Uncertainty Quantification in Machine Learning: Measuring Confidence in Predictions

featuring Eric Heim as Interviewed by Suzanne Miller

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.

Suzanne Miller: Hi, my name is Suzanne Miller. I am a principal researcher here at the SEI in the Software Solutions Division. Today I would like to welcome Dr. Eric Heim, a senior research scientist in machine learning at the SEI's Emerging Technology Center [ETC]. Today, we are here to talk about Eric's latest work in artificial intelligence regarding uncertainty quantification. So welcome, Eric, and before we get started, tell us a little bit about why are you here? What is it about the work at the SEI that brought you here and what is it that you do here on a day-to-day basis.

Eric Heim: Sure, well, first, thank you for having me. As you mentioned, I am a machine-learning researcher at the SEI. I characterize my work as focusing on bridging the gap between some of the scientific and mathematical advances from the machine-learning research community to the people who are more practitioners, the people that are actually going to use these in real-life systems. So, this means that we are taking stuff from the Machine Learning Department and the Computer Science Department here at CMU, that type of really awesome fundamental research, and trying to bridge the gap between them and people like software engineers, software developers, data scientists, and the people that are developing these ML systems that are going to get deployed. I will note that this is actually both a science and an engineering and also probably an art to some degree to bridge such a gap. So, there is actually a lot of work to be done both on the scientific and the engineering side here.

Suzanne: What brought you here to the SEI instead of doing this somewhere else?

Eric: Yes, so in my background, I got my Ph.D. at Pitt [University of Pittsburgh], so I am familiar with the area, and then I worked for 2 1/2 years in the Air Force Research Lab's Information Directorate. So that is how I got my foot into government research, specifically machine learning. There I led a basic and applied research group in machine learning. We wrote papers, applied it to some of the Air Force's problems. We developed systems, and to get back here to the SEI, I wanted to stay within the government-research framework, and the SEI is a great place to do that. Also, I am familiar with the area. So it naturally called me back here to Pittsburgh.

Suzanne: Good. I always like to get a little bit of that from people because we have got lots of listeners that are probably like five years behind you, and we want those good people to come to the SEI too.

Eric: Absolutely. Yes. Yes.

Suzanne: All right, so we are here to talk about AI. We know that is a priority in the DoD, in the federal government, and there is a specific aspect of that called quantifying uncertainty. Why don't you frame what the boundaries are of that area of research within artificial intelligence for our viewers.

Eric: Yes, so just taking a little bit of a step back, when you talk about machine-learning models, a lot of people view them as these things that make predictions about data. At a very high level, you can view machine learning, a model, as one that is trained on some data to make some informed or inference on some input data. So it is going to make some inference or prediction about that data.

Suzanne: And the data that it is getting is not necessarily the data it was trained on, right?That is the uncertainty part.

Eric: That is right, yes, very much so. Maybe grounding it a little bit more, a popular example of this is image classification. The model accepts some images, input and outputs, some label that indicates what is in that image. An example that is fairly accessible is you can train an ML model and give it some data then give it some new images of, say, a dog, it may be able to predict the breed of that dog. So, give it an image, the output is *golden retriever* or something like that. In a practical sense, one of the challenges for these ML models that do this type of task is that the tasks they are trying to do are naturally challenging. It is not always going to be the case, and it's actually very much expected, that there are certain situations or certain inputs for which the model will be wrong.

Suzanne: If you give it a picture of a cat, we would not expect it to be able to do the job, right?

Eric: Right, or even golden retrievers that you would expect it to do well, it still for various reasons, it could output the wrong answer. You have given a great example of something that is out of distribution. It doesn't have any idea of a cat, because it was never trained on a cat. That is a situation where it is going to be wrong. It is not clear when you are developing and training these systems always that when these models are going to be wrong. Basically, it is not easy to predict when it is unlikely to be correct.

One of the goals of uncertainty quantification is to develop these techniques that are going to be able to express some degree of uncertainty in their predictions. So, I will say the words *uncertainty*, and conversely, *confidence*. They are at the opposite ends of the spectrum, but basically the idea is we want these models to be able to express that alongside of these predictions. So not just *golden retriever*, but *golden retriever* with some confidence associated with it. In doing so, the result [is that] at the end, the consumers of those ML predictions—whether it be a human being who is trying to understand what the model is doing or use those inferences, or a software component—they are going to more reliably understand what is going on with these models and be more informed about when they are potentially going to be incorrect.

Suzanne: I will just give a very practical example for that in the DoD is, in image classification, for example, if I am presenting my model with an image of someone that I think is a threat actor, is a threat, I want to know, *Before I go and take this guy down and put the handcuffs behind his back, I want to have some idea, is this 30 percent confidence that this is the person I'm looking for, or are we in the 90 percent range, because I'm probably going to act a little differently between 30 percent and 90 percent.* So, that is the kind of thing that I look at as being necessary about this.

Eric: Sure, I give the toy example, but you are speaking to a point about where in the DoD this lies. To a point, intelligence analysts' tasks or types of ISR tasks are the most obvious ones. And, so, you mentioned a great one there. You can think about other intelligence analysts' tasks that are looking at data with the aid of some machine-learning model. Instead of just taking the black-box outputs like you mentioned, having some idea of confidence to make better informed decisions is key to a lot of different tasks, both in intelligence analysts but other potential tasks that you should think about from the DoD.

Suzanne: There is a multiplier factor here, right? I have got this model that helps me with eyes, and this model that helps me with noses, and this model that helps me with ears, and they all each have some level of confidence, but then when I am putting them together to try and identify a face, then I have another factor in terms of the aggregation of those different elements that may

also create, confound, the results. How does uncertainty quantification deal with that aspect, that we are not dealing with a single model in many of these complex cases?

Eric: Yes, so what you are speaking to is that fact that oftentimes, ML models aren't used in a vacuum but used in the context of a larger system, right? To your point, it is important to understand the various sources of uncertainty, whether it be from multiple models, interpretations of the output, noise of the input, things like that. Uncertainty quantification writ large, can consider just in the individual model case, which is really important to quantify there, but in the aggregate, how do you eventually take those and make some action based off that inference, you know, what's going on? Classically speaking, there's this way to think about it in a systems context, with your systems engineers, to understand how to interpret types of probabilistic events within a system. There are also particular ways you could model the aggregation of the data. So, ensembling methods or some specific way of understanding the individual outputs in some structured way to be able to produce some more consolidated estimate of uncertainty. It is something that certainly is possible.

Suzanne: Where is the research in this quantification of uncertainty, because there are things I know about. There is this probabilistic thinking kind of thing that floats all the way through everything in AI, and that's a foundational aspect of it, getting people accustomed to probabilistic thinking, but what are the things that are out on the edge that you are working on in the ETC?

Eric: Yes, it is interesting. You referenced this idea of that uncertainty in probabilistic models is something that the ML community, in a fundamental way, are very comfortable with. So, dating back to the roots in statistical machine learning, there are actually a lot of fundamental ways to reason about these models.

I think part of the research now is how do we get the same foundational understanding of these techniques, in the modern machine-learning realm, where these are very large, specifically deep neural-network models. A lot of the academic literature is trying to take some of the things that you are seeing, some of the types of reasoning about the simpler models in this deep-learning regime. So, that is one thing that you are talking about. I think it is important to recognize that what we are working at is trying to take some of the stuff, which is brand new, really a focus now from the academic literature, and trying to make it practical to the people I spoke about before. One way of doing that is taking this lens of use cases for which you may want to use uncertainty in your system or to express to a person, and deriving some principal statistical measure of uncertainty for which you can use. So, for instance, you may not care about the entire probability output of a model. You may care more about a specific subset. *I care more about these classes than these classe*s. Or, *I may care about this probability interval.* Or, *I may care more about knowing the top three predictions, not all potential hundred classes I could label*

something in. So taking that and taking the measures of uncertainty in which we'd evaluate techniques, and focusing them on those specific cases drawn from the first statistical principles and defining that particular measure you want to evaluate, is something that we're focusing on. So we're trying to look at ways that we can tell a practitioner, *For practical guidance purposes, if you care about this type of problem, here are the type of metrics you should care about*. Some of the work we are doing is defining those metrics and trying to make sure that there are ways to visualize them and understand them better, but also, where is the state of the art relative to those particular lenses? It is an empirical evaluation. We're not going to make sweeping, general proofs about the state of the art, but we are able to show some empirical evidence under those refined cases where you may care about those use cases, about the state of the art in a number of techniques and how they perform.

Suzanne: I can imagine, just as an example, that in safety, airline safety, we have got certain use cases, as you say, that we are more worried about. Wind shear under clear conditions is one of those things that is really, really hard for humans to detect. So that is a case where, boy, if I had a model that said, *If you are seeing these measures coming in from your sensors, you are likely to have some clear wind shear coming up*. Is that an example of the kind of thing that you would be focusing your models on, so that you can get that better accuracy, better confidence, but within a smaller bound?

Eric: Yes, I think what you are referring to are shifts in the dataset. So you may train your training set to build the model, which maybe focused on a particular set of use cases for which actually apply it. Like you said, this particularly difficult case may be rare or whatever, but your model may perform poorly. In those particular cases, you want your uncertainty quantification to be good, because your model is likely to be wrong in those cases. We are doing a little bit of work focusing on cases, and the specific case, which we can talk about in detail is about something called covariate shift. Quantifying uncertainty in the presence of covariate shift, meaning the distribution over your instances has somehow fundamentally changed from training until test. That is the type of problem. Your instances are fundamentally changed in this case. That is one of the problems we are focusing on, and I think it has, to your point, many practical use cases within the DoD, for which they are going to deploy a lot of these AI systems, in environments that are going to be changing or have a lot of dynamic elements to them.

Suzanne: Or have a lot of noise in them.

Eric: Or have a lot of noise, right? So, having your model be able to detect when it is incurring enough change in the environment, whether it be noise or something else, when something is changed fundamentally but now it is uncertain, it is a really important practical problem.

Suzanne: There is a user-interface aspect to this I don't know if you guys have dealt with yet, that was just coming to me as we are talking. If I am a user and I am getting accustomed to using these models, and let's put the pilot situation—I have a whole lot of data coming in, and I use models like this, essentially, to give me alerts. That is one of the reasons that we use these models. But, if I have a bunch of models running and saying, *Oh, you have a 45 percent chance of having wind shear in the next 10 minutes. Now you have a 33 percent chance. Now you have a 65 percent chance*. Are you doing work in terms of understanding how the user interface and how—they call it *alert fatigue*—is one of the things. I can see these models really being a contributor to that alert fatigue if we are not careful about how we present the information from the models.

Eric: Yes, there is actually a lot to unpack there. In general, I think it is important to understand, and this is some of the guiding principles of the work, is that how people are using the models determines how you should evaluate them. To your point, I think there are a couple things.

One, the uncertainty in the model can be used to guide when you should provide alerts. If there is stuff, and this is low-risk, things you can automate, if the model is super-super confident in something, like maybe you don't bother someone with that. Or, if you are super confident that some event where you know you need human intervention, like the type of alerts you care about, those are the ones you show. If you have a budget for the human beings, that human operator, whatever.

Suzanne: The cognitive load.

Eric: Right, if there is some budget for the cognitive load under somebody, you want to prioritize the things you are most confident that they should be viewing. In some sense, if you have this distribution, there is so much data coming in for which you may want a human being to look at some of it just to assess what is going on. You want to stack it in some way that they are making the best use of time. In some sense, uncertainty quantification is going to help do that because that provides you some way of doing that. Now the other thing I think you mentioned that is important is the interaction between how the user interface and how the model behaves is really important. So, again, from a statistical perspective, calibration is one measure of saying, *Is your model outputting the correct probability distribution of your model's predictive probability?* In a strict sense, you are talking about the entire distribution. *How is it predicting over the entire set of class labels it can predict over*? But, if you are only going to show somebody three of the top predictions because the interface only allows for that, maybe you want to evaluate that particular case. And so, understanding those probabilities more is something that is guided by things like interfaces or how people are going to interact within the ML system, that should be brought down to the evaluation stage of these methods. Part of the work we are doing is doing that. We are trying to look at how people are using some of these models and then really

trying to go, *OK, so what are the metrics and techniques we should be using for those particular cases of interface and other things?*

Suzanne: We often have people from our customer community that are listening to these. I can imagine some of the people that I work with being interested in taking some of the models that they have built and having some evaluation of this, if they haven't done it already. Is that something that they can contact you for, contact the SEI for? Are you looking for interesting models?

Eric: Oh yes, absolutely.

Suzanne: I kind of thought you might.

Eric: I will say the thing that we are doing right now that is of most interest is we are doing that evaluation, like I said. But the code base that we are building up is supposed to be somewhat modular in a way that if someone came in with a model, we could swap it in and do an extensive evaluation. When we developed these metrics and ways of assessing these models, we can, hopefully, be able to quickly spin something up. Given some dataset you care about, under some certain conditions, and a model you care about, maybe we could potentially swap this in and do this type of evaluation.

The hope is that when we are done with this work, we are going to have this solid code base for it, and we are going to talk about some people throughout the DoD and the intelligence community about their models and what they care about, about picking some of the stuff they care about, putting it in this framework, and doing the type of evaluations and giving them that insight into their models, about whether they are expressing uncertainty in a way that is acceptable for their tasks. In short, we have code, and we are not afraid to use it for these types of things.

Suzanne: One of the interesting areas—I am just curious if you have worked in this area—is the confluence of security and safety. There are certain elements, making something safe, that are very common with making something secure. Then there are things that are very separate in those areas. Do you do that kind of aggregation? Can you look at models from two different perspectives and help people understand, *Yes, this model gives you some prediction of the security and the vulnerabilities, but it really isn't helping you with the elements related to safety that are inherent in this kind of model*. Is that something that people might think about?

Eric: I think this is a very important topic. My intuition of what I understand what you are talking about relative to the machine-learning research is that there is this whole swath of research in the adversarial machine-learning realm. They talk about things like, I*f you are able to*

be robust against certain attacks, you also maybe reveal something about your model. So there is a tradeoff there.

Suzanne: OK.

Eric: I think that largely what is hidden in the uncertainty-quantification task is that to achieve some of this quantification of uncertainty, you also need to achieve a level of robustness. That means you are going to be somewhat robust to the type of security and privacy concerns but maybe not. For this particular project, we are focusing largely on the uncertainty-quantification bit, but I know that there is certainly work in the SEI and within the academic community that is more focused on those types of topics. And it is certainly an important thing to think about.

Suzanne: You may be hearing from them when you get a little farther along with this research.

Eric: Yes.

Suzanne: We talked about what kinds of effects some of this quantification of uncertainty can have in national security. Are there any areas I haven't mentioned that you would want to bring up in terms of positive effects that understanding your level of uncertainty could provide?

Eric: Yes, we talked a little bit about some of these. The ISR [intelligence, surveillance, reconnaissance] community obviously, there are certain situations where intelligence analysts, where you care about, and I know it is baked into a lot of ISR task work, and analysts have to provide confidence in the things they infer. Certainly, that is something that community is very comfortable with as far as I know. It seems very important for their pipeline. Now there are other parts within the DoD community that I think are going to be interested in this. One of them is the cyber community. Every time I talk to someone who is in cyber, cybersecurity, just the cyber domain, they will tell you that the data comes hard and fast. You are getting so much data. In those particular instances, you have to make decisions very quickly. Then, human budget is really at a premium there. We talked a little bit about the case where uncertainty can be used to curate, if people have only so much budget, it can be used to curate what they see. So maybe to automate a large portion of the data flow that is coming to somebody, and then overall, your processing throughput is significant, because you are able to automate some of that while only showing the hard cases to potentially that human being. So that is one. I think the other one that is obvious is uncertainty quantification is really important in things for reinforcement.

We know a couple people within the DoD. They care about reinforcement learning for things like robotic navigation. That is one of the things. And so, exploration of reinforcement learning, understanding where you are and what state you are in, and where to explore versus exploit is based on this concept of model uncertainty. That is something that is really important, and being able to quantify uncertainty in those cases is vital to a lot of those navigation tasks for robotic

platforms. Those are the ones that come immediately to my mind, but I think that the more I talk to people the more people go, *Oh, this is what you should be focusing on,* and so I think, obviously…

Suzanne: Everybody wants to help.

Eric: That's right, yeah, and I'm happy to get that type of feedback. But, I think that the point I take away is that this actually is pervasive in a lot of different applications.

Suzanne: So for our viewers that want to learn more about this, more about what you're doing specifically or just the machine-learning and uncertainty realm in general, are there any SEI-specific resources or other resources that you would recommend, in terms of a getting-started kit, if you will, in this area.

Eric: Yes, this project, at least on the SEI side, has probably been going for just over six months, so we probably don't have a lot out publicly yet. Like I said, we're building code and running experiments. Internally, we have tons of results that we're hopefully going to show in the next couple of months. There is almost certainly a project page associated with this on our website, but in terms of publication, look out in the next upcoming months from us, our group.

Suzanne: Maybe the Research Review.

Eric: Yes, that is a perfect example. I am certainly going to present something for that, and we have plenty of results we're really excited to show, but in terms of outside the SEI though, I think the first place to get your feet wet in this is a grounding in statistical machine-learning literature. As boring as it sounds, a lot of the foundations for the type of things we are talking about in this project are based in statistical machine learning. If you are familiar with that, ignore this whole thing, this whole rant here about this, but grab a textbook on statistical machine learning. They have tons to say about this. Now, if you are already there, the next step into it is to look at the literature from the machine-learning community. I am going to forget some people here, but Yarin Gal's group at Oxford does really good Bayesian neural network and other types of quantifying-uncertainty-within-models work. That is something that I think is really important and something that people can look at. Google has a couple groups that [do] this type of thing. Off the top of my head. Dustin Tran's group did a paper that was really great, and Jasper Snoek's group also has worked on this. And then, we really have this paper that came out at AISTATS in 2019. It is published by a group. I am blanking on where they are from. Uppsala University. It's a paper that they do this type of...they derive these metrics from statistical measures that are really, really useful. I think if you have a stats background, it is a really nice way to get your mind around some of these calibration and uncertainty-quantification techniques. That is where I

would go, but hopefully, if someone watches this in six months from now from when we are recording it, hopefully, you can just Google our group and maybe find some work from us.

Suzanne: Fair enough. No no, that is good. So you are an example of a podcast where we are kind of on the leading edge of the project. Other projects that have been around for a while have lots of resources. But there are different people that like to hear about both. You are in the right space. What is next for you? What are you working on that you may have something to tell us about in a few months, just to give people a little tease?

Eric: Yes. Like I have been mentioning, we have this calibration evaluation that we are doing. Again, the whole point is to provide some empirical evidence to practitioners about how to understand the landscape of uncertainty quantification and specifically, calibration of models. We are putting the finishing touches on that, and the uniqueness of that work is that we are trying to pose it as, *Here are some principled measures and empirical results*, and based on those measures, posing this as, *If you care about these types of uncertainties, here's how you evaluate it. Here is how the current state of the art works on a number of classification problems.* Hopefully, this is the next first step in a long line of work that we are going to be doing to ground the state of the art in some of this particular lens of viewing uncertainty. The idea is that that is going to be a jumping-off point for both speaking to people within the DoD about, *What problems do you care about uncertainty and how do we solve them with different types of techniques?* But also, I think it is important, in doing that work, we have found a number of research gaps that are really important. Specifically, we talked about covariate shift and dataset shift with their... That is one that I think is really important. But there are also examples, like you said, out of distribution detection where you have classes you haven't seen before that appear during developments or during testing.

Suzanne: During operation.

Eric: Yes, and so what do you do there? How do you reason about things in that regime? There are a lot of fundamental…

Suzanne: How do you account for a pandemic descending upon us? [laughing] Do you have a model that predicted that one.

Eric: That is right. Yes. We have some collaborators actually, or at least I know some people within the biology or the bioinformatics community that covariate shift is…I mean, so COVID is a very good example dataset shift, where suddenly, your clinical regime like you had before COVID and during COVID is entirely different. Your models need to update. So, that is a good example, but there are examples in the DoD context as well. But yes, that is the type of work, a

number of research gaps we have identified that I think coming up over the next couple of months, we are going to work on.

I also need to mention we have two academic collaborators from the [Carnegie Mellon] Machine Learning Department, Zach Lipton and Aarti Singh, and they are all doing some awesome work. It is starting to come out now, some of the work they are doing. So certainly look out for, again, Zach Lipton and Aarti Singh. I don't know if I said that too fast. But yeah, so they are going to do some awesome work, and so look out for some of the work they are going to be putting out.

Suzanne: Very good. All right, I want to thank you very much for joining us today, Eric, and this work is fascinating to me. I am not a statistical thinker natively, but have been kind of getting into that groove a little bit with some of the work that I have been doing. So I really appreciate this aspect to it, us bringing this to everyone. Because I think that is actually one of the things that is part of this whole machine-learning space is that people that weren't really thinking probabilistically before, there is going to be room, especially, maybe not as much in our personal lives, but certainly in our professional lives, I think we are going to be touching a lot of different things with this. Some of the things that you are doing now are going to make it easier later for people to be able to adopt modeling and adopt machine learning as an OK way for us to engage in some of our professional activities. I want to put that push in for everybody to look at your statistics. It's a different way of thinking that we need to be more aware of.

Thank you again for joining us. The resources that you spoke about today, we will include in our transcript for the podcast, so our viewers will be able to get access to all of those. I want to thank all of our viewers for listening today, and have a wonderful afternoon. Thank you.

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 do not hesitate to email us at info@sei.cmu.edu. Thank you.