Bias in AI: Impact, Challenges, and Opportunities

featuring Carol Smith and Jonathan Spring

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.

Jonathan Spring: Welcome to the SEI Podcast Series. My name is Dr. Jonathan Spring. I am a senior vulnerability researcher here at the SEI's CERT Division. I am joined today by Carol Smith, a senior research scientist in human-machine interaction at the SEI's Emerging Technology Center. Today, we are here to talk about bias in AI [artificial intelligence]. Welcome, Carol.

Carol Smith: Thank you. I'm glad to be here.

Jonathan: Great. Do you want to start off by telling our audience a little bit about the background and what work you do here at the SEI?

Carol: Sure. Yes. Human-machine interaction is really looking at problems between people working together and working with technology and trying to figure out how to best not only design the system, which obviously is very important so that they can use it easily and find it to be helpful for whatever it is they need to do, but that they really understand the system in an appropriate way, and that the systems are really supporting them in a variety of ways. So, looking at aspects of human-computer interaction, thinking about user experience, and doing research to better understand the needs of the people and their problems, and then using that information to make recommendations on systems and prototyping. We do mostly prototyping.

Jonathan: Oh, I think that is really great. Right now I am mostly with the vulnerability-management team, so we coordinate vulnerabilities in all kinds of systems. One of the things we have talked about lately is what it would be like to manage or coordinate a vulnerability in an AI system. But then, of course, we get into this question, which I think is really topical but is not what we are going to go into—when is something a vulnerability in a machine-learning or AI system, and what does that mean? So, bias is different than vulnerabilities, but I am hoping that maybe we can talk a little bit about where those lines blur.

But as a start, I know that you are familiar with this NIST study from 2019, the National Institute of Standards and Technology here in the U.S. They have put out a finding that commercial facial recognition, different apps that were on the market, misidentified people basically based on their race, their age, their gender, and specifically, people of Asian origin or people who are African American are consistently misidentified more often than white men. There are a lot of examples of bias in AI systems, but this is one that is very topical for people because it was in their phones, and it was not letting them unlock their phones and that sort of a thing. So what can you tell us about how bias evidences itself in modern ML and AI systems?

Carol: Yes. It is really is dependent on the data that is put into the system, and in that case, most likely they were working with an initial set of data that had been trained on primarily white and primarily male faces, which is very common because those are the people who typically have made these systems in the past, so they often use their own images to train the system. The bias comes in because, not only is the data itself very narrow in nature, but the people making the system don't always even realize it. So looking at a set of data that looks like a bunch of people I work with, I may not initially realize that people are missing. That is the bias that we are seeing in that sense. Once that system is being used by the general populace, all of a sudden that bias becomes very obvious. Part of the issue there is the data, and then part of it is the people making the systems, both of which aren't necessarily representative of the end use and the people who need to use the system.

Other examples of that are found in systems that are being built on historical information. A lot of that historical information, if you look at systems regarding mortgages and regarding all kinds of information about patterns in society related to financial or health or other aspects of our world, because many of those decisions and the data related to those decisions around those systems have been based, unfortunately, on bias. So, for example, with mortgages, often Latino and Black borrowers have been discriminated against in ways that result in higher mortgage rates. So if you are using the historical data to make a new system to determine mortgage rates, you are very likely, and it is almost impossible to remove the bias that is going to result in the system that continues to create that same system of bias and racism against those individuals.

Jonathan: Help me unpack all of the great information that you just presented there. Bias and data collection are not new. There have been all kinds of various problems with the data that's been collected over time. But we are able to see that in a lot more striking way now that we are automatically using that data to make decisions. I have heard some people in good faith at conferences say that they didn't include racial identifiers in their data, and so they don't have bias in their outcomes. Can you maybe talk a little bit about how correlates to sensitive information can still mean that you are discriminating or having bias based on that sensitive information, even if it is not explicitly in your data?

Carol: Certainly. Yes. One aspect when you remove those pieces is that you can't track it, so you don't know if it's racist or not. You can say that, but you don't know. Similarly with gender and all those kinds of indicators of specificity. Removing those doesn't remove that bias, and the reason for that might be, for example, with the mortgage example. If you were tracking race and then removed that from the data, you are not removing ZIP codes. You are not removing other aspects that can indicate information that you may not even be aware of, but that the system will see a pattern in that is then exposed through the AI learning, because the system will look for patterns naturally. An example people use often is with images where an AI system, instead of learning about the foreground image of an animal, for example, it learns about the background and associates more with that because of where it is seeing a pattern. I don’t recall the animal, but there was a system that was trained on identifying an animal, and instead it was actually identifying the snow in the background for one animal and lack of snow in the background of the other animal. So it wasn’t actually learning what people intended it to learn, but it seemed to work. It was working. Every time there was this animal with the snow in the background, it recognized the system. So there are all kinds of things in data, things in images, things in all of these creations that we have made, the data that we have decided to collect and label in certain ways and organize in certain ways. There is bias there that we don’t always even become aware of until we notice it in these systems.

Jonathan: Yes. So how can we test for bias in results of AI systems in a way that is a little bit more reliable and a little bit more forward looking than just someone else tells us that there is a problem?

Carol: Ideally, yes. We ideally do not want to have to wait until the systems are being used by people, so part of this is being more skeptical and really thinking carefully about the data. Really understanding its provenance. How it was collected. Why it was collected. What the goal was with that information. Really looking at the details of that data. And there are methods for that. Datasheets for Datasets is an example of a paper that was written that describes a method for really looking closely at that data and understanding it, and then also being speculative. So looking out after you have built the system, imagining what some of the consequences might be. So particularly when multiple datasets are being combined, there could be a risk of personally identifiable information being created because datasets that haven’t been combined before are combined, and now all of a sudden you can actually identify individuals in that dataset. There are other reasons why a system might become more vulnerable once it is accessing certain data sources. So really thinking through what the consequences are and the worst-case potential scenarios can help you to mitigate those risks and to ideally either, if possible, reduce the risk in the data—usually, that’s very difficult—or change data sources, if that is possible. Or maybe an AI system isn’t the best solution in that particular situation.

Jonathan: So there are a lot of intersections with other parts of machine learning right? There is this whole thread on adversarial machine learning, that an adversary can invert the model and recover information out of the model. That is a big problem with sensitive information, personally identifiable information that might have been joined together. Is that intersection something that you have thought about from this sort of bias and ethical perspective as well? Do the people creating a machine-learning model have a duty of care to make sure that it’s relatively robust against these sorts of attacks?

Carol: Yes, and I don't think you can necessarily prevent everything, but at least thinking through what the potential negative impacts might be and being prepared for them. I think in the past, a lot of organizations have created systems and set them loose, if you will, and then been surprised when they were attacked. That kind of naivete, that kind of lack of forethought, lack of imagination, is really dangerous. People making these systems do have a responsibility to make sure that they have at least done their due diligence as far as thinking through the risks and the ways that the system could be attacked and then how they are going to manage that. How are they going to protect individuals? How are they going to really protect the data that is in the system and think through that? The systems don't have rights and responsibilities. The people making the systems, the people operating the systems are the ones who have to be responsible for them.

Jonathan: Especially, I know that the legal stuff in the U.S. versus the, say, the E.U. is going to be very different. For any of our listeners that might be subject to, like, GDPR [General Data Protection Regulation] stuff, are there differences in what data you are allowed to use? Or do the aspects of bias being promulgated by machine-learning systems have legal consequences under some jurisdictions now?

Carol: Certainly. There is standing law—and I am not a legal professional—but there are standing laws that protect certain things and certain aspects across all kinds of different industries. Those laws still apply, of course, to AI systems. GDPR is helpful in many ways with regard to protecting people in the E.U. because of the ability to look at a decision that has been made by an autonomous system and to potentially be able to disagree with that and to get a different decision made or at least have the decision reviewed by a human. That is certainly one way of approaching it, but that also requires the person to have the resources to be able to do that. Even though those protections aren't as great as we would like. It does currently still fall on the people making these systems to make the right choices. But there are a lot of movements towards more regulation and more clarity there. But the existing laws also should protect, to some extent anyway, if they are applied.

Jonathan: So is there a code of ethics for machine-learning engineers?

Carol: There are many. There are many different organizations creating codes of ethics. The Association for Computing Machinery, ACM, has a set of ethics for engineers working in the field. The Department of Defense has a set of ethics that they are working on, doing more with, and right now they are working on adding more tools to the system so that people can really be able to use those ethics more clearly. There are probably over 200, at this point, sets of ethics across various industry organizations, nonprofits. Some governments are even coming up with sets of ethics, so there are lots out there. And some of them are very vague, unfortunately, and talk about *do no harm*, which is, really difficult to implement. Others are much more specific and helpful for the development of AI systems. So figuring out what that nice balance is to really help and empower the people making the systems to do the work that they want to do and not have to spend a huge amount of time interpreting the AIs is certainly the more helpful way to do this. But with each system, it is going to be a potentially a different set of ethics and a different way of looking at things. There is not going to be one answer that is going to work for every organization or every problem or every AI system. It is going to have to, to some extent, be a discussion. It comes down to people making hard decisions, but at least having those discussions and making those decisions based on a very thoughtful process versus not thinking about it at all.

Jonathan: Yes. So what I hear you saying is that machine-learning engineers, if nothing else, need to engage with and listen to their stakeholders.

Carol: Yes.

Jonathan: You mentioned the Department of Defense is putting together a code of ethics. Are there national-security ramifications for these sorts of bias in ML systems?

Carol: Yes. Definitely. Certainly from an aspect of just being able to use the systems. The wording they are actually switching to is more on *responsible AI* and thinking about *making systems that are responsible*. But the aspects that are truly important are just making sure that humans remain in control so that, if the system is not working as it was intended to, they can shut it off. That is probably the most important idea around responsible AI is the ability to override it and that humans are always in control, and that we aren't making systems that can do things without our oversight. That doesn't mean that the system has to ask for every single activity it does but, rather, that if systems aren't working the way they are intended, that a human can override it. A human can turn off the system, and work can continue as it needs to. And, specifically, when systems are changing—and an AI system, by nature, is dynamic, it is going to be constantly changing—being able to observe that change, understand that change, and, again, potentially revert to a previous version or do some other type of activity if it is not doing what is expected is really important.

Jonathan: How much of these are good programming principles and good software engineering principles for all systems? How much of it is stuff that is exacerbated by some of the features of AI and ML systems, and how much of it is totally new that we haven't seen before?

Carol: A lot of it sadly is new. I think, unfortunately, a lot of people in this industry and in software development haven't been forced to be more critical about the decisions they are making. Part of that is because AI is new. AI has the capability of bringing much broader sets of data, much more information together and then looking for those patterns within it, which is not something humans have been able to do in the past, necessarily. The other aspect is the impact that these systems can have is much more broad. So, both of those are very different. The scale and that sort of thing. A lot of things do change with these systems and, particularly, the dynamic nature. That creates a situation where, unfortunately, things have gone wrong, very wrong, with many systems and has exposed the risk that is there. That has driven a lot of these conversations. I wish these conversations had been had 50 years ago. Some of them were, but it didn't become part of the nature of software development until more recently, unfortunately, and so that is new. Newer than I would like.

Jonathan: Yes. So it sounds like we have had these problems for a while, but it is getting harder and harder to ignore them the more prominent the systems get. Is that part of what is going on?

Carol: Definitely. Yes.

Jonathan: If there is one conversation that you wish that the folks listening—software developers, whatever, business leaders—would have around these systems, what is the starting point for how we start to get a handle on improving this?

Carol: I think first just really looking at the problem or the situation that is there and understanding how people currently are doing the work. What is expected with an AI system? What is this AI system expected to do? Why is an AI system the right solution? And really just taking those steps to think through how people are going to use the system. How are people going to interact with it? How is it going to support them and make them more successful, more effective, whatever those measurements are? And being able to use that information then to build the right system and to build it in a way that is going to be effective and empowering for the people using it. It is often a step that is skipped, and people just say, *Hey, let's make an AI*. But really being more critical will not only save them a lot of time and energy, but result in a much better product. Ideally, because they are doing that process, they will along the way be able to identify those biases that are probably in their data and be more critical about rooting them out and making better systems.

Jonathan: That is really good. I know that we did a bit of a guide for people who wanted to implement ML in cybersecurity, which has a lot of its own specific problems because people will definitely be attacking, like, an antivirus detection system. And detecting bias is so connected with getting the results that you want anyway, right? Doing all of the things that you need to do to understand what can go wrong with the system by accident is also making sure you get the answers you are expecting in a lot of ways. Do you know of any good resources for a general audience on what initial steps, or what those questions are, that we could share?

Carol: We have been working on a checklist that is fairly approachable as far as just starting these conversations and having some conversations about the system, thinking through what you are trying to do, what the provenance of the data is. Really thinking about all these different aspects and trying to create a situation where the people building the system are able to have those difficult conversations together. That is really the first step is being more open to questioning the work and being more critical about the work as we go.

Jonathan: That is really good. What else are you working on that will help people root out this sort of bias in ML systems?

Carol: More activities around being...trying to make it more engaging and fun, to really be critical about the work. One of the activities that I have been working on is called *abusability testing*. It is the idea of really being more speculative, more creative in thinking about problems and going all the way to just really worst-case scenario that you can possibly imagine because that will help you to really think about more frequent, more common incidents and also prepare for those worst-case scenarios. Certainly by doing these kinds of imaginative activities, it also builds team cohesiveness. The team is more likely to be comfortable having more difficult conversations after those types of activities. It is a bit of working and having the team work together and also doing the work to identify those issues around bias.

Jonathan: That is really good. Now that we have got maybe some ability for the engineers to talk and have these hard conversations, what would we share with people about AI engineering? What can we do about guideposts or whatever. It's a very complicated...it is both simple and complex, right? The questions are not super hard to come up with the checklist, but the answers and how to answer them is complicated. So what else could we provide people with as far as AI engineering advice, do you think?

Carol: Yes, and that is such a great question because it is a huge challenge to even figure out where to start in a lot of this. The SEI has been given an ability to really focus on AI engineering as a practice and to really dig into some of these problems more deeply. That is what we are working on right now is figuring out how can we address some of these issues. How can we provide tools and at least really open up these conversations to our broader communities, to our DoD partners, and really help them to do this work in a more effective way and to utilize the tools that are available to them in ways that will help them solve problems and continue on their journey?

Jonathan: Great. Well, Carol, thanks so much for joining us today. To our listeners, we will, of course, include links in our transcript to all of the resources that we've mentioned in this podcast. Thank you 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.