Implementing the DoD's Ethical AI Principles

Featuring Alexandrea Van Deusen 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. I am a senior research scientist in human-machine interaction in the SEI's AI [Artificial Intelligence] Division. Today, I am pleased to welcome Alex Van Deusen, an assistant design researcher, also in the AI Division. We are here to discuss a project we both just finished up supporting in the Defense Innovation Unit to develop responsible AI guidelines. These guidelines can serve as a guide for organizations in industry and government to implement responsible AI considerations into practice in real-world programs. But before we delve into that topic, let's tell our audience a little bit about ourselves, our backgrounds, and the work that we do here at the SEI. Alex, would you like to introduce yourself?

Alex Van Deusen: Hi. Thanks, Carol. I am so happy to be here today to share some of our work. I am Alex Van Deusen, and like Carol said I am an assistant design researcher. I was hired here at the Software Engineering Institute as a multimedia designer on the Communication Design team. There I was able to kind of grow my skills in visual communication and deepen my understanding of user-experience design. I was also able to recognize that elements of human-centered design and design thinking methods are kind of the same key elements to the process of solving creative briefs, which I have learned during my education in creative advertising. This includes things like identifying tensions, cultural truths, and asking *How might we?* questions. So getting to have that experience in problem framing, communication, and user experience, I was prepared to transition into my role as a design researcher.

Now a lot of my work focuses on collaborating with others to understand challenges, using my expertise to design solutions that meet our research and business objectives. Often we find at the Software Engineering Institute, we have these very intricate and interesting solutions, but they can be really complex and difficult to communicate. So, I also focus on developing ways we can showcase our work and demonstrate our impact.

Carol: Excellent. And as I mentioned, I am a senior research scientist here at the SEI's AI Division, and I focus on human-computer and machine interactions. My career in human-computer interaction began about 20 years ago. I have worked across many industries, conducting user-experience research and doing prototyping to support fast learning so that we can use those as experiments to learn.

In 2015, I turned my focus to improving experiences with artificial intelligence systems and emerging technologies and explored really beyond that user experience to look at the ethical, social, and safety issues that are part of those systems. When I joined the SEI about two and a half years ago, I wanted to come here to focus on the important challenges that our government customers have regarding human-machine interaction and human-machine teaming, and really making human-centered artificially intelligent systems. I also teach courses and lecture at Carnegie Mellon University's Human-Computer Interaction Institute.

So I have contact with the students and professors there. I am enjoying the idea of really implementing a lot of this work So, that brings us to the topic of today. And Alex, if you could, start by talking about the Defense Innovation Unit and what the catalyst was behind this work how we became involved, and I'll chime in as well.

Alex: Yes, absolutely. Thank you. Carol and I were both kind of brought onto the team at DIU as technical advisors, researchers, collaborators. DIU is the Defense Innovation Unit. It's a Department of Defense [DoD] organization that helps the U.S. military make faster use of emerging commercial technologies. They work to accelerate DoD adoption of commercial tech, transform military capacity and capability, and strengthen the American national-security innovation base. DIU launched a strategic initiative to implement the DoD's Ethical Principles for AI into their commercial prototyping and acquisition programs. So, that is where Carol and I came in on that work.

Carol: Yes, and it was really exciting because we had seen these ethical principles released, but there really wasn't a lot of guidance available. Being able to really look at these more deeply and think about how people could make these elements, these principles work for them in making systems is really exciting. For those of you new to the DoD ethical principles, those are *responsible, equitable, traceable, reliable,* and *governable*. They come with descriptions of each, but there wasn't a lot of guidance as far as how to really make this part of the work. So this work

was really to be able to develop those guidelines, working from the real-world experiences that we have here at the SEI as well as DIU's and drawing upon the best practices that we've seen from the government, from nonprofit organizations, the academic community, and our industry partners. So we wanted to talk a little bit about the process that we use for developing these guidelines and how we solicited input from vendors as well as colleagues. Can you start with that, Alex?

Alex: Absolutely, yes. I want to echo what you had said about those DoD ethical principles for AI. The big part of these is that they were released, but they don't have a prescribed methodology for how to implement them, and that is kind of what you were saying there, Carol. They don't have a concrete direction on the way that they should be implemented, but they do identify that clear need for practical implementation guidelines for technology development and the acquisition workforce.

DIU had launched a strategic initiative to implement the principles into their commercial prototyping and acquisition programs, and the whole overall goal was to ensure that the AI ethical principles were able to be integrated into planning, development, and deployment of all DIU programs and prototypes. This would also enable all of their stakeholders from their program managers to commercial vendors and government partners to effectively examine, test, and validate that all the programs and prototypes meet DIU and DoD ethical guidelines and principles.

So what we were trying to do was establish a process that all of these different stakeholders could use that was reliable, replicable, and scalable across DIU programs and prototypes and then be able to be expandable into other DoD organizations. Like I said, that is where Carol and I came in, and DIU had done so much work identifying methods for implementing the principles. They had some ideas about what was important to ask and talk about with the vendors, but they didn't have a clear direction on how to make this system reliable, replicable, and scalable. What Carol and I did was attempt to figure out how we could take that ongoing work from DIU and make it something that people could actually use for their projects. The effort that we worked on with DIU is shaping how DIU works with their vendors, works with their partners, and is influencing the contracts between those parties. Do you want to talk a little bit more specifically about what these guidelines are, Carol?

Carol: Yes, sure. As you mentioned, we have got the three different stages of work: the planning work, the development work, and then the work to actually implement or actually get these systems out into the public. Part of the issue there is really trying to define early on those tasks: *What is someone going to be able to do with the system? How are they going to evaluate the system?* And thinking about the ownership access to the provenance of the data in the models. That is something that really often is overlooked in early stages. And so by bringing some of

these big questions into that planning phase, making sure that people who will be responsible are identified early and that there is some harms modeling done to really look at the magnitude of harm that the system might create, and being speculative about that as well as looking at system rollback and error identification. These are all part of that planning phase that is just so important to make a responsible AI system and really making that clear to people what work is to be done in that planning phase and then in the development phase as well, reflecting back on those planning questions but also looking at manipulations that could occur, defining procedures for reporting issues, thinking about who would be able to make and certify changes to the capability, and then looking at plans for verifying outputs. Finally, defining roles for this third-party system audit and other ways of really making sure that the system is continuously being monitored, and that humans are in the loop at all times. Then finally, we worked through the deployment phase.

And through all of this, again, these same comments and questions and conversations are coming up. That is to make sure that this is really intentional work, that people are really thinking through the entire system as a full system and not just one particular aspect of the system. So in deployment, it is looking at that continuous task and data validation. *Is the system still doing what it is intended to do?* Looking at functional testing, *If it's still meeting the desired functional goals, if it's performing the way it's intended to?* and then also looking at harms assessment and quality control. Again, *Is the system being monitored well and is that monitoring supporting the work? Are you continuing to identify any harms that may be occurring and doing that work to really think through what this could be and thinking about who is responsible to manage and maintain any of these systems?*

So lots of really deep and important questions that need to occur at each of those three stages. That is what we worked through. It was really refining not just what questions need to be asked but making sure that there were support systems for people to be able to understand what they were doing. Alex worked very hard on really improving these graphics that help to show how people are interacting with the systems. Then also, we worked on creating worksheets to help people go through these various stages and keep it at an approachable level but also making sure that people know the various steps that they need to get through.

Alex: Yes, and I just want to also add to that before we move on to our next question or topic is that planning part is really important. I think it is important even before we start thinking about, *What AI capability are we going to build?* But to ask the question, *is AI the right solution*?

Carol: Yes.

Alex: Because a lot of times, we are really excited to implement AI wherever we can. And we kind of skip the part where we think really hard about, *Well, is AI going to solve my problem? Or is AI the most efficient or effective way to solve it?* Some of those planning questions even get

down to the root of, *Is this really the right direction you should go with your strategy for solving a problem?* So, I think that's a really important part of the planning too is thinking, *Well, you know, how are we going to go about solving this and is that the right way?*

Carol: Yes. Excellent point, and along with that, not being afraid to stop working on a project when you find that indeed, the harms are too much or that the system isn't performing as intended. Excellent point. All of this is intended to really prevent making the wrong system, prevent making the system in the wrong way, and certainly to prevent any unintended consequences from the system. Excellent point.

The human-centered artificial intelligence is one of the three pillars of AI engineering that the SEI has identified. It is now recognized also as being important to successful AI system development more and more broadly across government and industry, and guidance is still needed for how to achieve it. So, thinking about the DIU guidance and the major tenets of this guidance being around being human centered, how do you think about the organizations being able to achieve human-centered AI?

Alex: Yes. I think this is a great question, and I think a lot of times we really have to go back to thinking about the user. I think that is what we were bringing to that team was a consistent, *Let's think about the people who are going to use this solution that we are coming up with or these guidelines*. It's a big part that is often an afterthought to design, to the design processes. When it comes to thinking about user experience and user interaction, oftentimes we think about those things too late in the game to actually make effective change. But there is a lot that we can change or improve to make that interaction more smooth and efficient for the user and make it something that people will actually use.

The place I see this work best is when it is evident that a great deal of care was taken toward the design of how information is presented. I think that was a great thing for us, getting to work with DIU, especially with our partners, Jared and Bryce from DIU. They had a deep understanding and a deep care for the people that they were creating this for. I think that really came through when they brought us on the team because they wanted us to be able to represent those people and think about those people. Our whole goal was to make this accessible and usable, like you said, looking at the questions, looking at the workflows, and even the visuals and thinking, *What are the different steps of this process? How can we reorganize and start to think about what's the most natural way that people are going to progress through answering the questions with those necessary stakeholders?* So thinking about from information architecture to information design, how we organize, structure, label the content. We always have this goal of helping those users understand and find the information and be able to complete their tasks, so we had to figure out how those pieces fit together into the larger picture of DoD ethical guidelines and actual AI capability creation. That was part of it was us thinking about the users, the context, and the

content and being able to make something that's reliable, replicable, scalable, but most of all usable.

Carol: Excellent point. It is so important for any type of tool being made to help people feel that it is approachable, that they can understand and use it and get to work making better systems so that then the end users can also have better experiences. Particularly when you are introducing an artificially intelligent system that in itself is so complex and placing it in a very complex situation, having tools that are helpful to you to understand that context is really important. Then taking the time, of course, to do the work to understand the end users' situation and how they will be using the system, what the larger situation is like, the environment. Really doing the work of a good design so that again, it is the right system for the right people. It's usable. It's useful and really supporting their work in whatever ways that is.

One of the other areas is within the Human-Centered AI paper that we put out, that "Engaging in Critical Oversight" piece. I was really happy that that also was reflected in the responsible AI guidelines and that there's so much information there and focus on making sure that humans are always in the loop, that humans are in control, that they are monitoring and maintaining these systems, that these systems aren't ones where you set and forget like most software from previous generations has been. You make an update, and then it's done, at least temporarily. Rather with this work, it truly is a constant work. It is constant updates and oversight and management. Making sure that these systems are behaving the way they are expected to and doing the work that they're expected to do. Not just looking at numbers and the accuracy and those types of metrics but truly, *Is it behaving the way we expect it to? Is it really meeting the needs that it's supposed to address?*

Alex: Absolutely. I think that main takeaway or at least one of those main takeaways is to get the vendors and stakeholders that are going to be involved in the conversation aware of the landscape of responsible AI and help them be able to answer. Or, if they have to on some sort of problem question, be able to do a deep dive and think about, you know, answering some of those questions because we want to build confidence in AI systems, building confidence in understanding how the systems can and will be used is really important when we design them. So, like Carol said, being clear in our intentions, and that means saying, *What do we expect this system to do? What do we expect it not to do?* You know, *What should we not expect an AI system to do that humans can't even do?* So it is being clear in the intention of our solution, identifying those people who are responsible for the AI capability and again, clearly defining what we expect it to do and being able to visualize how those ethical principles apply to each stage of the cycle, from planning to the development into deployment of AI technologies.

Carol: One other piece is the focus on the data and the models, and really understanding the implications that those can have. I think many times as mentioned earlier, there's a lot of focus

on creating an AI system, but not necessarily understanding the content, the data that the system is based on, and how bringing different types of data together or bringing different datasets together can potentially create new information or have unintended consequences in that way. Really focusing on people defining exactly, *What is this information? How do we get it? Who owns it?* Particularly with the models as well, being clear about that upfront and making that clear to everyone involved in the project is really important for that to be a trustworthy system, a system that people are willing to interact with and that they understand. As you mentioned *What is the system supposed to be able to do, and what can it not do? Or what shouldn't it do anyway?* And being able to report if there is an issue with the system, if it's not working as it was intended.

We know that organizations in government and industry, we hope anyway that they are going to use these guidelines in their own work with developing AI systems. We have got lots of resources that we'll be linking to from this information. Are there some things that you would point out as first steps for them as they get started?

Alex: Absolutely. I think a great place to start is, you know, at the culmination of this work we've been talking about. So, DIU just released their *Responsible AI Guidelines in Practice*, and Carol and I were coauthors for that. We are so excited to be able to share this work. I mean, we have been working on this for a long time. So it is exciting to see it out there, and it is exciting to see people's reactions to it. So yes, like I said, it was written in collaboration with members from the AI ML portfolio at DIU, and Carol and me from the Software Engineering Institute. The report really provides a step-by-step guidance for AI companies, DoD stakeholders, program managers, and ensures that AI programs are going to be able to reflect the DoD's ethical principles for AI and really goes into ensuring that fairness, accountability, transparency are all considered at each step of the development cycle. This was so interesting. For me, I just had a great time doing the research, like looking at what the current state is, getting to talk to different people and hear from different people in industry, government, academia. It was just a great experience, and I'm really looking forward to continuing to work in that area. But that's definitely one place I would point people is go check out that report because it was a labor of love.

Carol: Indeed, well said. It was a lot of fun and a lot of work. Well, thank you so much, Alex. This was great. I always enjoy talking to you. And for our audience, we will be including links to the transcript and resources that we mentioned during this podcast. And thank you again for joining us, and make responsible AI.

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.