Measuring the Trustworthiness of AI Systems

featuring Katherine-Marie Robinson, Alex Steiner and Carol Smith

Welcome to the SEI Podcast Series, a production of the Carnegie Mellon University Software Engineering Institute. The SEI is a federally funded research and development center sponsored by the U.S. Department of Defense. A transcript of today's podcast is posted on the SEI website at sei.cmu.edu/podcasts.

Carol Smith: Welcome to the SEI Podcast series. My name is Carol Smith, and I am a senior research scientist specializing in human-machine interaction at the SEI 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 the trustworthiness of AI-enabled systems, and we will explore the practices that can support trustworthiness and whether metrics can be used to measure the trustworthiness of AI-enabled systems. This is a particularly timely topic in an era where technology advancements and tools such as ChatGPT have opened the field of AI to so many uses, some good and some troubling.

Welcome to you both.

Alex Steiner: Thank you.

Carol: We will start by having our guests tell us about themselves. Katie, you are new to the podcast series. Welcome. Why don't you start by telling us a bit about yourself? The work you do here, and what brought you to the AI?

Katie Robinson: Perfect. Thanks, Carol. Hi everyone. My name is Katie Robinson. I joined the SEI at the end of September 2022, so I have been here almost a year. Since joining the SEI I have worked on a variety of different projects with Carol and Alex and other members of the AI division in responsible AI. So whether that is looking at tools that we can implement or curriculum as ways to introduce responsible AI to people. I have been involved in many of these different projects. My background is in engineering, and I did my masters in the ethical implications of AI. I am incredibly excited to be at the AI division and at the SEI as well.

Carol: Excellent, and Alex?

Alex Steiner: Yes, thanks, Carol. I joined the AI division as a design researcher in 2020. I plan and implement forward-looking design research that supports our current technical research work and our strategic vision. My background is in communication design and human-centered design. I work closely with customers, researchers, and engineers to understand challenges and identify requirements in order to translate those needs into research strategies, prototypes, and project plans. A lot of my work right now centers on responsible AI engineering as we, as a division, look to try to develop the discipline of AI engineering. This includes topics such as how to make AI capabilities trustworthy and usable, which is something we are here to talk about today.

Carol: Excellent. Yes, and just a little bit about myself as well. I joined the SEI four years ago. I have spent my career working to improve complex technologies and helping to make them more usable and useful to the people using them. I hold a master's degree in human-computer interaction. I have been doing this type of work for over 20 years across industries. I have been leading research on AI systems since 2015. I have had the pleasure to work with you both as well as our other researchers in the AI Division during that time.

Let's start by explaining what we mean by the word *trustworthiness*. A search of the literature on *trustworthy AI* reveals authors often use the terms trust and trustworthiness, interchangeably, and they use different definitions. We define trust at the SEI in the AI Division as *a psychological state based on expectations of the system's behavior, the confidence that the system will fulfill its*

promise. We will be focusing on trustworthiness in our discussion today. Could you share how we are using that term *trustworthiness*?

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

Carol: Excellent, and, Katie, did you wanna add anything there?

Katie: No, I think that was perfect. I think we have highlighted both of those definitions in our blog. So great.

Carol: Tools built on these large language models I mentioned, such as ChatGPT, have brought about new advancements that five years ago would have seemed unimaginable. Paul McCartney fairly recently used AI to complete the last Beatles song using John Lennon's voice. AI tools, such as Github Copilot, are being used to help developers and engineers generate software code.

And there are examples of people over-trusting these systems unfortunately. For example, a lawyer in New York used ChatGPT for legal research and presented that in a case, and the judge pointed out there were bogus judicial decisions with bogus quotes and bogus internal citations. That resulted in significant embarrassment and potential disciplinary action. At the same time many researchers worry that these tools are advancing too quickly in the wake of surprising and sometimes dangerous behaviors as well, such as the use of deepfakes to impersonate global leaders. Katie, given these developments, what is the general state of trust in AI at the moment.

Katie: That's a great question, Carol. Thank you so much. I think with the emergence of new AI technologies, we are being exposed to a wide variety of capabilities. These technologies will continue and have brought some very exciting opportunities and challenges and changes. For example, we can look to the cases where AI is being used to preserve languages. If we look at New

Zealand, Chile, Kenya, AI is actually being used to preserve those indigenous languages, to help people continue to learn them and to continue to progress them as well as future generations, so people can maintain these languages going forward. But on the opposite side of the spectrum, we are seeing some really concerning uses and outcomes of these technologies in different domains, such as education, art, and programming to your point. So if we look at the tool like Copilot, a code suggestion tool, we are starting to see—and people are starting to have the conversations about how the code it is suggesting seems very similar to code that people have already written, but people aren't receiving credit. We are starting to see that a lot in other domains as well, such as the text-generation tools and art-generation tools too. Overall, optimism is high, and I think it will be ever present. We are starting to see those understandable concerns that raise questions about how trustworthy these systems are. Therefore I think the general state of trustworthiness of AI currently depends on a few factors: how familiar people are with these tools, the context they are using the tools in, the impact the tools that are having on them in their day to day life, and just how much they understand them in general. I think, depending on the people you talk to, they might have different answers to those questions which really fluctuate the level of trustworthiness. I think that points to a term that we use in the blog post when we talk about the we talk about the calibrated level of trust, and we think about how trust is fluctuating. If people are approaching trust with successful interactions, they are having great ways of interacting with these systems, of course their trust is going to go up. But if they have these interactions that are negative, they don't understand it, they get frustrated, trust is going to go down in these systems as well.

When we talk about in general the current level of trust and the state of trustworthiness in AI, I think it really depends on the person, the impact is having, and the context they are using it.

Carol: Excellent, and to get us to the nuts and bolts of AI trustworthiness out. Trust is going to go down in these systems. Alex, what are the measurable aspects of trustworthiness?

Alex: As I mentioned before, trustworthiness can be broken down into those measurable aspects. These would be the building blocks of what make a system trustworthy. There are a number of aspects that make up trustworthiness. These might include things, such as validity and reliability, safety and privacy, security, resiliency, accountability, and transparency among a number of other aspects as well. Some of these components can be assessed through…Whether the are through quantitative or qualitative

measurements, we can try to make a measurement of each of those aspects so that we can build up to give us an answer about how trustworthy a system might be.

Carol: Excellent. We do talk about this in the recent blog, *Contextualizing End-user Needs: How to Measure the Trustworthiness of an AI System*, and we will include a link in the transcript as well. Alex, building on that there are so many aspects of trustworthiness, how do you balance them all?

Alex: Yes, that is a great point. Some of these elements of trustworthiness might sort of appear to be intention or conflict with each other, and they might actually be intention or conflict with each other. We can look at the example of transparency and privacy as a good example. To have transparency, we want to provide information describing how that system was developed. When we consider privacy, we know that the end users should not necessarily have access to all the details of how we have built that system. We can see how there may be necessary trade-offs between those characteristics. As we evaluate a system that performs well across each of these components, users still might be wary or distrustful of a system due to the interactions that they have with, and that kind of goes to Katie's point about calibrated level of trust. Trustworthiness can be a pretty tricky thing to measure. So negotiations among the team are truly necessary to determine how to balance those aspects of trust that are the intention, and understand what trade-offs we might need to make as we prioritize the system's trustworthiness.

Carol: OK. Katie, in that recent blog post, we outline the questions that organizations should ask before determining if they want to employ a new AI technology. What are key questions that you think encourage people to do that exploration?

Katie: Yes, so in the blog post we break down a variety of questions into these two different groups, where one group is looking at what is the intended use of an AI product, while the other group is [looking at] what is the process necessary to audit and verify the AI product performanc. When we are looking at intended use, we are really urging organizations to think about the context the AI product will be used in, how the AI product was trained, and the products capabilities and limitations. We believe these questions are really important to consider as they can lead to further questions and discussions about the possible impacts that the AI system may have on end users, stakeholders, or just the organization more broadly. Now when we are looking at questions relating to auditing and verifying the AI

product performance, we suggest organizations look to the products measurable characteristics, for example, maybe the implicit biases that are embedded in the technology, the product performance metrics, or the interpretability of the output. Additionally, organizations should also consider questions about how the product will be monitored and maintained once it is deployed, as well as the feasibility and opportunities for retraining and reevaluating the product in the future. Now overall we are encouraging organizations that want to employ new AI capabilities to ask questions to their stakeholders, meaning their employees, their customers, that they have or just their partners or anybody else who interacts with them, what they would want to know before using, implementing, integrating these new capabilities into their workflows.

Carol: When we are looking at these questions, these are more focused on people who are acquiring these systems. Is that right?

Katie: That is correct. Yes. When we are looking at these questions we are encouraging people who are adopting and acquiring these systems and having these questions and having the conversations that they need to with the people who are responsible for developing and designing these systems as well as the people who are going to be using these systems, so trying to bring that conversation together.

Carol: Yes. That is a nice segue to talking about some of the research efforts that are underway within the SEI AI Division that look at how to measure AI trustworthiness and particularly on explainability, which is an aspect that we have been mentioning throughout for both the end users the people who need to determine what to buy and other questions and understanding that people need to have trustworthiness. Alex, would you share about those?

Alex: Yes, absolutely. There is a lot of really interesting work going on here at the AI Divison. We have Anusha Sinha, a machine learning research scientist. She is leading work to leverage our expertise in adversarial machine learning and to develop new methods for identifying and mitigating bias. Identifying and mitigating bias in machine learning models will enable the creation of fairer AI systems. We will transition our methods to stakeholders interested in applying ML Tools in their hiring pipelines where equitable treatment of applicants is often a legal requirement. Then Dr. Eric Heim is leading work to examine and quantify the likelihood of failure. Users would be able to use this information as evidence of an AI system's capability within that current context, making that system more trustworthy. The clear communication of that information, it supports stakeholders of all types in maintaining

appropriate levels of trust in the system so that they can understand how likely is failure of that system and how can we quantify that.

Finally, Carol, you mentioned explainability. Explainability is a significant attribute of a trustworthy system for all stakeholders, engineers, and developers and users, and decision-makers who are involved in the acquisition of these systems, all should be considering explainability.

Software developer Violet Turri is leading work to support these decision makers in meeting purchase requirements, purchasing needs, by developing a process around requirements for explainability. Like I said, there is a lot of promising work in the area of trustworthiness going on at the AI Division, and it is really great to have all those smart people working on tough topics.

Carol: Well said. What about trustworthiness of newer systems? These large language models we have been talking about briefly: Midjourney, Dall-E2, ChatGPT, etc. LLMs. We hear about this every day now it seems. Can you tell us a little bit about that, Katie?

Katie: Yes, of course. I really like this question, because I think it sort of goes back to what we were discussing earlier where it is dependent on the person. For example, when we think about new advancements in AI, to go to your examples Carol, an example that was picked up very quickly I think by younger people was ChatGPT, especially in domains such as education. We saw ChatGPT as a way to generate prompts, to write essays or theses maybe, actually to write those essays and theses or just to answer random homework questions that people could throw at it. I think what is really funny, and the reason that we know about all these use cases, is people actually handed in this work they trusted it enough without checking it themselves. When you actually go back and see what people have handed in, you see that there were erroneous errors in the homework questions, or the topics didn't make sense. There were fake citations and references. So we are really seeing the trust that people had in these technologies and then what that actually means in turn. When we are talking about trustworthiness of these systems and with everything that has occurred previously, I think that these systems do have a long way to go. When we consider the different aspects that influence trustworthiness, as Alex and just discussed, we see that developing a trustworthy AI system is more than just telling users how much data you used or how you trained the model, or what you did throughout the process. It is really about considering and addressing the different aspects that are going to be affecting people using the system and how those aspects should be addressed throughout the different stages:

planning, design, development, and maintaining the phases of those AI systems.

Carol: Thinking about the future, of making AI systems trustworthy. What do both of you see happening in the space? Where do you think we are going next?

Alex: I will start. I am hearing an ask for more quantitative measurements. I have heard that quite a few times recently. In a research area such as responsible and trustworthy AI, a lot of times we have to evaluate, based on qualitative metrics, and measurements. Understandably stakeholders want metrics that they can more readily understand and interpret. To me that signals a need for us to think about ways we can develop quantitative numerical metrics for responsible AI including trustworthiness. I am certainly interested to see where others in academia, government, and industry identify those measurement opportunities. I believe that's something that this team will continue to work on as we move forward.

Carol: For sure. Katie?

Katie: Yes, just to add on to that, I think what Alex said was perfectly on the nose. I think it is really encouraging to see what people are doing in all of these different spaces. We see conferences or just get-togethers and just mind melts of people coming together with ways to address trustworthiness and responsible AI, and how we can progress that forward. I think the AI Division at the SEI—of course Alex highlighted with all these different researchers doing all this different work—is really exciting and really encouraging. Moving forward.

Carol: Yes, it is a really interesting and exciting time to be doing this work. I am excited to be working with you both and all the other individuals that we have mentioned. Katie and Alex, thank you for talking with us today. We will be including links in the yranscript to resources mentioned during this podcast and a reminder to our audience. There are podcasts are available on SoundCloud, Apple Podcasts, Google podcasts as well as the SEI's YouTube channel. If you like what you hear and see today, give us a thumbs up.

Thanks for joining us. Have a great day.

Thanks for joining us. This episode is available where you download podcasts, including SoundCloud, TuneIn Radio, Google Podcasts, and Apple Podcasts. It is also available on the SEI website at sei.cmu.edu/podcasts and the SEI's YouTube

channel. This copyrighted work is made available through the Software Engineering Institute, a federally funded research and development center sponsored by the U.S. Department of Defense. For more information about the SEI and this work, please visit www.sei.cmu.edu. As always, if you have any questions, please do not hesitate to email us at info@sei.cmu.edu. Thank you.