Letting Go of the Numbers: Measuring AI Trustworthiness

Carol J. Smith
Carnegie Mellon University, U.S.A.

Abstract: AI systems need to be designed to work with, and for, people. A person's willingness to trust a particular system is based on their expectations of the system's behavior. Their trust is complex, transient, and personal – it cannot easily be measured. However, an AI system's trustworthiness can be measured. A trustworthy AI system demonstrates that it will fulfill its promise by providing evidence that it is dependable in the context of use, and the end user has awareness of its capabilities during use. We can measure reliability and instrument systems to monitor usage (or lack thereof) quantitatively. However, AI's potential is bound to perceptions of its trustworthiness, which requires qualitative measures to fully ascertain. Doing AI well requires a reset – letting go of (some of) the numbers and learning new methods that provide a more complete assessment of the system.

EXTENDED ABSTRACT

AI systems need to be designed to work with, and for, people. Despite this obvious requirement, the performance of AI systems has been primarily focused on numeric (quantitative) values such as accuracy and F1 scores. These measures are important, but consider what we can learn about the system's fit with user needs from an accuracy score (not much). How about measuring the system's trustworthiness to end users from an accuracy score?

Quantitative metrics alone are not capable of providing a holistic view of the system's design, performance, and usage. If end users have an issue, quantitative information cannot typically provide enough information to fully understand the issue, nor can they provide enough guidance to address it. Unaddressed issues add up and eventually will affect system use. Despite this, only minimal effort is typically made to develop measures to determine if people using AI systems find them to be helpful and trustworthy. Prioritizing a good user experience during development and while the system is deployed, will help support a successful AI system.

An AI system's potential is bound to stakeholders' perceptions of its trustworthiness, which requires qualitative measures to fully ascertain. I use the term stakeholders to include those developing, acquiring, using, and being affected by AI systems. I will primarily discuss the people using systems and those who are affected by AI systems and I will present alternatives that will support you in conducting more complete assessments of AI systems.

Trustworthiness is a property of a system that demonstrates that it will fulfill its promise by providing evidence that it is dependable in the context of use and end users have awareness of its capabilities during use (C. Gardner, et al., 2023). Users gain an understanding (or misunderstanding) of an AI system's capabilities and limits as they work with it, within their context. Their awareness of those capabilities may be informed through training, their direct experience, and their colleagues' experiences, and they will use this information to develop a justified level of confidence - or calibrated trust of the system.

When humans develop calibrated trust of the system – a psychological state of adjusted confidence that is aligned to end users' real-time perceptions of trustworthiness (C. Gardner, et al., 2023) – they can be productive using the system and use it appropriately. Calibrated trust is neither over or under-trust – it is a true understanding of the systems capabilities and limitation. When people over-trust a system, they are likely to use it for tasks it was not designed to complete. For example, a generative AI system will excel at tasks where creativity is desirable and is less likely to be successful for tasks requiring retrieval of specific wording. My colleague Robin may choose to use a generative AI system for this task due to positive experiences using it in other contexts. Robin is likely to find the system to be ineffective in this new activity and will potentially distrust the system as a result of this poor experience. Robin may be less likely to use it - even in situations where it could be helpful to them.

Smith, C.
Letting Go of the Numbers: Measuring AI Trustworthiness.
DOI: 10.5220/0012644300003654
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 11-12
ISBN: 978-989-758-684-2; ISSN: 2184-4313
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

Human trust is complex, transient, and personal – it cannot easily be measured, and we cannot cajole or coerce people to trust an AI system (much less anything else). Each individual has a different perspective and a different way to determine the trustworthiness of an AI system. Given this we need to consider the tasks that they are doing and how to best support those activities through their interface with the system. For example, the model's accuracy may be found through research to be helpful to the development of confidence for end users. The next step would be to start with prototyping what we think are helpful and effective designs and then conducting additional user experience research early in the development process to inform the overall development. There is a possibility that due to factors outside of our control, stakeholders may be wary or distrustful of the system regardless of our efforts, however, a system that is not built to be trustworthy is much more likely to fail, to increase risk, or to be harmful in other ways.

An AI system's trustworthiness can be measured both during development and during deployment. During the design phase of an AI system, program managers, human-centered researchers, and AI risk specialists should conduct activities to understand the end users' needs and anticipate requirements for AI trustworthiness (C. Gardner, et al., 2023). Teams can integrate user experience studies at the earliest stages with prototypes to gauge usability, explainability, and interpretability. Teams need to consider both the end user experience and the effect the system may have on others to prevent unintended harms. In this talk I'll discuss additional user experience studies that can be conducted to collect feedback on the components of trustworthiness, to inform and validate design decisions, and to explore the potential effects on society as the system nears release.

Doing AI well requires a reset – letting go of (some of) the numbers and learning new methods that provide a more complete assessment of the system. Trustworthy design considerations must be embedded from the initial planning stages through release and maintenance (C. Gardner, et al., 2023). With intentional work to create trustworthiness by design, organizations can capture the full potential of AI's intended promise (C. Gardner, et al., 2023).

REFERENCES

C. Gardner, K.-M. Robinson, C. J. Smith and A. Steiner, "Contextualizing End-User Needs: How to Measure the Trustworthiness of an AI System," Carnegie Mellon University, Software Engineering Institute, 17 07 2023. [Online]. Available: https://insights.sei.cmu.edu/blog/contextualizing-end-user-needs-how-to-measure-the-trustworthiness-of-an-ai-system/. [Accessed 07 01 2024].