
# Designing Trustworthy AI

*featuring Carol Smith as Interviewed by Suzanne Miller*

---

*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.*

**Suzanne Miller:** Hello, my name is Suzanne Miller. I am a principal researcher here at the SEI. Today, I am very happy to introduce you to Carol Smith, one of our senior researchers in the Emerging Technology Center [ETC]. She is here to talk to us today about the Human-Machine Teaming [HMT] Framework, which she has been developing as part of our artificial intelligence research. I am very excited that you are here to talk about this because I also have a background in human-computer interaction. So, these two marrying up together is very exciting for me.

It is also very timely that we are talking because in a couple hours from now, the DoD [Department of Defense] is going to actually be announcing the adoption of a set of five principles for ethical use of AI [artificial intelligence] in DoD systems. So, we will talk a little bit about that. I know that is not the main part of the framework. But you are at the very cutting edge of policy, not just research. So that is very exciting for us, and I hope it is for you, too.

**Carol Smith:** Definitely yes.

Let's start off. Talk a little bit about what you do here at the SEI. What does a human-computer interaction (HCI) person do in someplace like the SEI?

**Carol:** In the teams I am working with, we are doing a lot of prototyping and really thinking through early ideas. So, the work I am doing is to understand the problem at its source. So, what is the individual who is going to interact with the system, an analyst or a warfighter, whoever it is, really trying to understand their point of view, their mindset, and then bring that information back to make sure that we are building the right thing and building it in the right way. So, by using that data collection, that research, I can better understand that information and really help everybody to really understand who they are building for and why.

**Suzanne:** In human-computer interaction and artificial intelligence, we do not always think about those things together, like *AI, that's what the machine does*. So what is it that connects, and why do we need a human-machine teaming framework? I am very intrigued by the way that is constructed.

**Carol:** Ideally, these systems are actually augmenting our intelligence and helping us to be better at the work that we are doing and doing the things that we do not do as well more quickly for us, while allowing us to be using the skills that we have as humans to partner and to really bring those skills together. The human-computer-interaction aspect is even more important with artificial intelligence, because there is a lot of distrust often of artificial intelligence and machine-learning systems. So, by really understanding what it is that is going to help people to better partner and to trust the systems more, we can help them to be more successful because they will be able to partner more effectively.

At the same time, we also want to make sure that the system is built in such a way that they *should* trust it. So, also part of that is bringing the humanity into the AI system, not that it can become sentient and or anything like that, not that we necessarily even want that, but rather to make sure that it is reflecting what we need it to, the reality that we want it to.

**Suzanne:** One of the things that I hear from people that are starting to interact—and, all of us are starting to interact with AI systems—if you use Google Assistant or Alexa or Siri, you may or may not realize that you are using something that has an AI backend. One of the things I hear people say is, *I don't trust it because it's like magic. I can't tell. Yes, it gave me the suggestion that I need right now, and that's creepy*. What is it about this framework that you are doing that is meant to help to allay peoples' issues with, *Yes, this is trustworthy, or no, you really shouldn't trust this and turn it off now*?

**Carol:** Part of it is just helping people to understand why it is making decisions it is making; what information is it using. But really, the first step is very early in the process—before an AI is even being developed, before the first line of code is written—is to have the team really think through the breadth of the work that they are doing and to be speculative about the potential bad outcomes that could happen because of the work that they are doing, as well as the good outcomes, and to really think through that so that they are better able to protect people. So, the system is built in such a way that it is really going to be trustworthy because it is being built with their ethics, the technology ethics that they keep in mind.

**Suzanne:** The way I talk about that, and we do this in measurement all the time, is we want to avoid, we want to understand what the unintended consequences could be and then avoid the ones that are really going to hurt our ability to use whatever it is that we are dealing with. That leads us into the idea of ethics in AI. This has been a topic that has really come to the fore in

recent years. We have research going on on campus in that area, we have research at the SEI. There is research all over the place. There was a Defense Innovation Board report that came out last October that is dealing exactly with this—ethics in AI in the DoD. That I think is one of the things that resulted in the DoD announcing today that it is going to adopt these principles. I am just going to read them for our viewers. We can just comment a little bit on how important these are to the kind of trust that we are trying to build across this whole space of AI, not just in the DoD.

*We want to use AI systems responsibly, with humans being able to exercise appropriate levels of judgment. We want them to be equitable,* so we want to avoid unintended bias. *We want them to be traceable,* which is what we were talking about, how did we get to where we are. *They want them to be reliable,* so they need to be within a well-defined domain and not sort of going out into the ether talking about how you make crepes when you are really not anything about making crepes. And *they need to be governable*.

I wanted to hit on that last one just a little bit. What does it mean for an AI system to be governable? Because I think the others are pretty intuitive for anybody that sort of worked in this space for a while. But the governable one was the one that struck me as being, *Yeah, I like it, but what are we really talking about there?*

**Carol:** I think it is going to depend on the industry. Particularly here, it is being able to know the breadth of its responsibilities and also where its limits are and making sure that that is understood by the people using it as well as the organizations that are getting data from it and really regulating, to some extent, what it is. I don't actually remember what the statement is there, so I'm curious.

**Suzanne:** *Engineer to fulfill their intended function but possessing the ability to detect and avoid unintended harm or disruption.* And then *disengage or deactivate*. That was the thing, is that the governable is *I can disengage it. We are not going to end up with a robot, robots taking over the world because*...So that aspect of being able to be disengaged I think is actually one of the things that enables trust.

**Carol:** Definitely. Yes, and making sure that humans always feel that they have that ability and that if the system is not doing what is expected, that they are able to shut it off, and that there are mitigation plans in place as well for those instances. Because it is likely with these systems that there are going to be instances where we are going to feel that, *The information it is giving me isn't accurate anymore. Or, it is making a correlation here that does not make sense. So, I'm not going to use that right now. Instead, I'm going to use a different system, or I am going to use my own experience to apply to this situation.*

**Suzanne:** There are other places in the SEI, some of our other colleagues are working on things like causal learning and understanding the causal structure of information, so you can combine it with some of the AI machine learning to say, *Yes, this is the data. This is the structure that actually has meaning. This one, not so much. Let's turn that one off because we really know the data we're getting over there, it's costing us money to collect it, and we aren't actually getting the benefit from it*. So that is another aspect that I think the *governable* comes into play.

I don't know how long it has been that we have been actually talking to DoD about these kinds of ethics issues. Has this been a very long journey, or is this something that actually happened relatively quickly?

**Carol:** The more recent aspects as far as artificial intelligence and ethics is more recent, within the past two years. But, the DoD has a long history of really talking about ethics and thinking through this, particularly with the types of situations that they are regularly in. There are laws of war. There are all kinds of regulations across the different services, and they have their own sets of ethics, if you will, as far as what they expect from the service people. That has been very common and well accepted for a very long time, and this is building on that long history. But this is really relatively recent and very exciting.

**Suzanne:** Defense Innovation Board studies don't just pop out of the air. They happen for a reason. The fact that this was considered to be a topic that was worthy of getting people together to really think about how do we do this in a coherent way is very, very important for all of us that are doing this kind of work.

Let's go back to talking about the Human-Machine Teaming [HMT] Framework. How public is this? If people want to start looking at it, start applying it, what kind of resources have you gotten? What are you looking for in terms of collaborators to help you further that research?

**Carol:** The work is available online. There's a paper on the archive, which we can link to. The checklist that came out of that is also available on the SEI website as a fact sheet. [We are] certainly looking for all kinds of feedback and ways to improve this and really to take this to the next level. The checklist is just a start. There are definitely some gaps. I was talking with somebody online today about how, I talked about, *Does it align to your values*? That is a very vague statement. How do you really think through that? How can we better place this terminology and state these things that are more clear and more easy to actually implement? When you are talking about a software solution, *values* is squishy, and people are not very comfortable with that, which is reasonable. So, trying to make it easier and easier to people.

**Suzanne:** That is a context-dependent thing, right?

**Carol:** Yes. Very much.

**Suzanne:** Because the values, when I am in an operational setting versus in an acquisition setting or development setting, they are not exactly the same. So, the understanding of context, how to apply context to that kind of thinking is one of the things that I think we try to provide guidance on in other areas as well. So that will be something that I think is going to be evolving. So, what is coming? What is sort of your next big thing that you are tackling in relationship to this human-teaming framework?

**Carol:** Working on improving this, working on really just trying to see how people are using it, potentially having different versions for different types of applications. There are a lot of different ways this can go. One of the things I would like to see is really helping people to mature the process that they go through as they consider and then implement an AI system and helping them to think through those issues because the work has to be done somehow. At some point, you really do need to think through those implications, and that needs to take a little bit of time. But, at the same time, working in an agile fashion and being able to be iterative and productive in a reasonable timeline is important too, and how do you balance the tension there and thinking through that as well. The process is really important, AI engineering in general and getting that moving.

**Suzanne:** The connection between this kind of framework and innovation research—we haven't talked about that before, but I am sort of dropping this in. It came into my head that this is actually very well connected to the whole idea of innovation because, with innovation, we are trying to get new ideas adopted. And often—the values conversation kind of triggered that—that we are dealing with misalignments. We have the same problem in HCI that when we have misalignments of expectations, that is when we get some of the unintended consequences. Is that a source of research for you in terms of taking some of the things out of that domain and seeing how you can focus on the HCI problem?

**Carol:** There have been a lot of situations that we can learn from in the past few years, unfortunately. Learning from negative consequences is one way of realizing how we could make improvements and change processes. Innovation is definitely the area that I am working in with the Emerging Technology [Center] and thinking through just new problems, problems that we have not necessarily thought about in these ways: how humans and robots are working together; how humans and just regular computing machines are working together. Really just trying to think about these areas but in the broader sense now, because with artificial intelligence, the systems have a lot more reach, there is a lot more data to deal with, there is just more and more implications. Really having people think through that is important.

**Suzanne:** I see you being busy. This is an area that is emerging. And we are just kind of starting to hit the high points of, *Oh, we're going to have to deal with HCI. Oh, we're going to have to deal with ethics. Oh, we're going to have to deal with big data.*

I used to teach a class in AI and decision making back in 1989, that is how old I am. I am just thinking about the issues we talked about then and what we are talking about today. We could not have had anything close to this conversation. It was all about forward and backward chains and logic chains and things, nothing like we are able to do today. So, this is a very exciting time, and I really appreciate having the conversation with you about your work. I am looking forward to actually bringing maybe some of it into some of the things I do. So, I really appreciate you taking this time out for us.

**Carol:** It was a pleasure. Thank you.

**Suzanne:** For our viewers, the transcript of this podcast will include links to the kinds of resources we have been talking about. It will be available on the SEI website. It will also be available where you get your other podcasts. And if you have any questions, please don't hesitate to send us an email at info@sei.cmu.edu. Thank you very much for viewing.

*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.*