A Dive into Deepfakes

featuring Shannon Gallagher as Interviewed by Dominic Ross

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.

Dominic Ross: Welcome to the SEI Podcast Series. I am Dominic Ross, instructional technology engineer for the SEI CERT Division as well as the SEI broadcast media team lead, and I am joined today with Shannon Gallagher, a data scientist for the SEI CERT Division. Today, we are talking about deepfakes.

Shannon, there were more than 85,000 harmful deepfake videos detected up to December 2020 with that number doubling every six months since observations began in December 2018. Shannon, we know deepfakes are definitely a problem for many organizations. It is such a good thing that we have you here today to walk us through and provide some clarity. But Shannon, since we are both new to the SEI Podcast Series, can we start by telling our audience a little bit about ourselves, what brought us to the SEI, and the work that we do here? Can you start us off?

Shannon Gallagher: Sure, and first of all, thank you Dom so much. It is great to be here today. As a little bit of an introduction, I got my PhD in statistics at Carnegie Mellon University from the department of statistics and data science. My thesis was about modeling infectious disease, and that was before COVID. So, it was kind of I guess timely. Then I did a post-doc at the National Institute of Allergy and Infectious Diseases [NIAID], and then came back here. I've been working at the SEI since June 2021 and have been working with some data-science machine-learning problems ever since. From a day-to-day aspect, there is a fun little quote from statistics that *We get to play in everyone's backyard*. I feel that is so true, it is really quite fun. Whoever has a spreadsheet, a set of data, they can come to the statisticians, the data scientists, machine learning and if you have specific hypotheses that you have in mind.

Or occasionally, you are like, *I have this dataset, can you help us explore it?* Those are some of the best. We do a lot of modeling of algorithms. This is one of the deepfake things we are doing today. That is where a lot of the statistics and machine-learning models come in. Really, I think one of my favorite parts of working here at the SEI is getting to work with so many different disciplines. That is the best part of being a statistician as well. For instance, I get to work with this really cool team media lead on this deepfake project.

Dominic: Well, I am sufficiently humbled. I took a different route to the SEI. I started as a multimedia artist. I primarily worked in broadcast and film and television. From there, moved to private production houses, moving over to corporate video and live-streaming productions. From there, I actually got hired at the SEI to be a digital media production specialist, which I had no idea what it meant at the time. I just knew that I would be creating distance-education offerings. From there, here at the SEI, they really recognize individual talent, and they allow you to pursue other fields outside of what you may have been hired for. So I was able to get into more engineering, design, and software research. I have worked on a number of research projects here at the SEI, specifically ones that are tailored more towards multimedia of course because that is where my interests lie. But getting to work with somebody as accomplished as you on creating a deepfake-detection model is quite thrilling. I really appreciate the opportunity to be just in your presence in this room but also to be able to work with you in research. Let's go ahead and get started. YouTube deepfakes are hilarious. Why would a data scientist study them instead of enjoy them?

Shannon: Well, part of studying them usually uses a holistic view so we do get to enjoy them and watch them. I agree, they are quite hilarious, and that makes the job pretty fun getting to watch these new YouTube videos. I am certainly kind of lucky to be able to do that, but there are a lot of reasons to study these statistically. First of all, they pose a real threat, they can pose a real threat to our society. Being able to impersonate someone's image and likeness, being able to not be able to trust what you see with your eyes and hear with your ears. That is something that we would like to be able to detect with the help of some machine-learning algorithms. From a statistical point of view, there is really interesting sorts of data. It's spatiotemporal. You need both the video. You have the image which is all very highly correlated with one another and the time as well. It makes it statistically a very interesting problem to work on.

Dominic: Spatial-temporal. I can't even say the word, it's so impressive.

Shannon: Well, thanks. But just time and space.

Dominic: What are deepfakes? How would you characterize the current state and technology that is being used to create them?

Shannon: Sure. Deepfakes are defined in the Mirsky and Lee paper as believable media that are generated by neural networks. This definition has evolved over the years, and depending on what field you are from, maybe you have a different idea of what deepfakes were before. Usually, when we are talking about deepfakes, especially today, we are going to be talking about humans and especially faces. We are going to focus a lot on the facial features of an individual. But it doesn't necessarily have to be a video or an image. It could be audio or text as well. So the definition of deepfakes seems to be broadening over time. In some ways, it is just an evolution of trying to fake someone out.

Dominic: So where does the neural network come into play?

Shannon: The neural net comes into play as the basis of which these images are created. Prior, you wouldn't be using like a statistical model to alter a media or image. But with the machine-learning algorithms, it really allows you to have this very firm basis of this machine, excuse me, this interpolation of an image usually on top a target source or person, so you can swap faces or swap identities and be able to really imitate a person that you want to.

Dominic: OK. Does a deepfake have to be created through using that neural network, or could it be created through other means like through a graphics-altering program like Photoshop?

Shannon: Sure, it depends on who you ask. Since I am in the statistics, machine learning, sometimes we have the bad habit of coming into a field that is already established and rewriting and saying, *Oh, yeah, we did this first.* I think you are very right. Altered-media came first and is really the basis of a lot of, with these deepfakes, and the deepfakes, a lot of it comes from being able to automate the process.

Dominic: Why did they become deepfakes anyhow?

Shannon: As a lot of interesting things from today come from the internet, *deepfakes* is one of them. This is a term that has been coined by a Reddit user. It is sort of a play on words between the *deep* from the *deep learning*, the neural networks [DNNs] come from that. Then a *fake*. There are multiple ways we could say *fakes*. But, put it together, it's kind of catchy, caught on, and here we are, *deepfakes*.

Dominic: Internet glory established, got it. Can you tell me more about the type of deepfakes present currently and what makes them different?

Shannon: Yes. So there are quite a few different types of deepfakes. Perhaps the most simple one that is available is the face swap. Perhaps you have been able to use that on some of the apps on your phone that you can just switch with a friend occasionally. It doesn't look particularly real because you are just kind of chopping off a little portion of your face and basically slapping

it on another person. There are some more sophisticated ones where you can swap like the whole head. You can even get a little more sophisticated with completely being able to puppeteer someone. There are some famous deepfakes on the internet available that do this where there is actually an impersonator who is being filmed. Then the face and likeness of a person is being placed on top of that person. So you are completely controlling that person's movements and expressions. That one is pretty dangerous. There is also just the fact that you can make completely new individuals with deepfakes.

Dominic: Is it easy to do this? Can anybody just pull out their computer and just start popping out deepfakes whenever they want?

Shannon: Yes and no. In some ways, if you just search on the internet for deepfake applications, you will be able to find them. A lot of them are very accessible. You just click on somewhere, and you can do some fun things. But on the other hand, I wouldn't say they are the most realistic. You would probably not be able to send those types around to your friends and family, and they probably wouldn't believe that the person you are trying to impersonate is actually speaking. So on that hand, no. But that said, there are some very sophisticated deepfakes out there. They take a lot of skill. I think it's really interesting because it's not just technical skills. It's a lot of artistry as well. I really think right now, you need both to make a deepfake that is pretty convincing.

Dominic: I would agree with you. I do think it is that mix between technology and artistry that makes the convincing deepfake of today. But I believe technology is quickly outpacing the need for that. Relatively soon, I believe the technology will be able to create photorealistic deepfakes on its own. But it's job security for me for at least another couple of years.

Shannon: It is kind of interesting how much deepfakes have evolved even in the past couple of years. They are getting pretty scary good right now, and a lot of that is from the technology.

Dominic: Yes. but here's the thing. I saw a deepfake of a leader asking civilians to surrender. It was so easy not to believe that. If I can easily spot those deepfakes, how easy are they to detect in the real world? And do we need a machine to actually do that detection?

Shannon: I think one of the reasons that one—I know what you are talking about—it was pretty easy to detect was because in some ways, it was supposed to be very an official video or at least mimic an official video. When we are coming from that, we expect a sort of provenance, a sort of standard for that video to have. I think the deepfake version of it did not meet that whatsoever. Now, imagine like if you had more of the found footage type of film, and you saw the Yeti walking through the woods. Then I think that would actually be a lot harder to discern whether it's real or not, or maybe it was just terrible video quality. I have faked myself out a million times and just excused like, *Oh, this is just bad filming*, when really there is something deeper

going on. Besides that, the number two is that there is something like 100,000 hours of video being uploaded to the internet every day.

Dominic: Wow, that's scary.

Shannon: It's concerning. But obviously, we can't check all that. So, we need some help from our computer friends.

Dominic: Yes, it's kind of like the analogy that anybody can make a pizza, but try making 1,000 pizzas, and let's see how well you can make that pizza at that point. Anybody can right now spot the majority of deepfakes out there that aren't necessarily the high-quality, media-artist-created photorealistic type. But show me 1,000 of them? I may miss a high percentage where we need a machine that is better than the human at doing that.

Shannon: Or at least for the machines that just take the first pass at it. I think that is important too. We don't need the machine-learning algorithm to do everything. Sometimes it just needs to help guide us in the right direction and then gives it to a human to make the final decision.

Dominic: I look at deepfakes almost like catfishing, and catfishing has been around for a very long time. Do you consider deepfakes the evolution of impersonation?

Shannon: Yes, I think so in a lot of ways. I will even take it back one step and say that it is just another forgery as well, so trying to get you to trust something that can actually not be trusted. I think this just takes it another step where you are now fooling people not just with words or an image, but even now it is going to be video as well. We use our eyes and ears a lot, and if we can't trust them, then it definitely sows a lot of doubt and perhaps discord.

Dominic: Yes, I wonder if I can like start using them to fool grandma that, you know, I can't make her 80th birthday because I am being tasked on this great mission, not that I want to go to a Formula One race.

Shannon: I mean, I guess but I wouldn't disappoint grandma.

Dominic: When most people think about deepfakes, they think about the real humans that we are talking about. What about AI [artificial-intelligence]-generated humans who don't exist?

Shannon: I would say there is a large number of AI humans that don't exist on the internet right now. I can imagine that there is actually a lot of social media profiles that do not have real individuals associated with them. It is certainly a part. From a statistical point of view, I think those ones are kind of very exciting for a number of reasons. One thing that it really brings to mind is that to make these algorithms, your model needs a set of real human photos to what we call it train on. So these photos are used as input to the model to make it work. Something

interesting about these new individuals is that sometimes maybe they are not as new as you thought but really just combinations of people who are already present.

Dominic: So you are talking about like six degrees of Kevin Bacon here that we can trace those combinations of features back to their original source imagery?

Shannon: Yes, and I think that is a great idea. I think there is a lot of open questions in this field. So the next data-scientist machine learner who has a lot of time, I feel like go ahead, keep working on that. I am ready for that paper, and we will gladly read it. Yes, but I think there is some certainly very interesting things in how you can trace back the origin of a deepfake.

Dominic: Yes, I mean these people that do not exist, it's present even in media artistry where UX designing software applications have plugins that you can just pull these random individuals, so you don't have to worry about copyright or stock imagery when you are creating these mockups of design. So, I would agree with you, it's probably pretty prevalent. But what about the pets that don't exist.

Shannon: Those are pretty funny, and here is an interesting tidbit. Sometimes a detector that can tell really well whether a person has been deepfaked or not, you can run that same detector on the pets that don't exist, and it will fail spectacularly. So it kind of shows how some of the intricacies of our models aren't really getting at the truth, but rather other spurious artifacts that are going on within the images.

Dominic: I saw a recent article that centered around trustworthiness, and that AI-generated humans are rated higher than real people when it comes to being trustworthy. What do you think about that?

Shannon: Yes. So that was a really nice paper by Hany Farid and his lab. They have been doing really awesome work a long time before it was called *deepfakes*, altered media and such. Yes, one of their studies found that the people from I think it is a model called StyleGAN and perhaps some others like it that they are found to be more trustworthy by actual human beings. I think there is actually a good reason for that because the images, the photos that StyleGAN was trained on, they are very nice photos of individuals. These were not like candid photos. So they are very pretty people and, they are very attractive. That probably helps with the trust level as well.

Dominic: So it is not like taking photos of me getting out of bed in the morning at my worst.

Shannon: I think if we put some of our Zoom photos on some of these models, we would have some scary-looking people, Dom.

Dominic: But it can also be one of those potential ways to detect these type of images as well, right? If they have an extremely high trustworthiness score or maybe just lots of symmetry, traits that aren't necessarily something that we see in the everyday individual, could that be a way to start discerning, *This is an AI-generated human* versus *This is a real human*?

Shannon: Yeah, and there is some stuff in StyleGAN that you see the asymmetry that you notice. One thing that is pretty hard for it to pick up is earrings. Normally, we expect those to be about symmetric. They are often not in the deepfakes. So usually that is a pretty obvious indicator. The thing is, we expect these generators to learn from these simple mistakes eventually. But right now, we are very grateful when they do the earring that is a star on one side and like a little stud on the other.

Dominic: What role does media literacy start to play in deepfake detection?

Shannon: I think it is important for just about anything that you would do on the internet. Honestly, you want to see where your source is coming from, who made it, what kind of idea that they are trying to portray? Are they selling something? You really just need to think through all that. I guess you just can't automatically trust anything.

Dominic: When it comes to deepfake creation, do I own my own personal identity?

Shannon: So, Dom, you asked a good question, probably one that I am not really qualified to answer. We will have to let the lawyers battle it out. But I will say, as a researcher, that we want to err on the side of caution, especially when…We want to respect the people behind the deepfakes because there are real people behind these models and images that we are using. That is very important, the ethics. We want to be researchers of integrity.

Dominic: Yes, it gets really strange I can tell you even in the media world. I am not a lawyer so this may not be right, but I believe that the photographer owns the copyright to the image of the person they take the photo of as opposed to the person that is in that photo. So, at that point, could anybody who has recorded video of me or taken photographs of me then own the rights and to be able to manipulate that any way they want? I would love to know what laws govern the creation of deepfakes. Have you heard of anything for that?

Shannon: I have heard of some legislation that was perhaps going to be passed in the U.S. but hasn't yet. I have been trying to keep an eye out. But it seems right now a pretty gray area. We are going to see this worked out sooner or later. Like you said, it is quite interesting. Copyright law is a big driver of it actually. Another is that, at least in Pennsylvania you do have that right to privacy and also your reputation. I don't know what that would mean if someone had a deepfake perhaps of yourself saying something that you would never say even if they were the one who

had the copyright to it. I don't know. Plus, there is the part where you have like art and satire. That is another attribute as well.

Dominic: Yes, the free speech protected in satire would be a difficult challenge I think for people who are creating deepfakes what legislation overcomes. So it will be interesting when it comes. That is a conversation for a different day, though. We could probably do a whole podcast with a lawyer present, of course.

Shannon: Yes, many lawyers perhaps.

Dominic: So, you talked about ethics, and guidance has been created concerning the ethics of AI here at the SEI. Should deepfakes be part of that conversation?

Shannon: Absolutely. I think this goes not only for what we are working on now but like anything especially that starts to touch upon humans and using them in our research. It needs to be really important. We are messing with real lives potentially, and we need to treat that with the care and respect it deserves. I know here at the SEI, we talk about the ethics a lot. We only use data that has been consented to and is publicly available. These are issues that we need to take seriously because if we don't, it will seep into our models and into the world. That is something that I think we should need to really care about.

Dominic: We have said a few things that we like about deepfakes but mainly negative things though. Are there positive use cases for deepfakes?

Shannon: I think there are. You know, the YouTube ones, the internet ones, they do bring me a lot of enjoyment. I do think there is a place for art, especially in the world of deepfakes. I will say I work with a lot of the bad use cases of deepfakes. For me, it is harder and harder to see some of the positive aspects. But I do think that ultimately we need that artistic expression as well.

Dominic: Yes, I see those same artistic expression positivities like where you restore old video footage as black and white, and you make it color. You actually de-age an actor to be able to play a role that maybe it has taken a long time for that film to get made. You want to keep the continuity, or you want to have subtitles where it appears in a native language, so you are using deepfake audio so that I can speak Mandarin perfectly. Then also the manipulation ellipsis happening the way that somebody would see it, so it is just easier for people to understand that medium that is being conveyed. Whether or not it can continue to spiral out of control, I guess will see. But I have also experienced a lot of joy seeing my favorite celebrities act as magicians or, you know, running for president is…

Shannon: Literally running.

Dominic: That is right. It is a lot of entertainment value as well. Let's talk about the SEI's work in this area though. To quote the Heilmeier catechism, *If we are successful, what difference will it make?*

Shannon: In my mind, there is two big categories that we are trying to make a difference in. I just want to preface this that we are in the area of detection right now. We are trying to make those algorithms that are finding if something is a deepfake or not. There are two types of targets that you would think. One is a popular, well-known figure. If you can control what that person is saying or doing, obviously that can be a big problem. Of course, though, it is also you don't want the average citizen to be able to be impersonated easily as well. That one is kind of a threat on the small-scale sort of side. Someone is uploading a video on social media perhaps in school or something, and they are trying to sow discord and doubt. That is not great either.

Dominic: Well, what do the current trends suggest about the future of deepfakes then in the efforts to detect and combat them?

Shannon: It is kind of a tricky situation these days. Generators I think as we know on images, just still images of videos, they are basically impossible for humans to detect. I would say we are still pretty good with videos. That is something that we haven't gotten there, but they are getting a lot better. So generators are getting better. That means detectors have to keep getting better as well. It is kind of a weird game going on because the generators can directly feed off the detectors. We don't even have to give them the algorithm. They just take the results of our detector, and they can get better images from that. It is kind of this loop of detectors get better, generators getting better, so on, so forth, but I don't think it is an infinite loop. And I think it is one where eventually, the generator can make something that is indistinguishable both to our eyes and to the detector as well, which is kind of a scary thought.

Dominic: Yes, it almost seems like it is infinite, right? It is almost like it is going to keep going on forever where they just keep outpacing each other like we will never catch up. Do you think that's the case? Or do you think…

Shannon: I do think it is a tough game to play if you want to call it a game. I know it is also very serious, but it is definitely hard. The good news is that generators don't, right now anyway, they don't get better without a lot of time in between these rounds. That is what's kind of keeping us afloat at the moment. But in 5 to 10 years, I am not so sure. This is a difficult problem and may not be able to be solved through the content alone, and it may be solved through other means, such as metadata and certificates as well.

Dominic: Well, if you like this content, and you want to hear more about detection methods and you want to get Shannon back here for another podcast, go ahead and hit that *Like* button that

will let the SEI know that you liked this research or like hearing about it. Hopefully, we could get Shannon back here again to talk about some more tailored, specific detection, deepfake-related work. But as we talked about that, one of the aspects of our work that we like to highlight in our podcast is transition. How might our listeners learn about deepfakes and about the work that the SEI is doing on detection models and software frameworks?

Shannon: Well, for better or for worse, deepfakes are currently in the news, so it is pretty easy to find a lot of information. For work that specifically we are working on, I hope everyone has seen that Dom here and another person from our team have written a really nice blog post on deepfake generation. I recommend checking that out. We are also planning on having a Deepfake Day in August, so keep your eyes peeled for more information on that.

Dominic: Deepfake Day, sounds like fun. Do we know if it is going to be restricted research or is it going to be fundamental research or yet to be decided?

Shannon: Yes, so, hopefully, that will all come very soon. So keep an eye out for that on the website.

Dominic: Just reach out to the SEI, and we will keep presenting you with good information as soon as we have it for you.

Shannon: That is right.

Dominic: Shannon, you have said it all. Thank you for taking the time to talk with us today. For our audience, we will include the links in the transcripts to the resources that we mentioned on the podcast. The SEI Podcast Series is available on Apple Podcasts, Google Podcasts, SoundCloud, Stitcher, a lot of places you can find this, the SEI's YouTube channel. So if you like what you see and hear, please give us a thumbs up, reach out, let us know that we are doing a decent job. We like letting you know about the things that we are working on here.

Thank you again for spending time with us today, Shannon. Thank you, audience, for listening. Have an awesome rest of your day.

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.