems. Our analysis introduces a novel schematization to quantify trustworthiness by assessing the discrepancies between expected and observed behaviors and how these affect perceived uncertainty and trust. The paper provides a formalization for measuring trustworthiness, taking into account both perceived and intrinsic characteristics. By detailing the factors that influence trust, this study aims to foster more ethical and widely accepted AI technologies, ensuring they meet both functional and ethical criteria.

Introduction

In contemporary discourse on artificial intelligence (AI), the concept of trustworthiness is paramount, especially as these technologies become more embedded in societal structures. Trustworthiness in AI refers to the assurance that AI systems will operate in a safe, reliable, and fair manner. The importance of trustworthiness stems from the need to mitigate risks associated with AI deployment, ensuring that these systems do not cause harm, and that their actions are understandable and predictable to humans.

Arguably, AI systems are distinguished from other technologies by their autonomy and adaptability. These systems are designed to make decisions and adapt to new information without human intervention, which can lead to complex interactions that are difficult to predict. Therefore, the trustworthiness of AI is critical not only for ensuring safety and compliance with ethical standards, but also for maintaining the integrity of the decisions made by these systems. As AI technologies become more autonomous, the potential for them to act in unforeseen ways increases, thereby amplifying the need for trustworthiness.

The European Union (EU) has recognized the critical importance of trustworthiness in AI systems. This recognition is evident in legislative and regulatory frameworks such as the AI Act, which seeks to govern the use of AI by establishing clear rules for its development and application. Moreover, the EU's Architecture of Standards of the European Committees for Standardization (CEN) and for Electrotechnical Standardization (CENELEC) specifically emphasizes the need to develop standards that ensure the trustworthiness of AI systems, notably for allowing the presumption of conformity with the EU AI Act. These initiatives aim to create a regulated environment where AI's benefits are maximized, and its risks are minimized, thereby fostering innovation, while protecting citizens' rights.

Previous literature has studied the conceptualization of trust in artificial intelligence (AI) drawing from theories initially developed to understand trust in human-human relationships. One traditional understanding of trust in interpersonal interactions encompassed three key components: ability, benevolence, and integrity (Mayer et al., 1995). Ability refers to the perceived skills and competence of the trustee in fulfilling a given task; benevolence indicates positive intentions not solely driven by self-interest; while integrity underscores the moral and just conduct of the trustee, ensuring consistency, predictability, and honesty. Initially, researchers attempted to apply this framework directly to assess trust in technology. However, debates arose regarding the feasibility of directly applying these components to human-to-machine trust due to the fundamental differences between humans and technology, particularly the lack of (perceived and actual) voluntariness and moral agency in technology.

McKnight and colleagues (2011) proposed an adapted framework focusing on three dimensions of trust in technology: functionality, reliability, and helpfulness. This approach acknowledged the unique attributes of technology, emphasizing its capability, consistency of operation, and usefulness to users. Some studies bridged the gap between human-trust and technology-trust, suggesting a conceptual framework centred on 'trust in automation' (Lee & See, 2004). This framework delineated trust across performance, process, and purpose categories, capturing both operational characteristics and users' goals and expectations.

Trustworthiness is also a crucial driver for social acceptance and societal adoption of AI (Shin, 2021; Shin et al., 2020; Söllner et al., 2016; Wu et al., 2011). Public trust in AI systems is essential for their integration into everyday life. If individuals and institutions believe that AI systems are reliable and fair, they are more likely to support and adopt these technologies. Conversely, a lack of trust can lead to resistance and hinder the potential benefits that AI might bring to society. Therefore, building trustworthy AI systems is not merely a regulatory requirement, but a strategic move to enhance public adoption and acceptance.

Despite the discussions highlighted above, the concept of trustworthiness remains loosely defined today. Many, including the CEN-CENELEC in response to the EC standardization request, view trustworthiness as the overarching framework for AI systems, encompassing both regulatory and ethical dimensions. There are two primary issues concerning this concept:

1) Trustworthiness, as an overarching framework, encompasses various other characteristics that AI systems need to possess, including robustness, oversight, accuracy, transparency, etc. However, no clear taxonomy has yet achieved the consensus required for universal acceptance (Graziani et al., 2023; Kaur et al., 2022). Furthermore, the definitions of these characteristics themselves are also debated. For example, the concept of transparency is ambiguous (Felzmann et al., 2020): it is unclear whether it refers to informing stakeholders about the system's goals and functioning, or if it relates to the system not being a "black box."

2) The concept of trustworthiness is often associated only with the technical characteristics of the system, such as robustness and accuracy. Conversely, the notion of perceived trustworthiness is also considered crucial (Schlicker & Langer, 2021). Human perceptions of AI systems are indeed of significant importance for ensuring social acceptance, adoption, and proper use of this technology.

However, the interplay between these two understandings of trustworthiness has not been theoretically worked out (Stanton & Jensen, 2021).

What Does Trustworthiness Refer to?

In this paper, we will concentrate on the second issue outlined above, namely our objective is to articulate a definition of trustworthiness that encompasses both the intrinsic (technical) characteristics of AI systems and the characteristics perceived by an observer.

Initially, it is crucial to acknowledge that trustworthiness is not solely an intrinsic property of a system, but is fundamentally relational. In fact, it requires the interaction between at least two parties: a trustor and a trustee (Jones & Shah, 2016). This perspective shifts the understanding of trustworthiness from being a static system characteristic to a dynamic, interactional attribute.

This relational aspect indicates that trustworthiness covers scenarios where certain processes of the system—or its interactions with the environment—are not completely predictable or controllable. Essentially, if a system were fully predictable and manageable, trust would be unnecessary. Therefore, trust specifically arises from the lack of complete knowledge or control, which, if eliminated, would make the concept of trust redundant.

Consequently, our goal is to refine the definition of trustworthiness to more accurately reflect both the technical capabilities of AI systems and the subjective experiences of those who interact with them. This comprehensive approach ensures that trustworthiness assessments are based on a full understanding of both objective system functionality and human perceptions, which are crucial for societal acceptance and the effective integration of AI technologies.

Building on this discussion, we first consider the ISO definition of Trustworthiness (ISO/IEC TR 24028:2020), which states:

"Trustworthiness is the ability to meet stakeholder expectations in a verifiable manner."

A note to this definition clarifies that, while trustworthiness is described as an ability, it is more aptly an attribute. In agreement with our earlier arguments, it is indeed more precise to classify it as an attribute. Specifically, an ability suggests a property that the system possesses and can be evaluated independently; it does not inherently require interaction. In

contrast, the term “attribute” implies a quality assigned by an observer, thereby emphasising its interactional nature. Thus, we propose a paraphrase of the ISO definition:

"Trustworthiness is an attribute, outcome of meeting stakeholders’ expectations".

This suggests that when stakeholders’ expectations are met, they attribute trustworthiness to the system. Based on this refined definition, we propose the following theoretical model:

a) trustworthiness is attributed when the AI system exhibits characteristics that, from the perspective of an external observer (the “stakeholders”), serve as proxies for intrinsic characteristics of the system. If these intrinsic characteristics indeed exist, they would enable the system to meet stakeholder expectations.

To establish trustworthiness, two distinct yet interconnected aspects must be addressed:

- The actual presence of certain characteristics within the system that enable it to fulfil stakeholders’ expectations.
- The ability of an observer (stakeholder) to predict the presence of these characteristics, based on certain behaviors that the system displays (proxies).

With this tentative reworking of the trustworthiness definition in mind, we can further elaborate our theoretical framework by delving into the concept of "expectations." To expect something is to conceptualise in a mental model the behaviour that an agent should exhibit to achieve a particular outcome (Lin et al. 2012) affected, for the observer, by a certain degree of uncertainty (Grimes et al. 2021). For instance, when I expect my dishwasher to clean dishes, I have a mental model of the internal processes necessary for achieving the final result, cleanliness. Conversely, I do not "expect" a pen to fall when it drops, because I am certain it will happen due to gravity.

Expectancy inherently involves a degree of uncertainty about potential outcomes. Hence, a good quantifiable measure of "expectation" might be the observer’s perceived uncertainty (Milliken, 1987) regarding the system’s behaviors (Jiang et al., 2022).

Summarizing the argument:

b) the need for an observer to expect something implies that they must construct a mental model of what should occur for a goal to be achieved. This necessity arises specifically when there is uncertainty about the connection between the system’s behaviors and the intended goal.

In light of claims a) and b), the attribution of trustworthiness operates as follows: trustworthiness is attributed to a system when the behavior it displays aligns with the observer's mental model. When this alignment occurs, the perceived uncertainty of the observer is low, leading them to trust that the system possesses the necessary intrinsic characteristics to meet their expectations.

To clarify this point, we will use the example of AI-driven robots. Robots can exhibit significant degrees of anthropomorphism, which helps illustrate our argument effectively.

Consider the example of proxemics and gestures, which are crucial features for building trust in social robots. If a robot is tasked with pouring water into a glass, but moves awkwardly and in a non-human-like manner, it may not conform to the observer’s mental model. The observer is likely to judge this behavior as insufficient for achieving the intended goal (successfully pouring the water).

To formalize this example, consider the following schema:

- A robot must perform a task in order to reach a goal.
- Mental model formation: Given i) the robot appearances ii) previous interactions with the robot iii) knowledge on how this goal is usually achieved, an observer creates a mental model of how a task should be performed by the robot. This model might include e.g. smooth, human-like movements during tasks such as pouring water.
- Behavior observation: The robot performs the task with movements that are either aligned or misaligned with this mental model.
- Perceived uncertainty evaluation: If the robot’s movements align with the mental model, the observer’s perceived uncertainty is low. Conversely, awkward or non-human-like movements increase perceived uncertainty.
- Trustworthiness attribution: Low perceived uncertainty leads the observer to attribute high trustworthiness to the robot, believing it possesses the necessary intrinsic characteristics to achieve the goal. High uncertainty results in a low attribution of trustworthiness.

- Outcome prediction: Based on the level of trustworthiness attributed, the observer predicts whether the robot will successfully complete the task.

More schematically:

- at time 0 the system starts performing an action to reach the target goal Z (object of the expectation),
- At time 1 the system exhibits behaviours X,
- This behaviour might prompt the observer to predict the presence of characteristics QY in the agent, since those specific behaviours, in the observer's experience, is a proxy of possessing QY.
- If QY are characteristic allowing to reach the goal Z (the expectation), then trustworthiness is attributed, else it is not.

Thus, two elements enhance our model of trustworthiness, allowing us to make a crucial distinction: the system must display behaviors that demonstrate possession of the qualities deemed necessary by the observer to 'meet stakeholders' expectations.' Furthermore, the system should exhibit these qualities through behaviors that reduce the observer's perceived uncertainty. This uncertainty is generated by the discrepancy (Δ) between the system's behavior and the observer's mental model.

We claim that this principle applies to any AI system that exhibits behaviors in a collaborative environment: the system must act in a manner that aligns with the observer's mental model of how the action should be performed to achieve the intended goal.

Before moving to the discussion on perceived trustworthiness, we point out one relevant issue, stemming from the above reasoning. One of the most important triggers of trustworthiness in the case of embodied agents, as AI-powered robots, is the match between what the observer expects as a behavior and what the robot actually performs as actions. A highly anthropomorphic robot will be expected to perform highly anthropomorphic actions, while a more mechanical-looking robot will be expected to perform poorly anthropomorphic actions. In fact, as said earlier, the observer mental model is not only based on previous experience on the task competition (e.g. pouring water), but also on agent-specific expectations, related to the agent shape and look, and on previous interactions with the agent. This claim is also consistent with Kätsyri's (et al. 2015) and Bisconti's (2021) claims about uncanniness, happening due to a mismatch between expectations and actual perceived behavior.

The Perceived Trustworthiness

Up until now we discussed that a system should possess and display some characteristics in order to align to the observer's mental model, and thus, consequently, be considered trustworthy. We did not touch upon which are these characteristics that the system should display and possess. As discussed above, it seems that the "ability, benevolence, competence" framework falls short (Mayer et al., 1995) in capturing this process.

However, a paradigm change is slowly emerging, to capture those characteristics that are crucial in the perceived trustworthiness of AI systems. We now discuss some key insight from the recent literature on the concept of perceived trustworthiness, identifying three perceived characteristics of AI systems that seems to influence the perceived uncertainty. It is important to note that the research in this field is still under constant evolution..

Transparency

In the relationship between human beings, the individual inevitably forms a mental model of the other agent. Said model becomes more detailed with each interaction and the more information is possessed about the person, the more the sense of understanding and predictability increases. Ultimately, gathering information constitutes a strategy to reduce perceived uncertainty. When a sufficient amount of information is available, the person becomes more "transparent" and it is possible to create a mental model that is close to the real way of working of the other; therefore, predictions about future behaviors align with the actual behaviors. Similarly, in human-machine interaction, this translates into a desire for information regarding the functioning of the system. This can be satisfied by endowing the system with transparency. It has been found that, as a consequence of reducing uncertainty, the transparency of a system can result in increased levels of trust (Liu, 2021). Generally, explanations about algorithmic models have been proven to enhance users' trust in the system (Zhang & Curley, 2018). However, some other studies have shown how providing information is not always helpful. For instance, Kizilcec (2016) showed how too much information negatively influences trust; Papenmeier et al. (2022) found how users can have more trust and insight into AI when there is no explanation about how the algorithm works, rather than when there is. Certainly, one aspect worth noting pertains to the adequacy of the information provided, depending, for example, on how it is framed and on its level

of complexity. In any case, reflecting on uncertainty, it can be assumed that for each system an ideal level of information could be identified so that it contributes to reduce uncertainty without confusing or overstimulating the user. Regarding this, the quality and quantity of information about the system should be tailored to the user's needs, e.g., depending on their need of knowledge to reduce uncertainty and on their ability to handle certain amounts of information without feeling overstimulated. Other than quantity and level of complexity, also the content of information should be chosen carefully. Transparency can be understood as giving information about the rationale of the system, as its rate of success in giving appropriate and correct outputs, or again as its speed in giving outputs. Users could feel the need of one kind of information more than others to form a mental model that is effective for them in contributing to reduce their levels of perceived uncertainty.

Agency Locus

However, even in the absence of concrete data with respect to operation, the users will still form assumptions in their mind. Such perception may impact the level of perceived uncertainty, especially by comparing mental images of the system's way of "thinking" that see it as more or less similar to human thinking mechanisms. The degree of humanness of a system's operating mechanisms may also derive from what the source of its model is. The machine's model or the cause of its apparent agency have been described as perceived agency locus, which can be external as created by humans, or internal, generated by the machine.

This distinction has become relevant with the development of AI technologies and the massive use of machine learning techniques, which are substantially different from systems programmed by humans, because they have the capacity to define or modify decision-making rules autonomously (Mittelstadt et al., 2016). This grants the system a certain degree of autonomy which makes it less subject to human determination and, as a consequence, less understandable and predictable to users. As, learned models are fundamentally different from human logic and machine agency could result in higher uncertainty and lower trust. In the case of agency locus, providing information to the user about the inner workings of the system could either reduce or increase the perception of uncertainty, whether its model is described as human-made or self-made. In fact, during an interaction, knowing that the system is programmed to follow human-made model seems to allow people to simulate its mental model and lead them to feel more knowledgeable about the decision-making process (Ososky et al., 2013).

Human Oversight

The presence of human oversight in the interaction setting requires a human agent to check and, in case, filter the workings and the output produced by the system. The fact that the process is not solely controlled by a machine could contribute to establish lower levels of perceived uncertainty, as the user is "reassured" by the presence of a third agent that is supposed to be perceived as more similar to them and therefore more understandable and predictable. The predictability of the human's actions and decisions will, therefore, transfer to the artificial agent.

Figure 1. Perceived trustworthiness theoretical framework

From these premises, a theoretical framework was developed (Figure 1) aiming to comprehensively represent the mechanisms of trust formation as mediated by uncertainty. This framework is intended to fully capture the dynamics of trust development in AI systems, by including characteristics of both the human agent and the AI system. Keeping an eye on both parts of the interaction is an element of novelty within the human-machine interaction research.

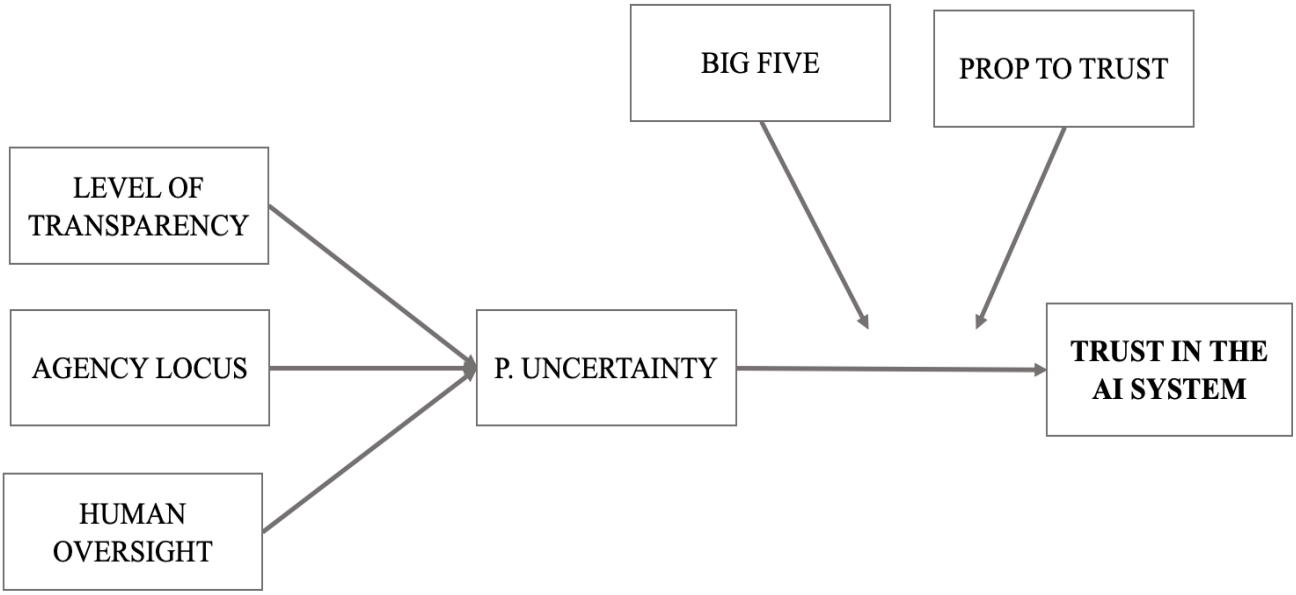
This theoretical framework interprets trust in artificial systems as being closely influenced by the level of uncertainty perceived by the user (perceived uncertainty). The latter would, in turn, be dependent on three basic aspects: the level of transparency of the system, agency locus and the presence of human oversight. In addition, other variables regarding individual user characteristics, such as personality traits,as understood by Big Five theory (John et al. 2008, and propensity to trust in general, would also intervene as mediators.

Personality Traits

Personality traits have been found to be predictive of the acceptance of AI (Bernotat & Eyssel, 2017; Esterwood et al., 2021), and, as a consequence, it can be hypothesized that they contribute to trust formation as well. The Big Five model (John et al., 2008) identifies five major traits: openness, conscientiousness, agreeableness, neuroticism, and extraversion (Rhee et al., 2013). Each of these traits has been found to determine characteristics in people that influence their attitude towards technology. For instance, Driskell et al. (2006) showed how significant openness traits can make people more receptive to new technologies and unfamiliar environments, whilst lower levels are linked to being accustomed to familiar. Therefore, it can be assumed that individuals with lower openness will also have a lower tolerance of uncertainty, needing the system to be more predictable to feel comfortable and trust it. Similarly, people with high degrees of agreeableness tend to trust others and new projects (Matsumoto & Juang, 2016). On the other hand, traits of neuroticism in people make them more likely to be hostile to unfamiliar things that require communication and detachment, while people with lower degrees are more emotionally stable in the face of new technologies (Jeronimus et al., 2016). Thus, users high in neuroticism will need more information and reassurance to reduce uncertainty levels when facing a new system. Taking into account the user's personality traits can help understand to which degree they will tolerate their expectations about the system's behavior not being met.

Propensity to Trust

Propensity to trust represents a person's general tendency to trust others regardless of the specific situation (McKnight et al., 1998). Dispositional trust is generally a stable trait as it is not immediately influenced by context and only undergoes changes due to learning from previous interactions (Colquitt et al., 2007). Users with higher levels of propensity to trust will be more prone to accepting and trusting new systems, even without needing to gather a lot of information about them. They could tolerate uncertainty better, being naturally drawn to trusting other agents (human or artificial).



A Formal Account of Trustworthiness

Our theoretical model of perceived trustworthiness is not intended to capture all the different relevant aspects of the observer's mental model. However, our claim is that the elements we identified are at least some of the relevant perceived characteristics, and that these are not ability, competence, benevolence. Having worked out this tentative theoretical framework for perceived trustworthiness, we can now plug it into our general model, and attempt a mathematical formulation.

First, we state the claim c, that along with a) and b) are the key claims of this manuscript.

c) To attain trustworthiness, an AI system must display behaviors that engender a sense of transparency, convey a human-like agency, and demonstrate human oversight in the eyes of the observer. Generally, these traits should align with the observer's mental models of anticipated behaviors, thereby fulfilling their expectations.

We can now deliver a formal account.

- **Overall Trustworthiness, T**: Represents the complete evaluation of the system's trustworthiness, influenced by both intrinsic and perceived elements.

i)

$$T = \frac{1}{1 + \Delta T}$$

ΔT is defined as the absolute difference between intrinsic trustworthiness (Ti) and perceived trustworthiness (Tp).

ii)

$$\Delta T = \mid T_i - T_p \mid$$

- **Intrinsic Trustworthiness, T_i,** will be defined as:

iii)

$$T_i = \frac{Q}{Q_{total}}$$

This component quantifies the proportion of intrinsic qualities actually present in the system ($|Q|$) relative to the total qualities needed to perform the action, in such a way that will meet stakeholders' expectations ($|Q_{total}|$). Therefore, it reflects the system's capability to meet stakeholder expectations based on inherent attributes. In the case of the robot pouring water, an actuator able to grab and move the glass, enough degrees of freedom to perform the action, computer vision, etc.

Under $|Q_{total}|$ we find all those technical characteristics that we discuss at the end of the first chapter, referring to the lack of an established taxonomy. It is highly unclear which are the characteristics needed by an AI system to meet stakeholders' expectations: no clear taxonomy is yet developed, and no clear relationship between trustworthiness characteristics gained enough scientific consensus. Moreover, T_i is also very much use-case sensitive, therefore for a thorough understanding of T_i we should contextualize it to the use case. Attempts in this direction are, in the International Standardization Organization (ISO) the concept of Profile or the concept of Operational Design Domain.

Perceived Trustworthiness, T_P, is the other element of ii). As for what concerns perceived trustworthiness, we claim that three main characteristics will have an influence (Transparency, Agency Locus, Human Oversight), mediated by *M* and adjusted for Perceived Uncertainty:

iv)

$$T_p = \frac{B}{\phi(L, A, H, M)} \cdot u$$

In this formula, *Tp* is determined first by the observed behaviors (*B*), that might be measured with quali-quantitative methods as questionnaires or focus groups. These behaviors are the ones that are displayed by the system and perceived by an observer, where the way they are perceived have of course epistemic priority. Second, iv) is determined by (*B*) alignment with a composite

measure of expectations of behaviors $\phi(L,A,H,M)$, which depends on Level of Transparency (L), Agency Locus (A), Human Oversight (H), the mediators (M) identified above (e.g. propensity to trust, personality traits, etc.) and u, explained below. This new definition of ϕ allows us to consider how these four factors collectively shape stakeholder expectations and, consequently, perceived uncertainty. Therefore, T_p is given by the relationship between the behaviors that are observed (B) and the ones expected (ϕ) in performing the task. As every stakeholder is an observer, this formalization could be theoretically applied to any stakeholder. On the other hand, this schematization cannot completely capture the complexity of stakeholders' perceived trustworthiness, since it is not sensitive a) to the perspective and the specific point of observation of each specific stakeholder and b) therefore to the fact that each stakeholder might value specific characteristics of the AI system differently. In fact, the stakeholders of an AI system go well beyond users, that are our ideal observers in the case of formula iv). Taking as example a medical decision support system, the stakeholders impacted directly are at least the physician and the patient. It is easy to note that already these two have profoundly different expectations, concerns, target goals.

Perceived Uncertainty u, completing formula iv)

v)

$$u = exp(-\gamma * | B - \phi(L, A, H) |)$$

The uncertainty function u articulates the decrease in perceived uncertainty with greater alignment between observed behaviors and the composite measure of expected behaviors, where γ is a constant that modulates this sensitivity. The formula $u = exp(-\gamma * | B - \phi(L, A, H) |)$ quantifies perceived uncertainty regarding the alignment of observed behaviors B with expected behaviors ϕ. Here, γ is a scaling factor that controls the sensitivity of perceived uncertainty to discrepancies between observed and expected behaviors. The expression $| B - \phi(L, A, H) |$ calculates the absolute difference between these two sets of behaviours. The exponential function, *exp*, is used to model the decrease in perceived uncertainty; as the difference decreases, the exponential term approaches zero, causing u to approach 1. This reflects a decrease in uncertainty as observed behaviors more closely align with what is expected. Conversely, larger differences result in higher values of the exponential term, driving u towards zero and indicating increased uncertainty. This modelling choice captures the non-linear impact that behavioral discrepancies have on trust perception, emphasizing that smaller deviations are significantly more tolerable than larger ones.

Trustworthiness Formula

The overall trustworthiness, denoted as T, is modelled to be inversely proportional to the discrepancy ΔT, which now encapsulates the combined effects of actual system qualities and perceived uncertainty as influenced by the integrated expectations. The expanded formula for trustworthiness is:

vi)

$$T = \frac{1}{1 + \left| \frac{|Q|}{|Q_{total}|} - \left(\frac{|B|}{|\phi(L, A, H, M)|} * \ exp(-\gamma * |B - \phi(L, A, H)|) \right) \right|}$$

This adjustment allows ΔT to capture not only the discrepancies in quantity and quality between T_i and T_p, but also the stakeholder's confidence in the system's behaviors aligning with expectations.

Conclusions and Limitations

The conceptual model formulated in this paper has the objective to theoretically ground the concept of AI trustworthiness in its double understanding of a system possessing some characteristics, and the perception of humans toward the system.

The formula reflects a foundational principle: the closer the system's behaviors align with stakeholder expectations and the lower the perceived uncertainty, the higher the attributed trustworthiness.

The practical utility of our trustworthiness formula is manifold. Firstly, by quantifying trustworthiness as a function of both perceived and intrinsic qualities, stakeholders can pinpoint specific areas where an AI system may fall short of expectations or excel. This dual perspective allows for targeted improvements in AI design and deployment, ensuring that both technical robustness and user perception are addressed simultaneously. For example, in the development of autonomous vehicles, our formula can help assess how well these vehicles meet safety standards (intrinsic qualities) while also gauging public trust through perceived safety and reliability.

Moreover, the formula's emphasis on reducing perceived uncertainty through alignment with stakeholders' mental models offers a strategic pathway for enhancing user engagement and trust. By systematically evaluating how observed behaviours of AI systems match expected behaviours, developers can adjust system transparency, agency locus, and the extent of human oversight. This alignment is crucial in sensitive applications such as medical diagnosis aids or financial advising systems, where the stakeholder's trust significantly impacts the technology's adoption and effectiveness.

Before discussing the limitations of this study, it is important to address a foundational assumption concerning the dichotomy between intrinsic and perceived trustworthiness. This sharp delineation between the objective and subjective aspects of trustworthiness is epistemologically contentious. Indeed, the characteristics deemed intrinsic to trustworthiness are, in fact, those the system's human developer has chosen to implement. Consequently, from an epistemological standpoint, there is no clear-cut distinction between the two: characteristics labelled as "intrinsic" are shaped by human designers, albeit not wholly determined due to the nondeterministic nature of AI systems. Nevertheless, for the purposes of clarity in this paper, we maintain this distinction to differentiate between characteristics that are essential for a system to achieve its objectives and those that are inferred based on the system's behaviour. At a specific point in time—time 0—when a system executes an action, there are some characteristics that the system possesses, and some characteristics that the system displays. These two sets can be different. Without simplifying the concept of intrinsic versus perceived trustworthiness, it would be challenging to consider this aspect.

Some *caveat* should be highlighted, along with some limitations of this paper:

i) the characteristics that a system should possess (intrinsic) can be different from the ones it needs to display (perceived) in order to actually be trustworthy. Therefore, future works should focus on how to ensure overlap between the two, as much as possible.

ii) As said, the fact that uncertainty is influenced by transparency, agency locus and human oversight will be subject of further experiments and specification, but in this paper we aimed at outlining a schematic and formalized theoretical model to build upon during experimental settings. Even if ϕ would be different in its components, this would not change the formula at large.

iii) Even if we would be able to capture the characteristics that should be perceived by an observer to judge the system trustworthy, these are subject-sensitive, and culturally variable. Our schematization does not represent, in its current stage, this issue.

iv) The formula only captures static trustworthiness, but trustworthiness is actually dynamic. Further specifications of the formula might embed the dynamicity of trustworthiness, both understood as dynamic in the perceptions of the user, and dynamic in the use cases an AI system might be deployed.

v) Trust in AI systems always happen in a sociotechnical fashion, while this formula only captures the dyadic interaction between one AI system and one observer. In fact, social trust is mediated by many other aspects that might only partially pertain to the technology and its functioning, as trust in institutions and trust in the deploying and implementing organizations.

vi) While we integrated propensity to trust in the model for perceived trustworthiness, this does not take into account that this differs among stakeholders. They might have different propensity, and also biases towards specific use cases influencing the outcome.

vii) Trustworthiness is use-case sensitive, and $\frac{Q}{Q_{total}}$ should be considered as a schematization that will change for every use case. The problem of intrinsic trustworthiness being use-case sensitive will require scholarly literature and standardization bodies to develop contextual taxonomies.

Given all this caveats, our integrative model aims to capture the essence of trustworthiness in AI systems, linking intrinsic system characteristics with the human perceptions influenced by uncertainty. By acknowledging and quantifying the impact of perceived uncertainty, the formula provides a tool for evaluating trust in AI systems, emphasizing the importance of aligning system behaviors with stakeholder expectations to foster trust.

In future work we aim to effectively incorporate issues i) to vii) into our formula, and experimentally validate our claims related to perceived uncertainty. As anticipated, the experimental testing of the presented hypotheses is set to be done through the manipulation of transparency, agency locus and the presence of human oversight. Specifically, what is to be investigated is how the level of transparency with which the system is equipped can influence the user's perception of uncertainty. Although

transparent systems have been found to be judged as more trustworthy (e.g., Zhang & Curley, 2018), there is no guarantee that providing complete transparency is an effective way to ensure optimal interaction. Indeed, people may not respond positively to a large amount of information because of the cognitive load it would entail. Therefore, it is interesting to test how much information people need in order to reach an acceptable degree of uncertainty, also in relation to their natural predisposition to trust. Further objective is to test how much human intervention in the operation of the artificial system can affect perceived uncertainty. Building on what has already been reported in the literature (e.g., Liu, 2021), we assume that operations characterized as human-like lead to a reduction in the perception of uncertainty. Such human likeness can be expressed either as external agency locus or as the presence of oversight by a human agent. In terms of experimental setting, what is intended is to propose to the participants an interaction with an artificial system under different conditions, namely: in the presence of low, medium or high levels of transparency depending on the amount of information provided; with systems presented as having internal or external agency locus; with or without the presence of oversight by a human. From the combination of these conditions, we then want to observe which elements are most salient in modulating uncertainty and which levels are optimal for improving the quality of interaction.

A solid understanding of the concept of trustworthiness is of paramount importance for the future of design and policies, since this framework is becoming the overarching one, at least in the EU, for development and adoption of AI systems.

The theoretical model and the formal account presented in this paper aims at practically implementing a clear, formalized and reproducible approach to trustworthiness measurement. The main added value of our proposal is considering both intrinsic and perceived characteristics of trustworthiness in a coherent theoretical model, along with the presenting a formal account.

Future work should concentrate on refining this theoretical framework and analysing its implications for the taxonomy of trustworthiness. Additionally, future studies should investigate the relationship between a system's inherent qualities and users' perceptions of these qualities. This exploration is part of another project, which we expect to produce significant findings in the near future.

Fundings

The research leading to these results partially received funding from the EU Commission under Grant Agreement No. 101094665.

This work has been partially supported by PNRR MUR project PE0000013-FAIR.

References

Bernotat, J., & Eyssel, F. (2017). A robot at home—How affect, technology commitment, and personality traits influence user experience in an intelligent robotics apartment. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 641–646. https://doi.org/10.1109/ROMAN.2017.8172370

Bisconti, P. (2021). How Robots' Unintentional Metacommunication Affects Human–Robot Interactions. A Systemic Approach. Minds and Machines, 31(4), 487–504. https://doi.org/10.1007/s11023-021-09584-5

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. Journal of Applied Psychology, 92(4), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

Driskell, J. E., Goodwin, G. F., Salas, E., & O'Shea, P. G. (2006). What makes a good team player? Personality and team effectiveness. Group Dynamics: Theory, Research, and Practice, 10(4), 249–271. https://doi.org/10.1037/1089-2699.10.4.249

Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., & Robert, L. P. (2021). Birds of a Feather Flock Together: But do Humans and Robots? A Meta-Analysis of Human and Robot Personality Matching. 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), 343–348. https://doi.org/10.1109/RO-MAN50785.2021.9515394

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and Engineering Ethics, 26(6), 3333–3361.

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., Yordanova, K., Vered, M., Nair, R., Abreu, P. H., Blanke, T., & Pulignano, V. (2023). A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. Artificial Intelligence Review, 56(4), 3473–3504.

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. Decision Support Systems, 144, 113515.

Jeronimus, B. F., Kotov, R., Riese, H., & Ormel, J. (2016). Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: A meta-

analysis on 59 longitudinal/prospective studies with 443 313 participants. Psychological Medicine, 46(14), 2883–2906. https://doi.org/10.1017/S0033291716001653
Jiang, J., Kahai, S., & Yang, M. (2022). Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. International Journal of Human-Computer Studies, 165, 102839.
John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. Handbook of Personality: Theory and Research, 3(2), 114–158.
John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. Handbook of Personality: Theory and Research, 3(2), 114–158.
Jones, S. L., & Shah, P. P. (2016). Diagnosing the locus of trust: A temporal perspective for trustor, trustee, and dyadic influences on perceived trustworthiness. Journal of Applied Psychology, 101(3), 392.
Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Frontiers in psychology, 6, 390 https://doi.org/10.3389/fpsyg.2015.00390
Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. ACM Computing Surveys (CSUR), 55(2), 1–38.
Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395. https://doi.org/10.1145/2858036.2858402
Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., & Zhang, J. (2012). Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In Proceedings of the 2012 ACM conference on ubiquitous computing (pp. 501-510).
Liu, B. (2021). In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human-AI Interaction. Journal of Computer-Mediated Communication, 26(6), 384–402. Scopus. https://doi.org/10.1093/jcmc/zmab013
Matsumoto, D., & Juang, L. (2016). Culture and psychology. Cengage Learning.
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. The Academy of Management Review, 20(3), 709. https://doi.org/10.2307/258792
McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial Trust Formation in New Organizational Relationships. The Academy of Management Review, 23(3), 473. https://doi.org/10.2307/259290
Milliken, F. J. (1987). Three types of perceived uncertainty about the environment: State, effect, and response uncertainty. Academy of Management Review, 12(1), 133–143.
Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 205395171667967. https://doi.org/10.1177/2053951716679679
Ososky, S., Philips, E., Schuster, D., & Jentsch, F. (2013). A Picture is Worth a Thousand Mental Models: Evaluating Human Understanding of Robot Teammates. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57(1), 1298–1302. https://doi.org/10.1177/1541931213571287
Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. ACM Transactions on Computer-Human Interaction, 29(4). Scopus. https://doi.org/10.1145/3495013
Rhee, J., Parent, D., & Basu, A. (2013). The influence of personality and ability on undergraduate teamwork and team performance. SpringerPlus, 2(1), 16. https://doi.org/10.1186/2193-1801-2-16
Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In Proceedings of Mensch und Computer 2021 (pp. 325–329).
Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence. *Preprint.*
Zhang, J., & Curley, S. P. (2018). Exploring Explanation Effects on Consumers' Trust in Online Recommender Agents. *International Journal of Human–Computer Interaction*, *34*(5), 421–432. https://doi.org/10.1080/10447318.2017.1357904