DM23-0860

Script: Measuring Trustworthiness of AI Systems
SME(s): *Katherine-Marie Robinson, Carol J. Smith, and Alexandrea Steiner*
Interviewer/Facilitator: *Carol J. Smith*
Interview Conducted: *Friday, August 18, 2023 at 10:00 a.m. ET*

<Canned Intro>

Carol: Welcome to the SEI Podcast Series. My name is Carol Smith, and I am a senior research scientist specializing in human machine interaction in the SEI's Artificial Intelligence Division.

Today I am joined by Katie Robinson and Alex Steiner, both of whom work with me in the AI Division as design researchers. Today we are here to discuss trustworthiness of AI-enabled systems. We'll explore the practices that can support trustworthiness and whether metrics can be used to measure trustworthiness of AI-enabled systems. This is a particularly timely topic in an era when technological advancements in tools, such as ChatGPT, have opened the field of AI to so many uses - some good, some troubling.
Welcome to you both.

Katie and Alex: *Thank you.*

1. **Carol**: We always start by having our guests tell us about themselves. Katie, you are new to the podcast series. Why don't you start. Tell us a little bit about yourself, the work that you do here, and what brought you to the SEI.

Carol: Alex…

Carol: Now, I will tell you a little about myself…

2. **Carol**: Let's start by explaining what we mean by the word trustworthiness. A search of the literature on trustworthy AI reveals that authors often use the terms *trust* and *trustworthiness* interchangeably and use different definitions.
We define trust as a psychological state based on expectations of the system's behavior—the confidence that the system will fulfill its promise.
We'll be focusing on trustworthiness in our discussion –
Alex would you share how we're using the term trustworthiness?

Alex: That definition of trust is really important in contrast and comparison to our definition of trustworthiness. As you described, trust is state—and it's a human state, if I have trust in something or someone, it means that I believe they will act in a way I expect. Trustworthiness, though is, as we've defined it, can be a characteristic—it can be a characteristic of a person, and of a system. If trustworthiness of a system is high, that means that its demonstrating that it will fulfill it's promise — and it's demonstrating that by providing evidence that it is dependable in the context of use, and that end users are aware of its capabilities during use. *Trustworthiness* is something that we can attempt to break into measurable aspects and quantify.

3. **Carol**: Tools built on large language models, such as ChatGPT, have brought about new advancements that, 5 years ago, would have seemed unimaginable. Paul McCartney recently used AI to complete the last Beatles song using John Lennon's voice. AI tools such as Github's Copilot are being used to help developers and engineers generate software code. There are also examples of people over-trusting these systems – for example a lawyer in New York used ChatGPT for legal research, which the judge pointed out were "bogus judicial decisions with bogus quotes and bogus internal citations" resulting in significant embarrassment and potential disciplinary actions against the law firm. At the same time, many researchers worry that these tools are advancing too quickly in the wake of dangerous ,

   Katie: Katie, given these developments, what is the general state of trust of AI at the moment?

4. **Carol:** To get to the nuts and bolts of AI trustworthiness, Alex, what are the measurable aspects of trustworthiness?

   Alex: As I mentioned before, trustworthiness can be broken down into measurable aspects—these are the building blocks of what make a system trustworthy. There's a number of aspects that make up trustworthiness, and these include
   validity and reliability
   safety, privacy
   security and resiliency and
   accountability, transparency

Some of these components can be assessed through quantitatively or qualitatively.

5. **Carol**: There are so many aspects of trustworthiness, how do you balance them all?

   Some of these elements of trustworthiness might appear to be in tension or conflict with each other: we can look at the example of transparency and privacy.

   To have transparency, we want to provide information describing how the system was developed, but when we consider privacy, we know that end users should not necessarily have access to all the details of the system. So we can how there may be necessary tradeoffs between those two characteristics.

   And we can evaluate a system that performs well across each of these components, and yet users may be wary or distrustful of the system due to the interactions they have with it. So trustworthiness can be a very tricky thing to measure.

   Negotiations among the team are necessary to determine how to balance the aspects that are in tension and understand what tradeoffs may need to be made as we prioritize the system's trustworthiness.

6. **Carol:** Katie, in a recent blog post we outlined questions that organizations should ask before determining if they want to employ

a new AI technology. What are key questions you encourage people to explore?

What is the intended use of the AI product?

- How representative is the training dataset to the operational context?
- How was the model trained?
- Is the AI product suitable for the use case?
- How do the AI product's characteristics align to the responsible AI dimensions of my use case and context?
- What are limitations of its functionality?

What is the process to audit and verify the AI product performance?

- What are the product performance metrics?
- How can end users interpret the output of the AI product?
- How is the product continuously monitored for failure and other risk conditions?
- What implicit biases are embedded in the technology?
- How are aspects of trustworthiness assessed? How frequently?
- Is there a way that I can have an expert retrain this tool to implement fairness policies?
- Will I be able to understand and audit the output of the tool?
- What are the safety controls to prevent this system from causing damage? How can these controls be tested?

7. **Carol:** We have some other research efforts underway within the SEI's AI Division that are also looking at how to measure AI trustworthiness. Alex would you share about those?

On fairness: Identifying and mitigating bias in machine learning (ML) models will enable the creation of fairer AI systems. Fairness contributes to system trustworthiness. Anusha Sinha is leading work to leverage our experience in adversarial machine learning, and to develop new methods for identifying and mitigating bias. We are working to establish and explore symmetries in adversarial threat models and fairness criteria. We will then transition our methods to stakeholders interested in applying ML tools in their hiring pipelines, where equitable treatment of applicants is often a legal requirement.

On robustness: AI systems will fail, and Eric Heim is leading work to examine the likelihood of failure and quantify the likelihood of those failures. End users can use this information—along with an understanding of how AI systems might fail—as evidence of an AI system's capability within the current context, making the system more trustworthy. The clear communication of that information supports stakeholders of all types in maintaining appropriate trust in the system.

On explainability: Explainability is a significant attribute of a trustworthy system for all stakeholders: engineers and developers, end users, and the decision-makers who are involved in the acquisition of these systems. Violet Turri is leading work to support these decision-makers in meeting purchasing needs by developing a process around requirements for explainability.

8. **Carol:** What about trustworthiness of newer systems like large language models (MidJourney, Dall-E, ChatGPT, etc.)? (Katie)

9. **Carol**: Let's talk about the future of making AI systems trustworthy, what do you see happening in this space? (Katie & Alex)

   Alex: Push for quantitative measurement in trustworthiness – and components of trustworthiness.

Carol: Katie and Alex thank you for talking with us today. We will include links in the transcript to resources mentioned during this podcast.

Finally, a reminder to our audience that our podcasts are available on Soundcloud, Stitcher, Apple Podcasts, and Google Podcasts as well as the SEI's YouTube Channel. If you like what you see and hear today, give us a thumbs up.

Thanks again for joining us.

<Canned Outro>