Explainable AI Explained

featuring Violet Turri as Interviewed by Rachel Dzombak

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.

Rachel Dzombak: Hi, everyone, and welcome to the SEI Podcast Series. My name is Rachel Dzombak, and I am head of digital transformation in the SEI's AI Division. Today, I am so excited to be joined by my colleague, Violet Turri to discuss the current state of explainable artificial intelligence [AI], as well as the challenges and limitations associated with explainable AI. Violet recently wrote a blog post outlining the strengths and practices of explainable AI, also known as XAI—we will use those terms interchangeably today—and she described XAI as a powerful tool to help develop user trust in AI systems and address rising ethical and legal concerns. I think we will have a lot to talk about today, and I am very excited to welcome Violet to the podcast.

Violet, thank you so much for being here today.

Violet Turri: Thanks, Rachel. I am looking forward to our discussion.

Rachel: Awesome. Let's start off. Let's jump right in, and have you tell us a little bit about yourself and the work you do here at SEI.

Violet: Sure. I am a software developer and researcher in the AI Division at the Software Engineering Institute. My background is in computer science, human-computer interaction, and also creative writing. Thus far, working at the SEI, I have worked on a handful of machine learning projects in more of a developer role. Throughout that process, I started thinking a lot about how we can build better machine learning systems, systems that are robust, secure, human-centered, and also scalable. And, especially, human-centered has been really interesting to me because there are so many issues related to trust, transparency, and then also human-machine teaming as well, that are obstacles to adoption of AI systems. I have gotten really interested in explainability and understanding how we can build systems that people want to work with.

Rachel: Great. Thanks so much for that introduction. I think the word *explainability*, we have already used it a few times here. It is popping up more and more these days across industry, across government, on Twitter, across all social media. But I feel like there is not necessarily a clear definition of what explainability looks like. Could you start off by sharing your perspective on what is explainability in the context of AI systems?

Violet: That is a great question. There really isn't a consensus on the definition of explainable AI or other related terms like interpretability or explainability. I see explainable AI as encompassing all of the techniques that make the decision-making processes of AI systems understandable to humans. This includes both the underlying statistical techniques and also the interfaces. There is a huge range of different kinds of interfaces. Explanations can look like heat maps or a neural net visualization, scatterplots. I have also seen people refer to explainable AI as really being those statistical methods that I mentioned, so, LIME, SHAP, Grad-CAM.

Another area that is open for debate is how we use interpretability versus explainability. Some take these terms to mean either a system that's transparent by design, in which case it's interpretable, or a system that is opaque by design, in which case explainability can be introduced. So, some associate interpretability with transparent systems versus explainability with opaque systems. Those are some of the distinctions and dialogues around current definitions.

Rachel: Gotcha. So sounds like not a whole lot of consensus and I imagine that the implementation of explainable AI looks really different as a result of that as well. Maybe that is a place to go next, if you could talk a little bit about what do XAI processes and methods entail? What does implementing them even begin to look like?

Violet: Yes, explainability is really about probing a model. That is what all of these techniques sort of boil down to. So you can look at specific examples and investigate how inputs result in outputs that are decisions, and you can examine what the relative feature importance was. So, which characteristics of the input were most important towards reaching that output? You can look at really specific instances. You can examine edge cases even as you are developing a machine learning system. You can also aggregate examples across a whole class to get a sense of what is important for instances across a particular category. That is the general feel for how explainability works. Currently, a lot of explanations that have been implemented are designed to support machine learning practitioners as they develop these systems and they tend to look highly technical.

Rachel: Gotcha. You started off that answer by saying that right now a big focus area is in probing the model. Could you double click on that? Say a little bit more about who would want to be probing a model? What types of questions would they be asking? What does that look like?

Violet: I think that this question takes a different form before deployment versus when a system is in deployment. Before a system is deployed, obviously you have your team that is working to create a model or system that meets specifications. For them, probing the model is all about making sure that it has learned the behavior that is the intended behavior of the system. A lot of the time, in typical sort of machine learning development practices, people select a handful of metrics, and then they just kind of track how well the model performs throughout training, and they keep working until they've hit some milestone. Then they say, *OK, the system's ready to go, we can use it*.

However, explainability helps you take that a step further and actually try to explore what the model has actually learned. A lot of the time the metrics are more of a proxy for a goal or behavior that you want to learn. Just because you have performed highly on a metric doesn't mean that your model has actually learned the intended behavior. It could have learned a shortcut of some kind. So for people who are employing explainability in this before-deployment kind of context, probing the model is all about vetting the system, making sure it behaves as intended. This can include looking for problems, such as bias or other unintended behaviors.

Then after deployment, probing the model is more about, or it can be more about, interactions with these end-users. So, for example, if you were in a human-machine teaming context and you are working with a system, and it recommends some course of action, you are going to want to understand potentially why that decision was made or a suggestion was made. You would require the same kind of information as you would of a human teammate; an understanding also of where the system does well and where the system fails, similar to how you would with a teammate as well.

Rachel: Yes, so definitely it looks different across these. What I am hearing you saying at least is that it looks different across these different stages, but that there are a lot of questions that practitioners need to be asking right now. So explainability can help with that. I think another area that at least I am hearing so many people focus on is about the need to develop trust in AI systems. That I imagine is another loaded term which people interpret very differently. But in our opening we talked about the provocation you gave about the connection between explainable AI and trust. I am wondering if you could share more on what does that connection entail and talk a little bit about how it manifests.

Violet: Trust is definitely a huge topic right now, similar to explainability. In speaking with people in industry and also in the government, there is this sense that trust is a really massive challenge, and that solutions to building trust are going to come from a lot of different sources. Some people feel as though they can trust a system after they understand how it makes decisions. Others need to have a positive track record of working with a system. So explainability can definitely help encourage trust, but it is just one component of this larger issue. I do think that

explainable AI is an important tool towards building trust. If you have people who are developing the system make use of explainability, they can make sure that the system is a high-quality system, and that it works as intended. So you have trust from the team that developed the system. You can also encourage trust on the part of end-users, because if they are interacting with the system and they feel like they can understand it and perhaps team with the system as they hope, that is another source of trust. So definitely explainable AI can help develop trust on a lot of different stages of development and among a variety of different groups, but it is only one part of sort of a larger challenge.

Rachel: I love that. I feel like a conversation we frequently have is how explainability is necessary, but it's not sufficient. But this trust challenge is a lot bigger, but also very multi-faceted based on how people interpret it.

Violet: Yes.

Rachel: Given the connection here between trust and explainability, we certainly have witnessed a rising demand for XAI within government sectors. The U.S. Department of Health and Human Services talks about an effort to promote ethical and trustworthy AI use in development, including explainable AI as one of the focus areas of their AI strategy. Certainly, in our work with the Department of Defense that is a major focal area as well. So I am wondering if you could talk a bit about how our work in the SEI AI Division helps the defense and intelligence communities to address this rising demand for XAI.

Violet: Yes. I would say that the demand more so than being for explainable AI is really about developing transparency and oversight in systems. This is so important in defense and intelligence communities because these systems are so critical, and failure can be really costly. We have a lot of different work that we have done, touching on a lot of different aspects of AI engineering that relate to transparency. A couple of projects that come to mind, one would be Carol Smith and Alex Van Deusen's project with Defense Innovation Unit. That is all about operationalizing the responsible AI principles. Another project is Eric Heim's work in uncertainty. Uncertainty is really a measurement of how confident the system is or unconfident the system is in its output. That is another measurement that helps people understand how to interact with a system and how much to trust it.

Related to explainability specifically, a couple projects also come to mind. One is earlier in the year, I worked on putting together a corpus of interactive explainable AI examples, which we hope to release in a future blog post. Then we are also working towards an upcoming project together that is looking at how to craft explanations for transition stakeholders. There is definitely a lot of work related to responsible AI, transparency, and explainability that bridge different areas of AI research that the SEI is involved in.

Rachel: Yes. I will add to that by saying I think what we are trying to do is within the AI Division a big focus of ours is doing AI as well as it can be done. Trying to figure out what that means requires us to take this holistic approach to understanding all of the different component parts, but especially the focus on what will it take for people to actually adopt these systems.

The AI division also has a growing area of focus on AI engineering, which is all about, how do we build a discipline around the growth and development of AI systems. What does the discipline need to entail? What are the set of processes, tools, and frameworks that can guide developers as they think about creating these systems? Not just from the craft phase of hacking it together, but to be well-engineered systems that are human-centered, designed to work with and for human needs, robust and secure, able to operate reliably in the face of uncertainty or threat, and scalable, which could be scaling up across an organization or scaling down into a specific context. I think that is one more area that is worth exploring is in what ways do you see this work on explainable AI intersecting with the pillars of AI engineering right now?

Violet: I think there is a really clear link to the human-centered pillar. You can use explainability to help develop trust by making the underlying decision processes of the machine learning models more clear to humans. There are instances where humans require explanation to know that they can trust an outcome, and they require explanation to develop trust in the system in general. I think that that link is really palpable.

As far as scalable AI goes, a lot of people think of scalable AI as meaning that you are scaling up or you are scaling out. But scalable AI can also be about scaling in and understanding what decisions need to be made in context. So, who are the stakeholders who are involved, what kinds of decisions are they making? What information do they need to make good decisions? And to also trust system outputs? These are kind of similar questions to the ones that we encounter with the human-centered aspects of explainable AI.

Then as far as the space of robust and secure AI goes, explainability really helps you understand what your model has actually learned. So if there are some weird kinds of correlations that your model has learned that are unanticipated, that will make your system be more robust; or only work in certain contexts; or if there's some kind of underlying relationships in the training data set that would pose a security risk if they were to be found out. These are all things that explainability can help you identify. It is definitely an important tool towards building robust and secure systems not only human-centered.

Rachel: I think that piece you just mentioned about the spurious correlations that could exist, and how explainability can counter that, or can at least make it transparent about what the decision is, is super important. Could you give an example of what one of those kind of odd correlations could look like? How does that show up in a machine learning system?

Violet: Yes, one example that I think a lot of people love is this one with cows and pastures. Basically researchers developed a machine learning system that was supposed to classify animals. They saw that it had really high performance for cows. So naturally they thought that it was good at identifying cows. What they realized is that what the system had actually learned was how to identify this kind of pasture setting that cows typically, within the training and testing data sets, were located within. So if they gave, for example, a picture of a cow in snow, it could no longer properly classify the animal. So you can imagine more kind of severe or high-risk sorts of correlations that either would render a system unable to work in new environments or would pose a threat to adversarial attack.

Rachel: Definitely, and my understanding is right now that many explainable AI methods, especially the statistical ones, are focusing on what are the features of interest and highlighting what are those main features that are being used in that decision. First, correct me if I am wrong, but second, how would that play out in the cow example? What would the explainable AI be identifying in those examples?

Violet: Yes, for examples that are involving computer vision systems, a popular technique is to develop a heat map that shows you which pixels within the scene were most important at reaching the decision. In this cow and pastures example, researchers or people probing the model could tell that it had learned to identify the background if they employed this technique because they would see that the cow is one of the least important features in the heat map, whereas the background is really critical.

Rachel: Gotcha. Cool. I think that brings alive how explainability can certainly help make some of these decision processes clearer or at least more transparent in how the decisions are being made by systems, which can help people make a judgment call on when and how they should trust and use AI systems. That leads into my next question, which is that despite this growing demand for AI in fields such as medicine and finance, adoption is still slow. I think for every day there is a hype article about AI, there is also an article saying, *Well, be cautious*, and *People aren't actually using the system yet*. What do you think is stopping the adoption of AI in these fields and others, and how could explainability help to increase adoption of AI systems?

Violet: To answer the first part of that question, *What is stopping adoption within these fields*? I think that in areas like medicine and finance, people are experts on existing processes, and they need systems that will work with their workflow. In these fields, it is also really important that the people who are affected by decisions and the people who are making decisions understand how these conclusions are drawn. They need to be able to make sure that they align with how they have been taught to handle the process, or what they know to be best practices. So, I think that AI poses a challenge, because sometimes, if you are working with opaque models, it is really

hard to figure out why exactly you have drawn a particular conclusion. So this is an obstacle, a major obstacle to adoption.

Also, on top of that, there are legal and ethical concerns related to accountability if you have people making decisions that are informed by AI systems. Sometimes these systems make decisions that cause things to go awry. It's really important that you have transparency in these domains, even when working with AI systems that seem able to handle everything on their own.

I think explainability could really help combat a lot of these issues. Currently, a lot of explainable AI work, as I mentioned before, is really about supporting machine learning practitioners as they develop. So these explanations are really technical. But I think moving forward with explainable AI, there is an increased effort to try to frame explanations for non-technical or non-AI experts, and I think that this work will really help with adoption.

But one other kind of caveat is just that I have seen a lot of discussion around if you are using an AI system to make an important decision, it is really critical that the system is transparent, and perhaps transparent by design. I have seen arguments that we should be working with interpretable models where we don't have to make these kinds of educated guesses through explainability about what the decision-making process is, but rather we should rely on something that we can definitively say is how the model works.

Rachel: That makes a lot of sense. Especially you use the word *important* problems, but I really think about it as high-stakes problems. When the stakes are low, if a system gives a recommendation and it is incorrect, then, if you are recommended the wrong song or your lights go out too quickly, you might be annoyed, but it's not going to be super consequential. But in fields like medicine, like finance, certainly in government, the consequences can be incredibly high. So the more we can translate how systems are working, the better.

One thing that we like to address in the podcast is transition. How do we get people to adopt, to pay attention to, to take on all of the important research that we are doing? If I am leading an effort to adopt XAI within my organization, where should I start? What resources would you recommend that I try to get access to ASAP?

Violet: I think that currently in the area of explainable AI research, there is a lot of discussion, but there are a lot of open questions about how to put these principles and techniques into practice. I am seeing more and more case studies emerge. I know that in areas like finance a lot of organizations are spinning up explainable AI hubs for experimenting with explainability in their systems. I would say it's still a bit challenging to figure out how to operationalize explainability.

From an organizational perspective, some things that I think it would be helpful to start thinking about would be reflecting on the resources that it would take to integrate explainable AI into your systems and also considering what kinds of roles that you would need to accomplish the goal. You are going to need to understand your stakeholder or stakeholders and what their needs are and figure out the best way to present these explanations. This will likely require people with more of a human-computer interaction or design background or roles that really work on facilitating communication between parties as opposed to just highly computer-science focused kind of developer roles.

Another kind of caveat related to how this looks organizationally is that a lot of people try to bolt on explainability at the end of their projects. It is really important to start thinking about the principles of responsible AI from the moment that you start developing a project because this could really shift which methods you use to attain outcomes. So in the instance of automated decision-making systems, for example, if you know that transparency or full transparency is going to be a requirement along the line somewhere, you may make the decision to work with interpretable architectures as opposed to ones that are opaque.

Also, as far as like what to read, how to start learning more about how you can implement explanations, I would suggest starting with a landscape analysis and just trying to find examples of where explainability has been integrated into projects successfully, like in the kinds of case studies that I mentioned. There are also some frameworks for practical implementation that are starting to be published, and I think taking a look at these would be another great step. Additionally, there are a lot of open-source tools out there that claim to help with explainability. I would say that it is going to likely be difficult to just take one of these tools off the shelf and start using it and have it handle all of your needs because explainability is so context and audience specific, but that could be a really good starting point.

Rachel: Excellent. I really love how you outline both what resources are available and how to navigate them cautiously, recognizing you can't just Google explainable AI and think, *Oh, that's going to be problem solved*. You have got to pull together a team. You have got to think about this from a systems perspective so that it can be implemented from the start of projects, not just bolted on afterwards.

Violet, thank you so much for being here to talk with us about explainable AI. I know I learned a lot, and I work with you all the time. So always great to be in conversation. And to our listeners, thank you for joining us today. We will include links in our transcript to all of the resources mentioned in this podcast. And as always, if you have any questions, please don't hesitate to e-mail us at info@sei.cmu.edu. 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 don't hesitate to email us at info@sei.cmu.edu. Thank you.