Trust and AI Systems

featuring Carol Smith and Dustin Updyke

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.

Dustin Updyke: Welcome to the SEI Podcast Series. My name is Dustin Updyke, and I am a senior cybersecurity engineer at the CERT Division of the SEI. I am also a graduate student at Carnegie Mellon's Department of Philosophy. The concept of trust is at the forefront of AI system concerns, and in 2021 the National Institute of Standards and Technology [NIST], whose research often helps us shape government policy, released an approach on how organizations can identify and manage bias in AI. NIST has stated that alongside research towards building trustworthy systems, understanding user trust in AI will be necessary in order to achieve the benefits and minimize the risks of this new technology. Joining me today to talk about trust in AI systems is Carol Smith, a senior research scientist in human-machine interaction at the SEI's AI [Artificial Intelligence] Division.

Hi, Carol. Welcome.

Carol: Hi. Thanks for having me.

Dustin: Let's start by telling our audience about ourselves, what brought each of us to the SEI, and what kind of work we do here? I will let you start, Carol.

Carol: I joined the SEI about two and a half years ago. I had been previously working in industry for 20 years primarily doing human-computer interaction [HCI]. I have a master's degree in HCI and have worked across many different industries. The past seven years or so, I have been working in artificial intelligence and more complex systems. I also teach in the Human-Computer Interaction Institute here at Carnegie Mellon. Right now I am working in the AI Division here at the SEI and working on really the connections between humans and machines and figuring out how to help make these systems in ways that humans can trust and are willing to be responsible for them.

Dustin: I also have 20 some years of industry experience before I came here in all sorts of different kinds of programming, systems development, applications, web technologies, that sort of thing in consumer and B2B [business-to-business] type organizations, but also healthcare was my last stint. I have been here about five years, [and] jumped into cybersecurity. We do different types of exercises and training for different DoD [Department of Defense] units. I love it. It has been really interesting so far. Then I decided to go to school here, and I have been spending a lot of time on trust in particular. That is part of my research, and so I am really excited to talk about that with you today.

Carol: Yes. Same here.

Dustin: Maybe we should start by clarifying what we mean when we use that term *trust*, particularly in terms of artificial intelligence and the autonomy that those systems might have. I think that for our audience we should probably define the sort of scope for this discussion. Are we talking about how we trust those AI systems or computers in general? Whether those systems trust one another? How we might build such a system or any or all of the above?

Carol: Yes, I do think it is a combination of things. From my perspective, I certainly am more concerned with making systems that are trustable, trustworthy, by people. So thinking about what is it that needs to be provided as far as evidence for the system to be trusted. What are its capabilities? What integrity does it have as far as how it was built, and how it is being maintained at a level that is appropriate for individuals. Then, by giving them that information, they can then determine if there is risk and how much risk and what the situation is. It is never a situation where we are going to be or should be building systems that are one hundred percent trusted because even with other people there is context. There are situations where things are going to change, and the system certainly is going to change, particularly the AI systems. So we need to make sure that they have a corresponding change in that trust. The literature points to that as calibrated trust, the idea of a balance between not overtrusting to the point where automation bias is something that people tend to overtrust systems. And not distrusting them or rejecting them, but rather calibrating that trust based again on the evidence that they have the capabilities, that the system has integrity, and is built in such a way that it can be trusted.

Dustin: You hit on something there that I don't think is often distinguished. There is trust, and then there is explicit mistrust or distrust of a system. Can you talk a little bit more about that?

Carol: Yes, we see this, particularly in industries where people are concerned about their job security or where they feel that the humans are much more talented or skilled in a particular area more so than the machine. In those cases, people just have a natural distrust of the systems because of their concerns about their personal livelihood or the work that they are trying to do. Then you will have distrust because a system does not perform the way people expect it to.

For example, I was talking to someone recently about security systems and if the security system is overly… if it is providing too many false reports that, you know, every single shadow that moves is reported; people quickly begin to distrust that type of security system and similarly with a computer. If it is constantly giving incorrect information, poorly timed information, those types of activities are then incorporated into the person's idea of the system and their ability to trust. So the trust will naturally go down based on that kind of reliability and effectiveness of the system.

Dustin: In the context of establishing trust with a system, maybe it is known, maybe it is new to that person. There have been lots of reported examples of bias and mistakes in AI systems, such as, I don't know facial recognition, predictive policing, those sorts of systems. The relevant question that you get often at the sort of holiday dinner table is, *Should we actually even trust these systems?* Like, *How do we start to engage with a system like the example of predictive policing or facial recognition?*

Carol: Right. Yes, and my answer would be no in this case with the current systems because the data that they have been trained with is so unrepresentative of the larger population. That is what is leading to those problems with the system because of the history that the systems are built on. So there is a lot of bias in historical data. There is bias in all data because humans are the ones creating it, but particularly when you are looking at historical patterns where we know that there has been racism or some type of unfair behavior that has been documented in that data; and then using that to create a new system that you are expecting to be without that bias is just not possible. The systems can only know what we provide them with, and you can't really take that kind of bias back out of the system, unfortunately. But on the other hand, there are systems that are very effective, so it really is very dependent on the type of system, the criticality of that system, and of course the data that the system is built on. We can't have artificial intelligence without data. If the data is problematic, particularly for the purpose it is being used for, it is going to create a system that is flawed and potentially adds and creates more harm than previously there even was in the system.

Dustin: I think you gave a couple of examples, but are there other roadblocks in either establishing or increasing human trust in AI systems that maybe some of the examples that we gave are problematic or suspect at the very least.

Carol: Yes, the relationships that people have, and I think you can add to this as well, really are important in their influence of the system. So if a person is working in a situation where they don't have a feeling of psychological safety, where they don't feel that they are trusted, or that they don't trust the other individuals around them, they are much less likely to have trust of an AI system or any other system. So those personal types of factors, their previous experiences with similar systems, things that they may not even be aware of, can affect their level of trust.

Additionally, as they work with a system ideally, they do attain that calibrated level of trust, but if there are negative experiences with the system that trust may go down. Or if there are overly positive experience—and that actually can be more dangerous—where someone actually becomes much more trusting of a system than is recommended or that is reasonable. We can see, unfortunately, accidents occurring because people are less observant of a system or less in tune to how the system is working. So the system begins to work in ways that are unanticipated or unhelpful, and that can be problematic. All of this points to, of course, the design of the system. If the system is designed with humans in mind, and it is designed in a way that really conveys the right information, allows them to have control over the system; to understand the information at the appropriate time, and that is dynamic in the way that the AI system is dynamic and really responds to the humans who are using it in good ways or helpful ways. Also when the system also is supporting the work that they are doing, then we get to the point where humans and machines will be able to team. We are still a little bit away from that. The current AI systems are very narrow in application and use. But I am looking forward to the day when we really are able to effectively team with these systems and have a little bit more autonomy but still maintain control so that we can trust them.

Dustin: One complexity that I am exploring in my graduate work is that humans, we historically have thought about how tools best suit a job, right? And computers, or machines in general, are really no different and so, as a result, I have come to trust this machine, this computer, because it is really good at this repeated task that I often ask it to do. For applications like a calculator, the inputs are always the same. The answers are sort of always the same across space and time, and it would be pretty easy to come and trust that sort of tool. Then, in a transitory way, you see that I get a certain result, you get the same result, right, and we start to build that sort of trust based on the computer being a tool. But with AI, you sort of blur the lines of what the two actually can do for me, right? I can ask certain applications all sorts of questions and they increasingly answer more and more types of questions. That I think really gets at the computer moving from a tool to an actual teammate. To your point about human-machine teaming, that is a sort of different kind of trust than this static tool that does certain things to this teammate that can do many different things. Part of the challenge is figuring out what that teammate might do and what they do best and how to engage with that tool. So do you have any thoughts on that? Or any experience in that particular area?

Carol: Yes, it is a really difficult balance. These are really complex questions, particularly when we start thinking about physical systems, but also even systems that an analyst might use at a desk. The way to understand what is appropriate for that particular situation, so understanding both for new users and users who have been using the system for a long time, but also thinking about the time cycles of the work. So if it is very short iterative work or very short individual like discriminate pieces of work, those may be very different interactions than someone who is

working on something over time; where they are adding information, where the system is providing new information or new information is integrated; and then a narrative or a larger story is being developed through those interactions. For example, healthcare is a good one where you are looking at potentially a long-term care for a patient and how maybe there is a new illness that is added to the complexity of the situation or new medications and interactions that you want to look at and thinking about how many different individuals may be caring for that patient and how much information needs to be in that system. Then when is the system making recommendations? When is the system pointing to new research? When is the system accepting new information about that patient, and how do you manage, protecting their information, protecting other people's information, and all of those complexities? So really, as with any system, you do need to think about how is this AI, which may be a very small piece relatively of a larger system, how are all of these pieces of information being brought together? How are the individuals who are interacting with it going to understand what is there and what is available to them and what someone else may have, and also the turn-taking aspects. So, if the human is in control primarily, what does that look like versus when the system is in control? Particularly with robotics, this can be a really important aspect to consider.

Another aspect of this work is thinking of how are humans remaining accountable and responsible for the systems. These systems don't have rights and responsibilities, nor do I think they should. So really considering how do we make sure that there is someone who is in control and responsible for the system during those various points of interactions. Then doing speculative work to make sure that the systems, risks, and benefits have been identified, and that they are really clear to everyone on the team, so that the team really understands not just the complexity of the system, but also the potential harms that can occur. So that they are thinking about those as they are developing the system and taking steps to prevent, when possible, those harms or mitigate them as appropriate. Then I mentioned security with regard to privacy, but also making sure that the system is respectful of individuals' information. Even when people are willfully providing information, that we are not collecting more than we need. As with any secure system, there is no reason for us to keep additional information about individuals. Particularly with an AI system, it becomes more important because the combination of multiple data sets can potentially create new information that creates more problems. Finally, looking at how to make sure that the system is honest, and that people understand that they are working with a computer, that they are working with a system, and usable, really human-centered design. Understanding the people that are going to be using the system and making sure that it is created in such a way that they do have that calibrated trust because they understand it is designed for their use.

Dustin: One of the things that is interesting is that if we were to stand up a system…Say we were on a team that is making decisions based on some amount of data, and we decide to

automate that and start to move towards an AI solution, do we have frameworks or tools to help think about how to transition that from an early application that maybe only does certain things? As we expand its functionality, the expectations of the users and the people impacted by that AI system change over time, right? Do we, as the SEI, have different resources and tools or maybe frameworks for how to think about implementing such a solution and then having people transition with it as it learns to take on new functionality?

Carol: Yes, we do. We have quite a few tools, and we are working on developing more. One of them I was actually mentioning some of the aspects of is the framework for designing trustworthy AI that we have available on the website. That lists those areas of *accountable to humans*, *cognizant of speculative risks and benefits*, *respectful and secure*, and *honest and usable*; and provides really a checklist, but it is not really a checklist. It is meant to prompt conversations and help people start to determine what work they need to do to make the system responsible in that way.

Then I am working with the Defense Innovation Unit. They have put together the responsible AI guidelines and worksheets. Those, particularly for government agencies, can be very helpful for them to again begin those conversations really understanding what it is that they are building, what the ways that they are going to measure the improvements. So looking at the existing systems and processes and using that to then set goals for the new system as far as improved performance or effectiveness, efficiency, whatever it is that they are looking for. Really helping them to set out some goals and ways to measure that. Also, that has three different phases: the planning phase, the development phase, and then the deployment phase, which really I think gets at what you were talking about, which is this constant work that does need to be done to make sure that the system is still performing as expected and that it is still providing the benefits to the end users or the affected communities as expected. That really entails quite a bit of work especially when new data is introduced to the system because the system may or may not perform in the way that it is expected to. Doing continuous work as far as evaluating its performance, looking at the data and the results, and also, of course, interacting with the end users and making sure that they are still at an appropriately calibrated level of trust due to the system's performance, and that that has not degraded or been inflated by unexpected situations.

Dustin: As an example, in healthcare, if we build a healthcare application, those users might be very different also, right? There might be hospital administrators. There might be doctors and healthcare professionals. There might be patients. They all have different expectations and different pathways to trust as well. Correct?

Carol: Yes, definitely. Yes, and really thinking about those individuals, their needs, different levels of fidelity of information, different types of information they may need access to, all of it may be contained in the same data sets, but very different access to that information. So that

should all be considered as the system is built, so that they do have the right experience, and so that you are not making things worse with the system but rather really making improvements that you hope to do.

Dustin: You seem so positive about the future of AI. It is pretty exciting. Here would be my example. In 1983, War Games comes out. If you are not familiar with the movie, the sort of fundamental question that the movie keeps dangling in front of you is should the characters in this movie trust what the computer is telling them. Is this World War III or is this a computer simulation? In some respects, things have not gotten much better. In fact, they have in some respects gotten much harder to determine whether I should trust a system that I am interacting with. Because there is a mix of things going on there, there is technology, there is human psychology, there are social issues. It seems like real progress seems sort of uniquely daunting in the AI space for some of these systems. Are there some fundamentals or some recent progress that you can point to that make you so optimistic in this space?

Carol: Yes, the more recent discussions about responsible AI and human-centered AI definitely make me very hopeful. But to your point it, all has to be done with caution. While I am very hopeful that these systems are going to be extremely beneficial in the future, I also feel that too little work has been done previously to really pay attention to things like trust. From a human perspective, in the past, there has been a lot of talk about humans just trusting, a hundred percent trust of systems, and now people are talking about zero trust. Neither of those for an end user is proper or appropriate. It has to be calibrated. It has to be made for the individuals that are using the systems. There is a huge amount of hype about AI, and we are certainly not meeting those expectations. Nor will we, I don't think, for quite a while, but I am very hopeful because we are starting to accept that these are more difficult problems than people anticipated, that making systems that work with humans is a hard problem. The AI systems themselves are not easy, but it is even harder to take that AI system and integrate it into an existing social situation or an existing business situation because the humans are even more complex. Accepting that and embracing that because that is what makes us awesome is really the part that I am most excited about, that people are really getting to the maturity in this work to understand that that is the problem to solve, that that is where the real work comes in.

Dustin: Something you said that made me think about our phones in general—they come with a certain series of applications already installed, and maybe we trust that, right? Maybe those applications are very simple, static, calculator-like applications. But then as we add a new application on to that device, it sort of changes its overall nature, the way we interact with it, maybe the questions that we can ask it. I often get sort of muddy responses as to what part of the phone I trust, right? That is sort of an interesting problem in and of itself. There are lots of

different applications running on that device, and you may trust them at different levels, but a lot of times people just see it as this one sort of static object.

Carol: Yes, and it is easy to think, *Well, of course I trust my computer. Of course, I trust my phone*. But to your point, if you download the wrong app or the wrong information or click the wrong link, all of a sudden, your entire system can be compromised. That balance is really hard for people to understand, and also if AI systems are introduced as just computers that is minimizing the potential harm that can occur. We are in this really interesting moment where a lot of different types of technology and different aspects are coming together, and it is really hard for people who are very knowledgeable about this stuff to differentiate, much less for the average person to really be able to even begin to approach these problems. They don't have time. So it is our responsibility as technologists to do a lot of that work for them and to help to protect them by making systems that are responsible, that are trustworthy, that help them do the things they need to do and don't do things without their permission that aren't a benefit to them.

Dustin: If I can switch gears for a second, I am just wrapping up a course in normative ethical theory this semester. It is exciting because there are a lot of different engineering students in there. I think the university has always offered not only specific ethics courses but ethics for engineers as well. I think what we have seen is a lot more interest in those types of courses. So, I have to ask you your experience in teaching at CMU and maybe some of the things that we have talked about today.

Carol: Yes, I have seen the same thing. Students and people in the community, in general, are much more interested in these types of conversations now because we have seen the harms that can occur when people really aren't thought about when systems are rolled out and when those implications aren't considered ahead of time. So more and more people are aware of the fact that it is important to think about these types of aspects as you are building the system. Not adding it on after the fact, but really from the very beginning thinking about what the potential implications are, the unintended consequences, the harms that can occur. Defining those as best you can, and then again, doing that work to prevent it. Even at the very smallest levels of code, even when you are just bringing in an API that is well respected, being able to be knowledgeable about the known issues with those pieces of information, particularly thinking through what happens when we bring them together is important. I am thrilled that more people are really thinking that through. In one of my classes, we have been doing a lot of discussions around the implementation of these new technologies and having them think about not only the intended use and why they are creating it, but also having them do a brief exercise about what harms could occur because of the system that they have built. When they bring the data and the interactions together and having them think that through creates, hopefully, a familiarity and a comfort with

those kinds of conversations that they will have in the future when they are on teams making new products.

Dustin: I don't have answers for this, but I am sort of trying to put together all the questions for how adversaries might specifically target trust. You can target to steal information from a system, to wreck a system, or something like that. But to actually want to get to a point where the system appears to be operating as normal, users are using it, but it is increasingly giving them the wrong information. That might change their behavior in terms of they come to not trust it anymore because what they are seeing in reality just doesn't match what the computer is telling them. Or, it might actually start to sway their reality in ways that could be harmful, confusing. I don't really have specific examples of how you might do that, but I don't have to think too hard about some of the things that we've seen in social media dating back to 2016 and those sorts of examples, it seems to me to be a real potential. If someone would take the time to do that in a way where the administrators of that system, the users of that system, wouldn't notice it, that could be hugely dangerous.

Carol: Yes, yes, that is a great point. When you brought up *War Games* that is what I was thinking too because they couldn't tell the difference. They couldn't tell what the truth was, and there was a real-life scenario that I don't know if the movie was based on it or not, but where there was an incident in real-life and the human thankfully did not trust the system in that situation. That is really important to maintain that thoughtfulness, that mindfulness that humans have of, *Hmm, you know, my gut is telling me something's not right here*. But how do we get them to do that? How do we help them to say, *Wait a minute, let's pause for a second. Is that really accurate? Is that really reflecting your understanding of the situation?* And figuring out at what moments those are needed. Are they from the outside? Is it a manager? Is it someone else that needs to intervene once in a while, just make sure that everyone is paying that kind of attention? It is really important that people remain appropriately skeptical but trusting enough to be able to be functional. Much like if you drive a vehicle, and you expect it to turn on every day when you go out to it. There needs to be some level of base expectations. But also, you know, the engine light, something, some kind of indication, but more meaningful than the light that everyone knows.

Dustin: I think a good example of that is Stuxnet, and it is not talked about in this particular way. But you have the Iranian nuclear program using older-generation centrifuges, right? Because of sanctions, they can't get the latest and greatest. So they are using these industrial control systems to monitor these cascades over time as sort of a preemptive maintenance partner, right? So you are the administrator of that network, and every so often a setting comes up that you say, *Oh, maybe that machine needs maintenance*. After Stuxnet is introduced on that network, slowly over time machines start to fail, but your computer is not giving you that same

read out of information. At some point you are not trusting what it is telling you because it says everything is OK, but these machines are failing at a higher rate than we expected. Obviously, we don't know all the details of what happened there, but you suspect that they shut everything down in terms of their nuclear program, even the adjacent nuclear power plants that they are building. I suspect that is because they really didn't know what was going on initially. The computer is telling me one thing, but things are failing. That is the sort of example I think that you can take to social media networks, healthcare systems. It can be a real danger.

Carol: Yes, and particularly if something really is wrong, and they are not aware, it is the system not being aware, but it is also the humans not being aware and everything in between. Yes. Huge challenges and very specific to the context of the individuals. A nuclear situation is very different from a robot or an autonomous vehicle versus my financial system or my stock recommender. These all have very different risk profiles and different effects on myself and society and true living beings. It has to be appropriate for that particular context.

Dustin: It sort of connects back to your comments about different users and different expectations and roles as well.

Carol: Yes.

Dustin: Well, Carol, thank you for talking with us today. We will include links in the transcript to resources that we mentioned throughout this conversation. Finally, a reminder to our audience that our podcasts are available on Sound Cloud, Stitcher, Apple podcasts, and Google podcasts, as well as the SEI's YouTube channel. If you like what you see and you hear today, give us a thumbs up please. Thanks again for joining us.

Thanks for joining us. This episode is available where you download podcasts, including SoundCloud, Stitcher, 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 don't hesitate to email us at info@sei.cmu.edu. Thank you.