Machine Learning in Cybersecurity: 7 Questions for Decision Makers

featuring April Galyardt, Angela Horneman, and Jonathan Spring

---

Welcome to the SEI Podcast Series, a production of Carnegie Mellon University's Software Engineering Institute. The SEI is a federally funded research and development center, sponsored by the Department of Defense and operated by Carnegie Mellon University. Today's podcast is going to be available at the SEI website at sei.cmu.edu/podcasts.

Jonathan Spring: Hi, my name is Dr. Jonathan Spring. I am senior vulnerability researcher here at the Software Engineering Institute's CERT Division. I am joined today by two of my colleagues, Dr. April Galyardt and Angela Horneman. We are going to be talking about machine learning in cybersecurity, specifically seven questions for managers and decision makers. Welcome, April and Angela.

Jonathan: Let's start off by telling the audience a little bit about what we do here at the SEI.

April Galyardt: I can talk about what I do.

Jonathan: Yes, go ahead.

April: I am a machine learning research scientist. I work with the Tactical and AI [artificial intelligence] -Enabled Technologies group.

Angela Horneman: I am Angela Horneman. I work on the Situational Awareness Analysis team. I am the team lead. We do a lot of work helping people understand their area of an organization in cyberspace and making sense of it to make better decisions.

Jonathan: I work on the Vulnerability Analysis and Threat team. We do a lot of things around vulnerability analysis, vulnerability management and coordination, and keeping an eye on the threat space and what is and is not really a big problem. Before we delve into the topic of machine learning, can we define for our audience what we mean by AI and ML?

April: This is not a trivial question. There are a lot of different definitions out there. One of the key AI textbooks has like eight different definitions in there.

But the one the Department of Defense has adopted is, *AI is a technology that can do a task automatically that would normally require human intelligence, such as translating a document*. Machine learning also has several different definitions, but machine learning is extracting information from large data sets, often to power an AI.

Jonathan: Thanks, April. Angela, can you tell us a little bit about the role that machine learning has to play in cybersecurity. Like, what is the state of the practice?

Angela There is a lot of using it to try to make sense of large data sets, because a lot of machine learning, specifically, is building information and models from large data sets. Cybersecurity has huge data sets. The idea is to try to use that to automate or to make better some of the processes that are currently going, whether it is incident response, malware detection, threat detection.

Currently, it is used well with some of the malware detection. A lot of the areas where it is trying to be used, for instance in incident response, it is a lot more iffy. It takes a lot of dedication, a lot of tuning, a lot of keeping up to date to get some limited utility from the solutions that are being put in place. A lot of organizations don’t have the resources and capability to effectively use it.

Jonathan: Anything you want to add, April?

April: Well, in some of the ways that Angela just mentioned, in cybersecurity it is not too different than how machine learning is in a lot of other applications. You have to stay on top of it, because things change moment to moment. There are a few things—maybe we’ll talk about them in a little bit—that do make machine learning and cybersecurity different than some other applications.

Jonathan: Let’s talk about our technical report, right? We co-authored this with Ed Stoner and Josh Fallon and Leigh Metcalf, and we outlined seven questions that managers and decision makers might want to ask. Can you talk a little bit of what the types of questions those are, like what they help the managers and decision makers figure out and do?

Angela: Yes. So, we have a few questions that deal with identifying your goals, and how you actually expect a machine learning or AI system to accomplish those goals. There are a few questions about how do you actually secure these systems. If we are talking about cybersecurity, securing these systems is part of cybersecurity itself. Then how do you actually evaluate what is happening with the tools? Are they really meeting those goals that you have and all the other needs around some sort of solution?

Jonathan: What are the ways that are different for a machine learning tool in cybersecurity for it to be protected than a machine learning tool in other areas?

April: The *to be protected* there maybe throws me off a little bit. Because when I am thinking about one of the biggest differences between cybersecurity and other applications, it is that often in cybersecurity the thing that you are looking for is a rare thing. That in particular makes it different than say machine learning on autonomous vehicles, because there you want to identify cars and people, and there are lots of cars and lots of people around. Looking for rare things is a little bit different.

Jonathan: Angela, can you tell us a little bit about contested environments and how that might be different?

Angela: There is also in the cybersecurity realm, a lot of…It is an adversarial environment. There are a lot of people trying to undo everything that you are trying to do. They want to deliberately deceive whatever systems you have. That doesn't go away when you have machine learning or an artificial intelligence solution. In some sense it gets worse because you don't have people, machines aren't yet at the ability to be able to detect interventions. Humans still need to do that. Humans are bad at that, but machines are even worse, so using machines in these areas just compounds the issues.

If you were thinking about something like a phishing email, it's one more area where machine learning is being put into good use. How do you detect some phishing, your email, your spam filters? The better you can make your email look like something legitimate, the less likely, and the harder it is, for any machine or an algorithm to detect the fact that it's phishing. We have environments where people are deliberately trying to make their interactions, their behaviors, their instances look specifically like something that it is not. That doesn't happen in a lot of realms.

Jonathan: Because this is so difficult, could we talk a little bit about the importance of sort of understanding what the tool is doing and getting an explanation for what the tool is situated in your environment to do and how it interacts with the other tools that you have in your cybersecurity environment?

April: Yes, so whether or not you can get an explanation at all and what kind of explanation you get, it depends highly on what tool you decide to use. That is actually one of the key things that you should consider when you are deciding what kind of solution you need, what your problem is. Because if you don't have, ahead of time, an idea of the explanation that you are going to need, and that kind of explanation, you might not be able to get it out at all, after you implement some sort of solution.

Angela: I think it goes back to the first few questions that we asked in the paper, which is, *What actually are you trying to accomplish? What do you need to know from your results? How is it* *going to fit into your larger context*? These systems can't exist in isolation. A lot of cybersecurity is very context dependent, so having just one specific data source or one specific topic, you're not necessarily going to be able to answer the question. For instance, about malware detection, you need to understand what actually can happen in your system, who things should be coming from, who they shouldn't be, what sort of programs you allow to run on your machines or not. If that context is not incorporated specifically into the solution, you have to be able to incorporate that, somehow, into a greater workflow. So that also impacts the interpretability you need of your results from your solution.

Jonathan: I think, maybe a couple of weeks ago you were saying that one other part of context is what the adversaries are doing now, and how much that informs what the adversaries might be doing in the future or doesn't inform? Because as you were saying, we know the adversaries are going to try to fool us. Can you talk a little bit about how time sensitive the training data then is too?

Angela: Sure. Cyberspace is a rapidly changing field. People change. They don't do the same thing all the time. We evolve our actions in response to our environments and the results that we get. A lot of things may need to be trained continually. The people who do malware signatures for something like a virus scan, they are training their modules continually because they are constantly getting new examples of malware, or phishing, or whatever the case may be, bad websites. Those models are updated continuously. There are some things that could go a little longer, if you can actually get a good data set to begin with, but because the area is so volatile.

April: Well, kind of on the malware example, and the taking context into account, there are a lot of malware tools out there and academic papers that are like, *Well, I trained my machine learning model to identify malware.* Well, they went to a giant malware repository and trained their model on that. That is all the stuff we can already find using the models and methods that we have. So, when you get high accuracy on that data set, that doesn't necessarily mean anything, other than that you can recover the tools that already exist and that you don't know what you didn't find. That is an example of something you don't want, because that is not the right training data.

Jonathan: That sounds like a similar problem to what we already have with evaluating black lists, right? It is already the case that it is really hard to say what new contribution some sort of detection mechanism has because they all are sort of finding different things, and that is the intention. Does machine learning do anything to make that problem better or worse?

Angela: Scalable. I think one of the most compelling reasons to use machine learning is scalability. How do you deal with all the data that exists? You can't do that as a human. You can't do that with most of the existing non-ML algorithms. The biggest draw for a lot of people for machine learning and artificial intelligence is being able to scale up to the size of what we are dealing with.

April: I think that is exactly it, but what I was thinking was that you are asking does it make it better or worse, and the answer is well, *it depends*. If you are not incredibly careful, it can make it much worse. If you are careful, you can make it better and achieve that scalability.

Jonathan: Can you maybe tell us a little bit about how the questions that we suggest people ask reduce these risks to increased scalability, and let people do it a little bit more safely?

April: One of the questions, *How would you find and mitigate unintended outputs and effects?* That is a big one: anticipating, or at least trying to anticipate, the unintended consequences. Going back to the rare events, if you have a lot of rare events, your algorithm is very likely to detect a lot of things that are not, a lot of false positives. Having something in place to deal with all of those false positives or training your model so that you can reduce that number of false positives, those are necessary things to make your system useful and scalable.

Angela: I think also the first three questions, which are *Topic of interest*? *What information will help you address the topic of interest?*, and *How do you anticipate that an ML tool will address the topic of interest?* Those are not surface questions. Those are actually looking for you to get down deep, to talk to both SMEs [subject matter experts] in the area where you are trying to solve the problem, as well as data science SMEs and software engineer SMEs. How can you actually build the system to make sure that you can not only address what you need to do, but get the correct output. Do you even have the resources you need to be able to tackle this?

Jonathan: Is there a strong sort of like data curation dependency on being able to successfully have a good machine learning deployment?

Angela: Definitely data curation, but even beyond data curation, even understanding, *Is the data that I have available, or that I can get, does it contain the answer, or can I use it to create a model for the solution I need*? That is still a huge question, and a lot of the use cases where people want to apply machine learning for cybersecurity, specifically in the realm of general threat detection, not necessarily malware detection, but malicious traffic, something like an IDS [intrusion detection system]. A lot of the data we have doesn't really include the information that you would need to make a determination, which would be something like context. *What is actually supposed to happen?* Whether your policy is, *Who should be doing what? Who should not be*?

April: Right. Think of even a simpler example: if all you have is the amount of traffic that is going through the network, that can maybe tell you there is something anomalous, but that doesn’t give you a whole lot of information on what is going on. Maybe you just have a lot of customers at the moment.

Angela: Yes. Most anomalies are not malicious. Most anomalies have nothing to do with security. They are just general ebbs and flows and new behaviors and explorations.

April: That’s why I bring that example up because, if that is the only data you are looking at, then, you are probably not looking at the right thing.

Jonathan: Changing gears a little bit, can you tell us a little bit about the attacks specifically against machine learning tools to get them to intentionally misclassify events?

April: It depends on what you have access to. If you have access, and you can inject something into the training data...

Jonathan: You mean me as the attacker?

April: Yes, sorry, you as the attacker, if you are trying to attack my system.

Jonathan: I would never.

April: You would never, but let’s suppose that you were.

Jonathan: Please.

April: You could try and attack my training data and inject things there and alter it in some way. You could try and attack my model itself and adjust some of the parameters in the model. If you don’t have the ability to attack either the data directly or the model directly, if you looked at what is coming out of my model and just look at that, that can maybe tell you where there are some holes. So, that is the kind of attack where you get, in the self-driving cars when you put stickers on stop signs, suddenly it looks like a speed limit sign, and the car does not stop and in fact speeds up. You can make adversarial examples that will then make my model misbehave on those examples. Those are kind of the three big classes. There is some variation there.

Jonathan: Within those classes, just to be clear, are those hypothetical examples, or are those demonstrated attacks that have worked?

April: There are many, many academic examples in each of those categories. We are starting to maybe see some of them in the wild, out in the world.

Angela: I think, specifically, for the one where you are attacking the training data, a lot of the methods currently used for building models, the training data is captured live, which isn’t clean to begin with. So, it is not a deliberate attack against the training data. But because it already includes some of the stuff that you don't want or that you are detecting, you don't necessarily know it is in there to have labeled it correctly or, depending on, maybe you don't have label data. You just don't know it is in there. It is an indirect attack.

Jonathan: That seems like if you are learning based on your network traffic and the adversary can send you packets. That seems like that is an open vector. Is that right?

Angela: That is correct. Another way that is not necessarily machine learning specific—but in the network traffic or IDS example, or even something like a phishing—if you just sent enough traffic in general something is going to get through because people can't deal with it. A lot of the output is still not auto handled. Even when it is, some of it is going to be misclassified. If you send enough, you are guaranteed to get some sort of misclassification that gets through.

April: Well, in the examples that you guys raised, there is one other thing that I wouldn't necessarily call an adversarial attack. Feedback loops can cause their own problems. Even if nobody is actually attacking you, if I have an analyst who is looking to code, and their analysis of the code goes into my next training data, which then feeds back into the next thing they see, sometimes feedback loops like that can wind up in very strange places.

Jonathan: Angela, you mentioned earlier, that it is rarely the case that anomalies are security events. Is this sort of feedback loop and ability for the adversary to change what is normal sort of another problem with treating weird things or anomalies or what is normal as what is security or what is secure?

Angela: Yes. Humans are imperfect in what they do just as the machines are. We also are very time constrained. We can't effectively go through everything that we are documenting. So sometimes we take shortcuts. We go through and say, *Oh, all these look the same, I am going to close them all with the same resolution*. If that is being fed into your feedback loop, and it was not 100 percent correct, you have just shot yourself in the foot.

Jonathan: One thing that we do love to emphasize with the podcast series is transition. So, if I am an audience member, no longer an attacker but an audience member who is interested in using machine learning tools in cybersecurity, can you tell me where I should start please?

Angela: From my perspective, where you should start is really think about what you are trying to do, what you expect to accomplish. *What do you need to accomplish?* Don't worry about whether it's an ML or AI problem to begin with. Just really try to document in detail what it is you are trying to accomplish.

April: I think that is absolutely step one. Step two then following from that is, *Do I have the data that I would need to start working on that problem?* Because as soon as you have that data and start collecting it and even just counting what is in there, number of events, defining what is normal, you start to actually have actionable information even before you brought machine learning into it.

Jonathan: Are there resources that are available to help with these sorts of things?

April: What your problem is and what data you need, that is very institution dependent. But, there are lots of off-the-shelf machine learning tools available, and a lot of them are much better and much easier than trying to build everything yourself. So, yes, absolutely. Once you know your problem, once you have your data, use the off the shelf tools to get up and running.

Angela: I wanted to back up a little bit about the data. One of the things is as you are evaluating the data you have, a good rule of thumb for machine learning right now is if your people can't find the answer in the data, the machine definitely isn't going to. Which can give you some insight into what other data you may need to put into, how you may need to combine it or where else you may need to look or even if it's a good topic to begin with.

Jonathan: But, what if I need a panacea because everything is broken, and I don't have the people that can fix it?

April: There still are no silver bullets.

Jonathan: Looking ahead then, where are we headed with this work?

Angela: I've been doing a lot of work on trying to define AI engineering. There's a paper out right now, the 11 fundamental practices (that's not the title of the paper), but basically, they're starting points about how do you actually start building your system once you have done some of your preliminary work. What else are we working on?

April: I have got two research projects that are really working on how we build and sustain the full ML system. One of the things that happens right now is that you have a machine learning scientist who, they build a model, and it's great. Then they have to hand it off to a developer and the developer doesn't know anything about machine learning. We talked about data drift and how data changes over time. You get to operations and you need to go back and check and see if it's time to update. But you can't check and see if your model is still good because the developer didn't know to put the right hooks in to make those checks.

So how do we streamline that communication from the ML developer all the way to operations and make sure that everybody has the information that they need so that nobody leaves the hooks out, and formalizing that communication. That is one of the projects that I am looking at. Then there's another one that, so now you have somebody on the other end who has gotten a report out of the ML system. Now they have to make a decision based on the system's recommendation.

How do we present that information to them so that they actually interpret it correctly? That is nontrivial. Humans don't reason well with uncertainty. There are tons of psych research that shows that. How do we build this whole system to function well?

Angela: Specifically, on the cybersecurity side, we are looking into trying to figure out how do you actually incorporate context into not just ML, but systems in general related to cybersecurity. Obviously, it's very a very big part of what you need for an ML system, but it is also needed in general. A lot of the analytics don't currently work. They are anomaly detection. You get a lot of stuff that has nothing to do with security. How do you actually incorporate your policies into your analytics or into your analytic processes and workflows?

Jonathan: It sounds like in general we have the questions that people should ask, but we don't have all of the answers fully laid out yet.

April: I think that is a good way of summarizing it.

Jonathan: I know that we are also, with the vulnerability management side, thinking hard about how the AI community will manage, respond to, coordinate, and fix flaws in ML-related software. So as with any other piece of software, your vulnerability management is going to be a key part of your resiliency for your ML tools as well. In some sense nothing new there. In some sense, we have a whole new community of people who are building software who need to be brought into that.

April: Your average data scientist doesn't know anything about building good software and architectures and that sort of thing, which brings us back to that pipeline problem.

Jonathan: Yes, you have got to bring several communities together here. I think that is one of the things we're good at with the SEI, I mean software engineering. That is the thing.

Thank you so much for joining us today. As we mentioned, Angela, April, and myself along with Josh Fallon, Leigh Metcalf, and Ed Stoner have published a technical report on this topic. Go to sei.cmu.edu and type *machine learning cybersecurity* into the search box, and it will definitely come up for you straightaway. We will also include links to all these resources in the podcast and the transcript. Thanks for joining us and take care.

Thanks for joining us. This episode is available where you download podcasts including SoundCloud, Stitcher, TuneIn Radio, Google Podcast, and Apple Podcast. It is also available on the SEI Website at sei.cmu.edu/podcast 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.*