Prior to Trust: Frequentist and Bayesian views of Trust in AI

Mattia Petrolo[1,*], Ekaterina Kubyshkina[2] and Giuseppe Primiero[2]

[1]University of Lisbon, CFCUL, Alameda da Universidade, 1649-004 Lisbon, Portugal
[2]Logic, Uncertainty, Computation and Information Lab, PhilTech Research Center, Philosophy Department, University of Milan, Via Festa del Perdono, 7 20122, Milan, Italy

Abstract

The notions of trust and trustworthiness in the field of AI are currently the focus of a collective, interdisciplinary effort for clarification. In this work, we contribute to this ongoing debate by identifying two senses in which an agent might place trust in an AI system. The first sense, referring to trustworthiness as formalised in previous work, considers the results of tests conducted on the system alongside the agent's expectations. The second sense, extends the former by factoring in the agent's "pragmatic" background when considering these tests. We argue that these two forms of trust can be understood in relation to well-known approaches in statistical inference: the first aligns with a *frequentist* interpretation, while the second reflects a *Bayesian* view of trust.

Keywords

Trustworthy AI, Reliable AI, Statistical inference

1. Introduction

The concepts of trust and trustworthiness in AI are currently the subject of an interdisciplinary effort of conceptual, formal and procedural clarification. This is evident from the increasing attention these notions are receiving across various fields of research. AI engineers, for instance, are working to incorporate properties into AI systems to enhance their trustworthiness (see, e.g., [1]). Meanwhile, philosophers are engaged in defining Trustworthy AI (TAI), exploring its epistemological and ethical implications, and debating whether it is even possible to discuss TAI without committing a categorical mistake (see, e.g., [2], and [3] for a critical discussion). Logicians, on the other hand, are developing formal systems to capture the complex and elusive concepts of trust and TAI (see, e.g., [4], [5]). Finally, sociologists are examining the societal impacts of trusting AI-based technologies (see, e.g., [6]).

The challenge of understanding trust and trustworthiness in AI is not purely theoretical. The widely discussed European proposal for the Artificial Intelligence Act [7], inspired by the Ethics Guidelines for Trustworthy AI [8], stipulates that AI systems and the information they generate must be reliable, transparent, and trustworthy, among other things. However, these terms have distinct and not always shared definitions. This lack of clarity, and the absence of a solid theoretical foundation, is source of potential misunderstandings that could affect the social perception of AI systems. Without a clear definition of these concepts, there is a genuine risk that the Guidelines and the AI Act may lack the practical relevance necessary for meaningful implementation. Given that these frameworks aim to regulate issues of critical importance to human well-being and governance, a deeper analysis and clarification of the notions of trust and trustworthiness in AI is essential.

In this paper, we contribute to this ongoing debate by identifying two senses in which an agent might place trust in an AI system. The first sense considers the results of tests conducted on the system alongside the agent's expectations. The second extends the former by factoring in the agent's "pragmatic" background when considering these tests. We argue that these two forms of trust can

3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE24), co-located with AIxIA 2024, November 25-28, 2024, Bolzano, Italy
*Corresponding author.
mpetrolo@fc.ul.pt (M. Petrolo); ekaterina.kubyshkina@unimi.it (E. Kubyshkina); giuseppe.primiero@unimi.it (G. Primiero)

be understood in relation to well-known approaches in statistical inference: the first aligns with a *frequentist* interpretation, while the second reflects a *Bayesian* view of trust.

2. Two accounts of trust in AI systems

When it comes to trusting an AI system, particularly a Machine Learning (ML) system, there are at least two distinct ways in which an agent can do so. To illustrate this, we borrow and slightly modify an example from [1]. Let us consider a simple example. Imagine a classifier $\mathscr{C}$ designed to identify pictures of wolves, and assume that $\mathscr{C}$ has already been evaluated as trustworthy according to some relevant metrics. Now, suppose we present two images to the classifier: one of a wolf and the other of a Siberian Husky. Upon processing, $\mathscr{C}$ classifies both images as wolves. Let us now consider two agents, A_1 and A_2, both aware that $\mathscr{C}$ has been deemed trustworthy, and both receiving the same classification output. The only difference between the two agents is that A_2 is an expert dog trainer, while A_1 is not. At this point, their reactions diverge: A_1 trusts the output, while A_2 does not. How can we explain the difference in their responses?

To address this, let us first examine what allows A_1 to trust $\mathscr{C}$. We will assume some conditions we consider necessary for $\mathscr{C}$ to be considered trustworthy. In the following we assume a notion of trustworthiness for non-deterministic computational system as the one proposed in [9], [4], [5]. In this framework, a trustworthy non-deterministic process for a given output is defined as one where the frequency of that output, over a specified number of trials, does not deviate beyond an acceptable threshold from its expected probability. This understanding relies on a series of tests performed on the system and the alignment of these tests with the expectations an agent has regarding the system's behavior. Note that this interpretation is not necessarily constrained to a sharp measure of probability and could be extended naturally to a graded version. In this context, trustworthiness is always indexed by both an agent and an output. Different agents may have (or assume) varying expectations about a system's performance, leading them to assess the trustworthiness of the same system differently. Similarly, a system may be deemed trustworthy for certain outputs but not for others. For example, a ML system might be well-trained to provide accurate answers about historical events, but not about current events. Thus, it can be considered trustworthy in relation to historical outputs while being untrustworthy for current ones. Necessary conditions for this notion of trustworthiness to induce an epistemic state are: first, A_1 knows that $\mathscr{C}$ is trustworthy – meaning that the behavior displayed by $\mathscr{C}$ is as *expected* by the agent in any epistemic scenario; and second, A_1 has an evidence for $\mathscr{C}$ being trustworthy, giving them a *justification* to accept the output as correct. With this understanding, we can characterize a first form of trust in an AI system, which we will refer to as $trust_1$:

An agent A $trusts_1$ an AI system S iff
A has an evidence that S produces an output in accordance with the behavior of S as expected by A.

This notion of trust is widely referenced in the literature on evaluating AI trustworthiness (see, for example, [5]). In this context, trustworthiness is viewed as a crucial component, while other aspects of trust are set aside.

Let us return to our example of the classifier and examine the reasons why the second agent, A_2, might not trust the classifier, unlike the first agent, A_1. Assume that both agents possess the same knowledge about $\mathscr{C}$. However, A_2 may have additional beliefs regarding the potential inaccuracies of the output, even if they acknowledge that such inaccuracies are expected. We argue that these additional beliefs, which lead to the divergence in trust between A_1 and A_2, stem from the specific *pragmatic background* of A_2. By pragmatic background, we refer to the set of beliefs an agent holds prior to interacting with the AI system. These beliefs may be shaped by various factors, including education, experience, cultural context, and moral or ethical principles. Based on this understanding, we can characterize this extended form of trust in an AI system, which we will refer to as $trust_2$:

An agent A $trusts_2$ an AI system S iff
A $trusts_1$ S and the output of S is compatible with the pragmatic background of the agent.

As evident from the previous characterization, $trust_2$ extends $trust_1$ by incorporating an agent's belief set and comparing it with the output provided by the AI system. With the definitions of $trust_1$ and $trust_2$ established, let us revisit our motivating example of the classifier and compare the two forms of trust held by A_1 and A_2.

Viewing the example through the lens of these definitions, we can assert that A_1 $trusts_1$ $\mathscr{C}$ to classify both pictures as wolves. Furthermore, A_1 also $trusts_2$ $\mathscr{C}$ for the same classification, because the received output does not contradict their pragmatic background. In contrast, while A_2 $trusts_1$ $\mathscr{C}$ to classify both pictures as wolves, their situation diverges. A_2 does not $trust_2$ $\mathscr{C}$ for this classification. This divergence may stem, for instance, from A_2's background as a dog trainer who has previously trained a Siberian Husky. In this context, a single error from $\mathscr{C}$ does not undermine its overall trustworthiness, and A_2 is aware of this. Thus, A_2 maintains $trust_1$ in $\mathscr{C}$. However, A_2 recognizes the Siberian Husky and believes, based on their education and experience, that a Siberian Husky is not a wolf. Consequently, they would not base further reasoning or actions on this erroneous output. In this sense, A_2 does not $trust_2$ $\mathscr{C}$, as the output contradicts their pragmatic background.

3. Trust in AI via statistical inference

As evident from our definitions of $trust_1$ and $trust_2$, these notions are not mutually exclusive; rather, $trust_1$ is included in $trust_2$. The primary distinction is that while $trust_1$ is based solely on the calculation of the correspondence between data obtained from a sufficient number of tests and an agent's expectations about an AI system, $trust_2$ incorporates the agent's overall background into the reasoning process. From this perspective, we argue that the two kinds of trust discussed in this article naturally correspond to two forms of statistical inference: frequentist and Bayesian approaches.

Frequentists define the probability of an event as the limit of its relative frequency over a large number of trials, whereas Bayesians extend probabilities to account for varying degrees of certainty about statements (see, e.g., [10] for more details). The fundamental difference between these approaches lies in their treatment of probabilities: frequentists analyze probabilities purely as calculations based on data located on a sample space of possible outcomes, while Bayesians include the dimension of an agent's knowledge about that data.

As previously noted, $trust_1$ relies on the knowledge of a system's trustworthiness, as discussed in [9], [4], [5]. In this context, trustworthiness is established through a post-hoc verification strategy that evaluates the reliability of an AI system's behavior in statistical terms, alongside adherence to an evaluation criterion (see [11]). Specifically, this verification employs two specific comparison terms. First, it uses a formal expression to denote the observable behavior of a (possibly) non-deterministic system over a finite number of executions. Second, it incorporates a transparent model of the expected behavior, which is normatively or ethically desirable based on the observed model and the input data. This second model serves as a benchmark for evaluating the first observed model, and the formal verification process measures the distance between the two models. From this perspective, assessing trustworthiness is fundamentally tied to considering the P-value of the results in frequentist terms. This means measuring the probability of obtaining the observed results under the assumption that the null hypothesis is true. This hypothesis posits that there is no statistically significant relationship between two sets of data. In this context, it emphasizes the need to evaluate trustworthiness based on a sufficient number of *distinct* tests performed on the system.

Let us reconsider our example in frequentist terms. In order to establish $trust_1$ an agent takes into account the probability of getting a result which identifies wolf as wolf (lets dub it *Result*), during a sufficient number of tests (*Data*):

$$P(Result) = \frac{\#Result}{\#Data}$$

Then, the agent verifies whether $P(Result)$ matches an acceptable threshold against the expected probability for $\#Result$. In our example, both agents evaluate that the observed frequency of *Result* sits within an acceptable threshold compared to its theoretical counterpart, thereby inferring trustworthiness. Notice, that even though for A_2 the output was not accounted in the number of *Result*, the difference is so insignificant that $P(Result)$ still matched an acceptable threshold.

Since $trust_1$ is included within $trust_2$, trustworthiness – and, by extension, frequentist statistical reasoning – plays a significant role in establishing $trust_2$. However, a distinguishing feature of $trust_2$ is its incorporation of the agent's pragmatic background in the evaluation. This pragmatic background reflects the current state of the agent's knowledge, not only regarding the results of testing an AI system (i.e., the data) but also encompassing prior information and hypotheses about the system and acceptable outcomes. From this perspective, the pragmatic background can be viewed as a *prior probability*, which represents the probability assigned to an output before receiving the relevant information. In this broad sense of pragmatic background, $trust_2$ seems to align with a Bayesian interpretation of the probability of the output, as it incorporates the dimension of the agent's prior credence or degree of the agent's beliefs, which can be updated subsequently.

Returning to our example of the classifier $\mathscr{C}$, we can say that A_1 and A_2 establish their $trust_1$ in $\mathscr{C}$ based on the overall trustworthy behavior of the classifier, which is measured in frequentist terms. In the case of $trust_2$, however, the agents A_1 and A_2 appear to have different priors –specifically, differing knowledge and assumptions about dogs and wolves – which influences their attitudes toward the output. From this perspective, we notice that the conditional belief (*posterior belief* in Bayesian terms) of A_1 and A_2 differs, once it is calculated via Bayes' theorem:

$$P(Result \mid Data) = \frac{P(Data \mid Result) \times P(Result)}{P(Data)},$$

that is the conditional belief in event ($P(Result \mid Data)$) is calculated by multiplying prior belief of the agent by the likelihood $P(Data \mid Result)$ that *Data* will occur if *Result* is true. Clearly, $P(Result \mid Data)$ would be significantly different for A_1 and A_2, given that $P(Result)$ is different for them, as well as the likelihood for A_1 is much higher than the one for A_2.

4. Conclusion

The association of the two forms of trust we introduced with established methods of statistical inference supports the distinction between $trust_1$ and $trust_2$. These forms of trust allow for a focus on different objectives when evaluating an AI system. Specifically, $trust_1$ can be seen as measuring the reliability of a system with respect to a benchmark, while $trust_2$ involves an agent's attitudes and hypotheses, which may not be directly tied to the AI system itself. A notable aspect of our analysis is the relationship between trust and trustworthiness, where trust inherently presupposes trustworthiness. In our framework, trustworthiness is always relative to the agent. From this perspective, it seems natural to assert that if an agent trusts an AI system, they must perceive it as trustworthy. However, the reverse is not necessarily true: an agent may consider a system trustworthy without actually placing their trust in it. Formally, both $trust_1$ and $trust_2$ can be applied depending on the desired level of abstraction in the model. The development of a formal framework to represent these two types of trust remains a topic for future research.

Acknowledgments

The authors would like to thank two anonymous reviewers for their comments. All authors acknowledge the support of the Project PRIN2020 BRIO - Bias, Risk and Opacity in AI (2020SSKZ7R) awarded by the Italian Ministry of University and Research (MUR). Giuseppe Primiero is further funded through the project PRIN2022 SMARTEST - Simulation of Probabilistic Systems for the Age of the Digital Twin (2022E8Y4X) awarded by the Italian Ministry of University and Research (MUR). The research of

Ekaterina Kubyshkina is funded under the “Foundations of Fair and Trustworthy AI” Project of the University of Milan. Giuseppe Primiero and Ekaterina Kubyshkina are further funded by the Department of Philosophy “Piero Martinetti” of the University of Milan under the Project “Departments of Excellence 2023-2027” awarded by the Ministry of University and Research (MUR). Mattia Petrolo acknowledges the financial support of the FCT – Fundação para a Ciência e a Tecnologia (2022.08338.CEECIND; R&D Unit Grants UIDB/00678/2020 and UIDP/00678/2020) and the French National Research Agency (ANR) through the Project ANR-20-CE27-0004.

References

[1] M. Ribeiro, S. Singh, C. Guestrin, “why should i trust you?”: Explaining the predictions of any classifier, in: KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[2] J. M. Durán, K. R. Jongsma, Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical a.i., Journal of Medical Ethics 47 (2021) 329–335.

[3] G. G Zanotti, M. Petrolo, D. Chiffi, V. Schiaffonati, Keep trusting! a plea for the notion of trustworthy ai, AI & Soc (2023). doi:https://doi.org/10.1007/s00146-023-01789-9.

[4] F. A. D’Asaro, G. Primiero, Probabilistic typed natural deduction for trustworthy computations, in: D. Wang, R. Falcone, J. Zhang (Eds.), Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021), CEUR Workshop Proceedings, 2021, pp. 1–12.

[5] F. A. D’Asaro, F. Genco, G. Primiero, Checking trustworthiness of probabilistic computations in a typed natural deduction system, 2023. URL: https://arxiv.org/pdf/2206.12934.pdf, coRR, arXiv:2206.12934 [abs].

[6] J. Dacon, Are you worthy of my trust?: A socioethical perspective on the impacts of trustworthy ai systems on the environment and human society, 2023. URL: https://arxiv.org/pdf/2309.09450, coRR, arXiv:2309.09450.

[7] Regulation (eu) 2024/... of the european parliament and of the council laying down harmonised rules on artificial intelligence (ai act), European Commission, 2024. URL: https://artificialintelligenceact.eu/the-act/.

[8] Ethics guidelines for trustworthy ai of the high-level expert group on artificial intelligence, AI HLEG, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[9] E. Kubyshkina, G. Primiero, A possible worlds semantics for trustworthy non-deterministic computations, International Journal of Approximate Reasoning 172 (2024) 1–24. doi:https://doi.org/10.1016/j.ijar.2024.109212.

[10] A. Hájek, Interpretations of probability, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, winter 2023 ed., 2023.

[11] G. Primiero, Brio: il ruolo della logica nella costruzione di una ia equa, in: L. Marinucci, C. Caporale (Eds.), Spiegabilità e Intelligenza Artificiale. Una prospettiva multidisciplinare, Etica della ricerca, bioetica, biodiritto, biopolitica, CNR Edizioni, to appear.