SEI Cyber Talk (Season 2 Episode 2)

Human–Machine Teaming and AI
by Ritwik Gupta, Carol Smith and Andrew Mellinger

Ritwik Gupta: Hi, everyone. Welcome to another episode of the SEI Cyber Talk. I'm Ritwik Gupta, your host. I'm a machine learning research scientist here at the Software Engineering Institute, and today we're here to talk about this really fascinating world of human-machine teaming. So we me today I have two researchers-- Andrew Mellinger and Carol Smith-- and they're here to tell us more about what it means to be a human and be a machine and how to work today.

Carol Smith: Thanks. I'm Carol Smith. I am a senior human-machine interaction researcher here in the Emerging Tech Center at SEI.

Andrew Mellinger: Hi, I'm Andrew Mellinger. I'm a software engineer, software architect at the ETC as well. I work with Carol. Actually we're officemates next door to each other.

Carol Smith: Yeah.

Ritwik Gupta: Yeah. I mean, I can always see you guys banter the whole time. So human-machine teaming-- it seems like this really complicated world, because not only are we dealing with human problems but also this world of machine problems, and all the kind of uncertain, really soft and sometimes really technical world in the middle. So what does it even mean when we say the words "human-machine teaming"?

Carol Smith: So I think about it as not just a machine, not just software, but rather some type of complex system, an artificial intelligence intelligent system, that has the ability to do some things and to support a human in the work that they're doing. So they're actually working together and sharing knowledge, sharing information, and determining how they're going to use that to work to each other's strengths and support each other in solving some bigger problem. And sometimes it's one human and multiple machines, and sometimes it's just one-on-one.

Ritwik Gupta: Gotcha. So why not just get a whole bunch people who research humans, get a whole bunch of people who research machines, and just let them do their own thing? How did this specific world of human-machine teaming-- I assume the fact that if I'm using a phone or I'm using a laptop or something, it's-- am I a team already? I am not (inaudible) working? Why do we have this specific world? What's going on there?

Andrew Mellinger: Yeah, so from my perspective, and to follow what Carol was saying, there's a difference between-- I'll put in terms of normal human-- in terms of communication versus collaboration. So when I think about communication, I might want to tell you something. I might want to tell Carol something, and you might tell me things. That's different than collaboration, where we're actually working together and we're actually concerned about each other's perspectives, more so than in communication; it's much more assertive. So in the case of human-machine teaming, really it's about-- the point where she said they work together. It's how can the machine be aware of what the person's doing, and whether it's through AI or whatever the system happens to be, but also how is the person aware of what the machine is doing, and that involves things like trust. So as an example of what I mean by how can the machine be aware of what the human's doing, there was some research into when people get tired, how do they behave. And so, for example, maybe I'm studying a screen, maybe I'm watching that, maybe my job is to be monitoring something for hour after hour. On three or four hours, the machine may change its behavior about what it shows you. And so it may filter out some of the less important things and accentuate the more important things. So it may change its behavior over time as you get more tired-- and so how can the machine actually adjust its behavior to what we want, which is more what you see between people working together, with shared mental models. So how do we get that shared mental model into the human and machine combination rather than just human to human?

Ritwik Gupta: So based on what you're saying, it doesn't sound like it's something new. As long as machines have been around, we have been teaming with them, and these interactions, communication, collaboration problems have been around for a long time. Would you say that the way I interact with just my user desktop, is that a valid human-machine teaming problem, or is it something a little bit more intense, where the machine and I are working towards a common goal and that's where the teaming problems come into play? I guess what I'm asking is: What's the difference between UX, like human-machine teaming, and where do they intersections actually lie?

Carol Smith: Yeah, so user experience is a big piece of human-machine teaming, but the human-machine teaming is different from a human working with a desktop because the system is actually responding and changing based on the situation. So it's not simply me using the machine, but rather the machine interacting with me, and it's that actual interaction and that working together on the same problems in a way where you're exchanging information, beyond what you would do with a desktop or even a website. It's actually adjusting its behavior based on the information it has and then that is furthering along the work. So the work is progressing in a different way than it would if you were just interacting with a regular computer system.

Ritwik Gupta: So when you say that the system is adjusting its behavior as the human's interacting with it, does that mean the system has to be intelligent, or can it be a much simpler-- can it be like a rule-based system? Does a system have to be AI-enabled, or do these human-machine teaming problems exist in simple systems, like, I don't know, a command-and-control system, or I don't know, like if I wanted-- like some robots at UPMC Presbyterian deliver medicine to different-- do these systems have to be smart in order to have human-machine teaming problems?

Andrew Mellinger: That's pretty tricky. So I'm not really sure what you mean by smart or what you mean by intelligence or even artificial intelligence, and I think that we have one of the problems in AI right now, where as soon as we solve a problem, people say, "Well, that's not really AI anymore." Right? "That's a solved problem. That's something we have an algorithm for, so therefore it's not AI." So we're kind of chasing this boundary of intelligence and artificial intelligence like we chase magic, right? So one person right now might say that human-machine teaming is the stuff that-- how we imagine a human and machine working together in the future that we haven't come across yet, like R2D2 and Luke Skywalker are flying the X-Wing together. Arguably Clippy, back in the Microsoft days, could be considered human-machine teaming, and it did somewhat send some sort of responses, it did learn a little bit. Is Siri really human-machine teaming?

Ritwik Gupta: Is that a dig at Apple?

Andrew Mellinger: Well, you can look at Siri and how it's been used in other platforms. In fact the software came out of SRI; it was used in the military perspective as well to actually help the war fighter adjust to looking at localized problems and be able to-- anyway, do a variety of searched in a more contextually-aware fashion. But what it ends up coming down to is we have this spectrum of what it actually means, and so one could be pedantic and say that, yes, Clippy was human-machine teaming, but really what's interesting is how do we-- what are the ways we can actually make that more useful, right? And so does it have to be smart? Does it have to be AI? Well, it has to sense. It has to be adaptive. That's the important part, is that the machines are capable of sensing what the human's doing in some fashion. It's not just a timer that goes off after three hours, "We're going to change the behavior." It has to have some sort of mechanism for sensing and some mechanism for reacting to that. Is that smarts? I don't know. I'm not really sure it's important to say that it's smart or intelligent, but to be able to identify the qualities that we need to have to make those systems work together well.

Ritwik Gupta: So what are some of these qualities that make the system work well?

Andrew Mellinger: I'll let Carol talk about trust a little bit.

Carol Smith: Yeah. So definitely trust, in that the human really has a clear understanding of why the computer-- the machine, I should say-- is making decisions and what basis it has for making those decisions, and how it's looking at the information it has, what information it has. All of that knowledge and awareness on the human's side gives them more confidence and trust in the system, that the human understands why the machine is making these decisions and why the machine is reacting the way it is, and why it's not doing some things in certain cases-- why did it not react in this situation, or why did it not choose this option. And so by building and providing more and more transparency, that starts to engender trust from the human in the machine, and that's the really important part: Do I trust my desktop computer? There's nothing to trust there necessarily, but in human-machine teaming, the trust becomes very important. So that's a large difference, I would say, with these systems.

Andrew Mellinger: Yeah, I think it's a bit more complex, and Ritwik, you have some robotics experience, right? And so there are some ways we can try to humanize the robot, so we like to give them more trust maybe than they actually deserve. I think we'll end pretty sure-- we're starting to develop a lot of distrust-- ignoring all the Heinlein sort of robotic distrust-- but I think that we're starting to develop a lot more distrust for AI as we start to see a lot of these systems violate privacy, make decisions about things that they really should not be making decisions about, and I think we're actually going to see kind of inverse of the hype cycle, which is going to go down for a while, till we learn how to actually develop valid trust cycles. There's actually some term I think in robotics about how you give-- or even in ethics-- about how you give the computer initially more trust. There's some-- I forget the name of that one, but yeah. So we're going to get over that soon I think and have a lot more issues with how people actually trust the AI.

Ritwik Gupta: So that's interesting, because you're saying that in order for me to successfully work with a machine, I need to establish some sense of trust with it, right?

Carol Smith: Mm-hmm.

Ritwik Gupta: If I don't know what-- I don't know, let's say I am (inaudible) robot, and instead of my really cute emotional support Corgi, I have my emotional support robot and it's taking me down a road, and you're saying if I can't trust it, I'm obviously not going to follow it, or I will be very wary of its instructions.

Ritwik Gupta: How can I begin to trust a machine when I don't even know how to properly trust some humans? Like Andrew's code, I don't know if it works. I don't know if I trust that behavior, right?

Carol Smith: Right.

Ritwik Gupta: If we haven't solved this problem of human trust, how are we proposing to solve this problem of human-machine trust?

Carol Smith: Right. These are really complex problems and there's a lot of things with humans that we still haven't figured out and that's going to continue to be a challenge with the machines and figuring out what is enough, what is the point where people are more trusting, and different people are going to have different levels of trust anyway, and so what works for one human isn't necessarily going to work for another human, and so that's going to also change and may change how much transparency the machine provides, because maybe one person needs a lot more information before they build that relationship, whereas another person may inherently have trust of the systems or has worked with the systems before and so has built up an ability to trust these systems more. But either way, we want to make sure that the humans are always in control of the situation, to the point where they can determine when it's necessary to take over from the machine. That's part of that trust as well, is enabling the human to be able to say, "I don't trust what's going on right now. Therefore I'm going to take control and do something different."

Ritwik Gupta: Right. For anyone watching, by the way, as an aside, if you do have an emotional support Corgi, I am very jealous. Bring it over here to CMU.

Andrew Mellinger: We should have some for the office, I think.

Carol Smith: I agree.

Ritwik Gupta: I completely agree.

Carol Smith: Hundred percent.

Ritwik Gupta: So trust sounds super important. Obviously you don't want to be trusting everything you're working with. What's another huge quality for human-machine teaming beyond trust?

Carol Smith: So thinking about the interpretability and just how do you convey the information in a way that's clear, and how do you transition that information between the human and machine? That's something that's going to be very important, is making sure that while a system-- a machine can process huge amounts of data, the human can only process so much at a time, whereas a human can change context very, very quickly and the machine can't always change context that quickly. So figuring out how to manage the strengths and weaknesses of both systems, the human and the machine, is going to be really important as well.

Ritwik Gupta: From a deeper software perspective, how do you build interpretability into a system?

Andrew Mellinger: Wow. From a software perspective. So a lot of it has to do with appropriate-- well, it goes all the way back to the design stage. It's like we talked about initially the security, how we would test in security at the end, and now there's the whole point about you design in security from the beginning. Interpretability has to be designed in. I think this is kind of one of the current arguments in the AI space about explainability versus actually-- "interpretability" I think was the contradictory term. So how do we go ahead and make sure the machines from the get-go are capable of being understood? This has to do to some degree-- assuming that most of our systems, our human-machine systems, are going to have some level of AI-- we're going to be training off data, and so where do we get that data from and how do we have some confidence that data is being unbiased? Or if it is biased, how is it biased in the way that serves the need of that human-machine team? So we are going to have to be more conscious of that, and I don't know if in the future we're going to have some sort of certification where this data came from some cited sources, some metrics for that or some way to know, because people are going to be asking those questions. Where did the data come from that was trained for this system? I mean, if you have a system that's assistive-- let's say medical diagnosis. How are you sure that this data set is appropriate for all races, or all genders? And that's going to be part of the overall architectural process, and then as you do get those data sets in, how do you continue to maintain those across time? We know the algorithms that are capable of dealing with that type of data are going to be available for working with that. So the engineering process then is going to be throughout, and then at the point where we actually have the UX designers understanding how to present that data, we'll have to tie in those steps as well.

Ritwik Gupta: So you're saying it's not a magic bullet. It's going to require architectural changes and very, very interpretability-aware design processes. It's not just something that I can just say, "Done. This thing will make my machine interpretable, and now I can just call it a day." Right? This is a fundamental problem to solve.

Andrew Mellinger: I argue yeah, but also you're going to find vendors I'm sure that'll come in and say, "I have that silver bullet for you. If you still our device in the middle of the processor, you still our processor in the middle of your entire lifecycle..." I mean, we saw that before in big data, we saw that in security, we see that in AI now; it's going to come around with human-machine teaming too.

Carol Smith: For sure.

Ritwik Gupta: It sounds like a fascinating world, right? There's so many moving pieces. Again, these are just two of those key qualities that we're talking about. But it doesn't sound like a lot of this is very organized. It seems like a very nascent field; it's still growing. Is there any framework for how to think about understanding human-machine teaming. I understand that you had work recently at AAAI that talks about this?

Carol Smith: Correct, yeah. I presented a paper and a checklist at the AAAI summit just a couple weeks ago, and it was really great to get the feedback from the audience and talk through what is that we can do to help ourselves at those early design stages, when we need to make those big decisions about the system's going to work and what the system's going to provide, and then as we're building it, making sure that we're reducing the unwanted bias, that we're really querying what are the things that we want to avoid, the unintended consequences that we need to make sure that we're mitigating and helping people to really query the system before it's in front of humans that are working with it.

Ritwik Gupta: Gotcha. And then from a deeper perspective, is this-- I mean, it's a framework, right? So there's multiple possible frameworks. As a practitioner, let's say I'm building a system that will interact with humans. How do I follow this framework? What do I do? How do I make sure that I am being not only specific to my use cases but also general to the overall use case of human-machine teaming? Or "use case" is the wrong word, but making sure that human-machine teaming is as synchronized and in sync as possible.

Carol Smith: Yeah, so part of it is aligning on a set of ethics, technology ethics, and really understanding what is important to the team and looking at that closely and then using something like the checklist that I developed to really, again, have those conversations and make sure that everyone's communicating and understands what it is that they're trying to build and what is going to be outside of the operation of the system and how to keep humans at-- the goal is for the humans to be able to be more successful and more powerful and more-- better problem-solvers because they're using this system, and not to put them at a disadvantage with the system.

Ritwik Gupta: I see. Interesting. And then I think you mentioned ethics as well, and I think that's-- we can have another Cyber Talk on that, except just water we'll have shots. But that's really interesting that you not only want to design systems that work well but work well in an ethical fashion. What happens when your mission goals and ethics are at complete odds?

Andrew Mellinger: Well, so I think for human-machine teaming, part of the goal at this point, and one reason why I think it's becoming more important, is the human becomes the human in the loop and provides that ethical governor across these machines. So I think we're going to actually see a tremendous rise in the human-machine teaming demand as we recognize that the AI is not prepared to deal with these ethical questions. And so the answer really is: That's why we have the humans there.

Carol Smith: Yeah.

Ritwik Gupta: I see. And so how does a machine then, in this framework, pick up a sense of ethics? It sounds like it's this very, very mushy social thing that necessarily doesn't lend itself well to being learned or hard-coded in. How do we make sure that these machines not only are picking up some sort of ethics but how do we make sure that those things are being updated and constantly being checked by the human? What's the process for that?

Carol Smith: Yeah, I think part of it is just keeping the human in the loop, making sure that the human is aware. So there's certain decisions that are not enabled by the machine. So the machine can't make certain decisions regarding human life, or regarding decisions that are going to alter a human's well-being-- or anything that's really important and affects humans should be a human-made decision versus a machine. Or that those decisions can at least be changed, so if the system is enabled to make a decision but the decision wasn't based on some more subjective information that is available, that humans can then override that decision. So it's not necessarily that the system itself has ethics, although that's ideal, in some cases those won't be as possible, but we can build in safeguards to make sure that there is a human who's making those decisions and that the decisions are as transparent as possible.

Ritwik Gupta: I think that's a very interesting point, because I do think that because of popular media and stuff, people start thinking that systems have ethics, but really what you're saying is that it's the people who design the system who have the ethics and it's their job to imbibe that into the system. So the machine are not doing anything just by itself hopefully.

Andrew Mellinger: Or to be higher-level detail, it's the choice of the data we use to train those systems that have the ethics in it, embedded in it, and unfortunately we may not know what those ethics actually are. I think we're going to start to evolve better ways of visualizing our data sources and understanding our data sources before they even go into the AI systems. I think we need to work hard on that.

Carol Smith: Yeah.

Ritwik Gupta: There's so much here I want to dig into. Again, I think we could just spend hours talking about data and what does it even mean to be ethical and all these things, but for the sake of the audience, I don't want to bore you too much. Are there any cool resources that you guys might have that you can point so that people, if they want to learn a little bit more about ethics, or learn a little bit more about UX or HMT, where should they go look?

Carol Smith: Yeah, there are lots of articles everywhere right now on ethics. On most of the major technology blogs, they're starting to talk about it a lot more. The Department of Defense just recently had-- the DIB presented some ethics with regard to artificial intelligence, which is really exciting, and AAAI's symposium from just a couple weeks ago-- those papers are online on archive.

Ritwik Gupta: Okay. And then from a little bit more of a softer perspective, is there anything that talks about data integrity or how to build these-- as you mentioned, interpretability is more of a software architecture process rather than a simple one or two methods. Anything that you could point out there?

Andrew Mellinger: We really published a paper on 11 principles of AI, and so assuming that AI is going to be part of this larger system, that would be a good place to look.

Ritwik Gupta: Awesome. I know that here at Carnegie Mellon we have an entire institute, the Human-Computer Interaction Institute, that focuses for a long time now on human-computer interaction, so if anyone's interested, you can check out HCI's website, fascinating research.

Stanford HCI just recently became a thing. So Stanford HCI is a great resource as well, but we'll add more things that we think of into the transcript as well, and some of them hopefully come up on the screen too. But yeah, that's fascinating. Thank you guys so much. Again, if you guys have any questions, please feel free to reach out to us. You guys can email me at rgupta@sei.cmu.edu, and I can get it over to these guys, or just if you have any generic questions, please feel free to email us at info@sei.cmu.edu, and I'll see you guys next time.

Related Resources

AI at the SEI: https://resources.sei.cmu.edu/asset_files/FactSheet/2019_010_001_539538.pdf
AI Engineering: https://www.sei.cmu.edu/research-capabilities/artificial-intelligence/index.cfm
AI Engineering: 11 Foundational Practices for Decision Makers
Designing Ethical AI Experiences: Checklist and Agreement
Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development