Carnegie Mellon University
Software Engineering Institute

SEI Podcasts

Conversations in Artificial Intelligence,
Cybersecurity, and Software Engineering

From Data to Performance: Understanding and Improving Your AI Model

featuring Nick Testa and Crisanne Nolan as Interviewed by Linda Parker Gates

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.

Linda Parker Gates: Hi. Welcome to the SEI Podcast Series. I am Linda Parker Gates, and I am the initiative lead for the Software Acquisition Pathways work under the SEI Software Solutions Division. I am here from the Washington, D.C. [Arlington] office, and I am joined by Crisanne Nolan and Nick Testa, who are both new to the podcast series. We all three worked together on a project that we are excited to talk to you about today. It is a tool for helping programs better understand and improve AI performance. I would like to start by asking you both to tell me a little about yourselves and your background, how you came to the SEI, and what kind of work you do here. Nick, why don't you start us off?

Nick Testa: Yes, I am a bit of an oddball here. I got my PhD in evolutionary biology, studying the molecular, genetic, and developmental mechanisms that underpin shape and size. I was looking at a lot of really cool stuff that was kind of scary to a lot of biologists. Size is actually pretty well understood. You just weigh something, right? That is easy. But shape requires a lot of weird machine learning and crazy like weird statistics that start to take you into places that biologists don't really work. That actually provided a really good bridge into data science whenever I decided full time academia wasn't for me. As I worked more and more in industry, I started to miss that sort of academic touch. The SEI here had exactly what I was looking for. We work in this weird mix of government, industry, and academia. I felt really at home here in the kinds of work we do. I love that. My current team, I am the senior data scientist on the team, and I do a lot of work on causal learning. That is kind of the one big thing that I got pulled in on.

Linda: You are going to hear a lot about that today.

Nick: You are going to. Yes, exactly.

Linda: Thanks, Nick. Crisanne tell us about yourself.

Crisanne Nolan: Yes, I am happy to. I have been at the SEI for quite some time in a number of different roles. My background is in communications and technology transition, so I started in a transition and training team. My current role is Agile Transformation Engineer, where I work in the Continuous Deployment of Capability directorate. I take agile and DevSecOps principles and bring them to bear in really complex cyber-physical system development—systems of systems that don't necessarily lend themselves to agility, but trying to shorten things like learning cycles, important things like that. I will say too that one of my favorite parts of being at the SEI is working with experts like Nick who have very different backgrounds, whether it is data science or cybersecurity or architecture, all sorts of expertise, subject matter experts across the board. We get to team in really diverse ways and look at these really hard problems together and think about holistic answers, holistic solutions in different ways.

Linda: Well, I have a liberal arts degree in computer science and a master's from CMU in architecture. While we will briefly talk about bridges today, that is a pretty interesting mix just of the three of us. And we have got three times as many folks on the team from really interesting backgrounds. We have come together to work on a project about understanding and evaluating AI, artificial intelligence and machine learning predictions, which can be really challenging. There are things like concept drift, data drift, edge cases, and other emerging phenomena that can really throw off training data, in addition to just the natural uncertainty of machine learning outcomes that can start to bring bias into predictions and compromise the decisions that you are making based on those predictions. We have this tool, it is called AIR [AI Robustness Tool], and it is based on a method developed here at the SEI that uses causal learning to look at problems that are built on data correlations. We are working with some partners, because this is fundamental research that we are doing. We are in a late-stage research project. We are working with some partners, and we are looking for partners to work with us, and you will hear more about that as well. Let me ask you first, Nick, why should we…If I have an AI or ML model, and I am using it to make some decisions, why should I be concerned about bias?

Nick: Well, one of the most important reasons, I guess, is when you are training an AI or ML model, what you are training it with is a snapshot of the current context of the data. You are trying to build a model that is really good at predicting the outcome of events. But those events are some snapshot. Even if you take it over time, it is still a snapshot in context, and it is still bounded. But the real world isn't bounded like that. The real world is really messy. You can miss a lot of context that can add bias or that can add some sort of weird confounding factor to your data. And if you don't understand exactly how and why your data are interacting, then you could miss something. You could be really confident in your answer, and you could still be confidently wrong. And that is kind of what our tool that this thing that we're building.

Linda: There is a lot of buzz about AI right now. I am going to bring us up a level again because there is a lot of buzz. In fact, there was a recent memo from OMB, about accelerating federal use of AI through innovation, governance and public trust. One of the things the memo does is it really recognizes the urgency of monitoring AI for potential adverse effects. We are really working in that space. The three of us recently—I promised we would talk about bridges—the three of us recently came across an article about AI and ML robustness and in particular, data drift. Nick, why don't you tell us about the article or kind of what the phenomenon is that the article discusses? Because that is really why I am bringing it up.

Nick: Yes. I mean, of course. We are in Pittsburgh. We are going to be talking about bridges. That is a given. That is why we are drawn to this paper. This paper was from some researchers in South Korea. What they found was that they were studying the effects of earthquakes and other seismic activity on bridge stability. They were trying to figure out when bridges are likely to fail. I think this is for, like, first responders and things like that. What they found was immediately after they built this model, everything was functioning really great. And so the models are predicting everything with high accuracy. Then, as the model aged, their predictions started to kind of skew. They were a lot less reliable than they should have been. Through some subject-matter-expert digging and investigation, they figured out that the reason is because a lot of these bridges are corroding at different rates or in different ways that they weren't expecting, whenever they actually made the model. You have what is actually data drift The data on the bridges is drifting because corrosion is causing the rebar to be less stable and the cement.

Linda: So your bridges are degrading, but your model is also degrading because it is based on data that is no longer accurate.

Nick: Yes. Exactly. That data drift is it is one really important thing that we need to watch.

Crisanne: It is such a great example for government organizations we have been engaging through this project, that they also have challenges like that where they start with great data, but they are making high consequence decisions for things like predictive maintenance or workforce planning, where they can't wait for degradation to be apparent and to apply a SME, subject matter expert, expertise to try to fix it. They need to know as early as possible, and so they are looking for best-of-breed options, to try to learn more up front.

Linda: We are working with a really wide…I mean, the applicability of the tool is really broad.

Crisanne: It is. Because this the sort of problems Nick is describing happens with data, in many different areas bringing that data to model.

Linda: As broadly as AI is used, you got these kinds of situations popping up.

Nick: Right. And then when they are high stakes like this where lives could potentially be on the line, that is really important to get right. We don't want to wait for the patterns of corrosion to appear and have to figure that out. We don't have time for subject matter experts to do this. We would like to get a step ahead of that.

Crisanne: It is critical to find that data drift.

Linda: Right. We are helping with more than just data drift, right? So, Nick, can you take us on a little adventure about bias and correlation and just explain what the environment is that brought us to bring causal learning to AI.

Nick: Yes, so obviously it is more than just data drift, like you mentioned. Bias can be introduced as well if we are not careful in what our data looks like. It can sneak in really silently. It can be kind of a silent killer for your models. Things like confounder bias and collider bias. We can have weird paradoxes.

Linda: Why don't you give us a scenario and then maybe walk us through the different kinds of bias that one might encounter?

Nick: Sure, that is a good idea. Imagine that you are in charge of this fleet of vehicles. You are in charge of maintenance maybe for whatever kind of vehicle we are looking at. You can have a thing called confounder bias. That is one that is one really good thing to pay attention to that we want to watch. Let's say you are trying to figure out what causes failure rates or higher maintenance events in your vehicles. You might notice, *Isn't it weird that that that the volume of engine coolant that we add every day seems to be correlated with maintenance events or failure events of these vehicles?* And if you weren't accounting for something like heatwaves or maybe just increased temperature, you might not notice that that temperature, those heat waves, are a common cause for both needing that increased coolant volume and then also the engine failures.

Linda: That is like the sharks and ice cream thing, right?

Nick: I love that one.

Linda: It is easy to understand. You have got an increase in ice cream sales that is tracking with basically shark attacks. You might start to think, *Oh, okay. The more ice cream we sell, the more likely sharks are to attack*. *We should close down these ice cream shops*. You can go crazy with that. Really there is a common cause. There are higher temperatures or something like that that is causing both the increase in people wanting ice cream and sharks hanging out nearby.

Nick: Yes. It is more than just temperature increases that cause this. Even though both of our examples are…

Linda: Simple. I am a simple thinker.

Nick: Keep it easy then.

Linda: So that is one. That is one common cause situation.

Nick: That is one way the bias can sneak in. Now, one might seem really obvious, like, *Oh, of course we will never make that mistake*.

Then you have other things like collider bias, which are a little weirder. Let's say you look at these vehicles every day, and you do a maintenance check every day on these vehicles. And so you prioritize a lot of the vehicles that have that are flagged as needing maintenance. Then you also prioritize vehicles that are flagged as mission critical. *I am going to need to use this vehicle in the next couple of days. I want to make sure it is ready to go*. What the analysis will end up turning out is kind of a weird negative association between how critical that vehicle is and how likely it is to need maintenance or to fail. You end up with this kind of false negative correlation because you are stratifying on, *I checked it today* instead of *the whole fleet*. That is called collider bias. Filtering through a gate like that can kind of create these weird spurious correlations in your data.

Linda: I imagine the bigger your data set is, and the more variables you have, the harder it is for these things to be obvious to a subject matter expert who is looking at the data.

Nick: Yes. And these things aren't going to always be quite that obvious. A lot of times they are going to be hidden down in some weird data interactions that you hadn't even considered. That is why that is important to look at.

The last piece of bias I want to talk about is one called Simpson's Paradox. This really happens when you have pooled data that you maybe you shouldn't be pooling. Let's say in your vehicle fleet, you have got vehicles that are active in temperate zones and also desert zones. And you are trying to figure out how well use correlates with the need the need for maintenance or failure events sort-of thing. You graph it out and you see, *Oh, it looks like the more I use this vehicle, the less it needs/requires maintenance*. And that is really odd, right? Obviously you would scratch your head and say, *Something is wrong with this*. What ends up happening is when you pool the temperate data in the desert data, you can get this weird false correlation because all the tempered vehicles have low failure rates because they are not operating in such an extreme environment.

Linda: They are not under the stress.

Nick: Yes, exactly. So if you are not accounting for those sorts of things, then it will look like that. Whereas really the high ones are just the desert, and the low ones are the temperate.

Linda: And then your correlations are correct.

Nick: Right.

Linda: That is wild.

Nick: If we miss these biases, we can spend a lot of money on the wrong fixes and we can schedule work wrong. We can still get surprised by these failures that we shouldn't be surprised by. None of this stuff should be surprising if we do our homework. If we correct them, we get closer to the truth of which actions actually reduce those failures for whom and under what conditions? I was just going to cap off this statement with this AIR tool that we are building helps us build out a causal map to understand what factors we need to account for and which ones we want to actively not filter on and or account for. Then we can estimate the maintenance effect as though we had run like a fair randomized controlled trial almost. So it has the effect of turning your AI/ML model into a randomized, controlled experiment without all the hassle of actually doing that. That is kind of the gold standard of what we want to do when we are trying to understand cause and effect.

Linda: Right. To put another cap on that. Correlation is the most common structure of AI and ML models. Hearing about the ways that correlation can be risky or problematic is the reason that we have this tool based on causal learning. Nick, can you take us under the hood a little bit on how we are bringing these causal learning methods together to try to uncover some of this bias that you described.

Nick: Yes, I would love to. It is really cool. What we are doing is we are taking a lot of kind of cutting-edge work in the causal learning space, and we are chaining together three kind-of disparate fields within causal learning into an end-to-end automated process. What we are doing is we are taking causal discovery, and that is the ability to take your data in and build out a map of the cause and effect within your data. That way we can see what the causal links between individual variables in your data. And so that is cool. And on its own that is helpful. But then we go into what is called causal identification, and we use some state-of-the-art methods there too and algorithms to do some graph manipulations to figure out just where that bias lives. So a lot of this confounder bias and collider bias can be detected. And we can detect the signals just using the graph that we have produced by that causal discovery graph. And then we can use that, and we can use those flagged nodes or variables, and we can input them into a causal estimation, which again we are using really kind of state-of-the-art technology here to estimate, *What if we had the effect of X on Y. What if we could predict that outcome without the effects of bias?* That is what our tool is really doing is it is taking you from your raw data all the way out to the prediction of your effect without any of that bias, like a randomized control trial.

Crisanne: Once you have that prediction, you can then take it back to the model that you are using in operational real-life contexts and compare and start to learn more about your model and where to go next.

Nick: Yes. Absolutely. We have a couple different options for the users. They can either let the tool take them on a ride. You know and it'll do all that that stuff for them. You know they can input their model, or they can, they can even do some calculations for treatment effects and then compare, *Hey, how well is my model actually functioning?* And if it is not performing as well as it should, it'll flag the proper adjustment sets and the proper, the proper variables on that map It is like X marks the spot on a treasure map. It will tell you exactly where to find the issues.

Linda: Crisanne, you talked a little bit about some of the domains where this is applicable, and there are many. We are working with people doing engine maintenance, people doing workforce planning. Can you talk a little bit more about who might be interested in using this tool, where it might be of value?

Crisanne: Certainly. As Nick has described, moving from data to predictions is really the key, and that happens across many fields and specialties: things like logistics, mission planning, even cybersecurity. Everyone is interested in knowing what the impact of *this* might be on *that,* and they are applying AI to that problem set. We have something to learn almost regardless of the context of where the data comes from. Really, we are just looking at the causal relationships. That is applicable throughout the government.

Linda: One of the things—because I mentioned we are kind of in a late-stage research project where we are starting to work with real projects—what makes a project a good fit for AIR? Because there is sort of the methodology then there is this tool we have out on GitHub. What makes a project a good fit for the tool, being able to run that tool?

Crisanne: Absolutely. Yes. It does start again with using a classifier, some kind of model that makes predictions. It might be a classifier, but some model that you are looking to help predict what happens because of a cause-and-effect relationship. *What is the impact of X on Y in the future?* You want first of all, to be interested in that kind of question. You should be able to get access to data, things that tell you about what is happening in either the time when you train the model, or perhaps today, at a later time than when you created the model, but the right kinds of data that reflect the factors of the causal relationship in that scenario. You want to get as clear of a picture of what is going on in real life as you can. Then a little bit of subject matter expert. Someone that can weigh in and start to guide the tool to understand some of the basic relationships between those variables. To go back to your bridge example. You don't need the sort of deep expertise of people who have been studying just those particular variables forever. The tool does a lot of work to help supplement the subject matter expertise, but a little bit going in is useful.

Nick: And they don't need to be experts on causal learning or anything either. It is the other thing. We take care of that part.

Linda: Right, That is a good point. Really it is a way for us to bring that specific knowledge to, to a problem space. There can't possibly be any caveats, can there?

Nick: Yes, this isn't a silver bullet, right? This isn't some sort of panacea or cure all. What it is really good at, it is really good at. So it is really good at being able to find and detect those sources of bias and tell you exactly where they are. What it is not going to do for you is solve those problems. It is kind of more like a check-engine light. It is going to tell you why your model is failing and where. It will tell you, *Your model is failing because of the rebar is corroding, and that is and that is a problem*. *The data looks wrong here. You need to do something about that.* That is not something that our tool will do for you.

Linda: Right. How about from a user perspective? Are there any scenarios where this kind of is like…I guess I am thinking about image data and some of the things that the tool is not quite ready for.

Nick: Conceptually, we could get there. Especially on image data and LLMs, we are not quite there yet. At least the version of the tool that we are building out now can't handle that. This is really geared towards predictors where we are really trying to understand what the outcome of an event is going to be and how an event can affect that. That is really what we are here for.

Linda: OK, great. If someone would like to use this tool or work with us, how would they do that?

Crisanne: Certainly. Well, if someone would like to try the tool themselves, we publish our prototype out to the SEI GitHub. That is available for anyone to access and try. We have instructions and user help there. to get people up and running and to start to explore what the prototype is capable of, where our project team is continuing to mature that. So we expect it will evolve. And so we continually publish updates there. So that is the best place to check out the tool itself. As you mentioned, we're also looking for partners. So an organization that might think this is an interesting approach or feel like this meets a mission need within their work can reach out to us Partners have the benefit of working with our team, SEI experts, they can kind of lean in and work shoulder to shoulder to get things kind of stood up and to make sure, things are fine-tuned appropriately. And we can help analyze the results. All of that kind of fine tuning that might make it more useful more quickly for someone out in the world.

Linda: And free of charge. Because this is a research project, and so we are fully funded. We are not looking for partners to fund work. We're looking for partners to just work with us. It is a nice offer.

Crisanne: Exactly. Yes, and we ask them not to pay us, but to keep us updated. Yes, share their experience to give us ideas of how it could be improved, how other organizations like them might find it to be more useful or user friendly. And so, making it as bulletproof as we can for a broader adoption is really the goal of the project.

Linda: Is there anything we didn't cover that you wanted to talk about today?

Crisanne: I'll just add that we are doing email announcements to people who are interested in this research. For ongoing updates on new publications or new updates to the prototype tool, if you email the SEI and asked to be included, we're happy to keep people updated on the work as this moves forward.

Linda: We will include those links and contacts in the podcast information. Thanks, Crisanne. For our audio-only listeners, if there is more that you would like to learn about what we're doing, please don't hesitate to email us at info@sei.cmu.edu. OK, so thank you both for this conversation today. To our listeners, thank you for joining us. Again, we'll include links and also a transcript of our conversation and the things mentioned in the podcast. The SEI podcast series is available in places where you can find podcasts, Apple podcasts, SoundCloud, Spotify and the SEI's YouTube channel. And as always, if you have questions, please don't hesitate to email us at info@sei.cmu.edu. And thank you so much.

Thanks for joining us. This episode is available where you download podcasts, including SoundCloud, TuneIn radio, 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 e-mail us at info@sei.cmu.edu. Thank you.