Barclay Damon Live Presents *Cyber Sip™*
Episode 39: "Deepfakes and Double Takes: Can AI Take Us All Down?," With Siwei Lyu

Speakers: **Kevin Szczepanski,** Barclay Damon,
and **Siwei Lyu,** State University of New York at Buffalo

[Kevin Szczepanski]: Hey, everyone, this is a *Barclay Damon Live* broadcast of the *Cyber Sip*. Practical talk about cybersecurity. I'm your host, Kevin Szczepanski. Let's talk.

[Kevin]: Welcome back. We are here with Professor Siwei Lyu of the State University of New York at Buffalo. He is a professor in the Department of Computer Science and Engineering. Professor. You were on earlier to talk to us about some of the history and benefits of AI. And today he is back, and we will talk about some of its downsides. How can AI go wrong? Before we get to one of the chief ways in really one of your central areas of expertise, Siwei, deep fakes, I want to ask you a general question, which is: why do you think that in society today, in media today, there is such a disparity between those who think nothing of AI, understand that we're using it in our phones and other technology today. And on the other extreme, individuals who are literally talking about the existential threats of this technology, how we could lose our entire civilization. Why are these two poles sort of pointing and arguing with each other now?

[Siwei Lyu]: Well, I guess, you know, you are... this is like many complicated issues. You become more articulate when you push something to the extreme. I think that the reality is we're neither on the brink of a total collapse of society or, you know just as good old days, you know, nothing going to happen. It's somewhere there, there are some fundamental changes, but the change come in a way that is gradual and it is somehow, you know, subconsciously, in many cases. But I think, you know, people... many of the argument that come from that, you know, to make these points clearer, we need to give the extreme you know, push them to the extreme. Then we'll see where we are. So I kind of liked the discussion because, I mean, this is a what is sort of have, what is thought to be in a healthy, democratic society that we talk about the implications. I think neither...but on the other hand, you know, that if we base our decision on either of the extremes, it would not be useful. You know, we cannot totally ignore to see AI, but we can now be kind of so concerned that we don't want to use it at all again, as a combination is a trade-off in is somewhere in between.

[Kevin]: Yeah. Okay. Well thank you for that. And I think that's that makes perfect sense to me. Like, like every... we, we were having similar conversations, I think, about nuclear energy, which may have gotten off to a rather alarming start in the 20th century. But has evolved to be entirely useful. So let's switch to the chief subject that I wanted to get to today, and that is deep fakes. Tell us briefly about your background, because I think you are the leading expert in the field and then give us a definition when we talk about "deep fake," what is it?

[Siwei]: Okay. So a little bit about me. So I actually I was not born in the US. I was born in China. I came to the US in 2001 to pursue my PhD study and I was just, you know, fortunate to study with Professor Hany Farid, who is one of the... who is the pioneer and the leading scientist in this field called "media forensics." So when I started working with my advisor 22 years ago, to be exact, it was an infant research area because back then, you know, internet is a new thing. We just start to use email and Photoshop is a shiny new thing, you know, new kid on the block. And people just start to realize, oh, we can manipulate images with computers with such ease, using a tool like Photoshop. So many times, you know, the fact that we can manipulate media for.

At the time only digital images, it feels something like a like a like a new phenomenon. And people mostly feel about the curious you know, curious part of it then the potential harm. So Hany was one among one of the first several researchers who realized there are, you know, harmful side effects of this technology development. So I ...he gave a talk at the graduate fair, first year grad students meet with faculties and I was he needed fascinated with this idea because I never thought about you know devolving the algorithm and doing something really cool, really fun, you know, as if you are a programmer. It turns out that it can be something harmful and so I started working on that, and that's my PhD thesis. And now we're... since I become a faculty member and focusing in that area and my research actually cover a whole wide range of roles. I study, you know, computer science, developing algorithms, AI and machine learning. But I also have a strong mathematical background. So all this I would like to put them in the mathematical models to study them and turn them into computer code to run those models and run those algorithms. I also I was also fortunate to train under a neuroscientist when I was a post doc. So I also understand the side of human reason, you know, and that's very helpful because it gave me the perspective of why, you know, we fall for certain media manipulations. And that's because, you know, our biological brain was not aware that way to recognize those manipulations. And then later I started work with social psychologist and, you know, people working on education. I sort of understand the social aspects more of the social aspects of manipulated media. So my work has been mostly, you know, go under... behind the scene because, you know, media forensics was never a big topic to the public or even to the research community until in the middle of 2012, around 2015, 2016, when deep fakes suddenly become such a phenomena, you know, AI algorithms... deep fakes puts more simply are just falsified or manipulate media. Here we're talking about images, videos, voices, and also attacks. And the combination of then there are created with the help or directly using computer algorithms, and particularly artificial intelligence algorithms. So that suddenly put our research area into the front row for all people all want to know can we tell what is real and what is fake using all this because, you know, they become so realistic.

[Kevin]: So, Siwei, can you give us a good example of what a deep fake is? What do they look like?

[Siwei]: Well, deep fakes have many different forms. For instance, there are images created by algorithms that belong to nobody. So you see faces of ...AI-generated the faces, but also know.

[Kevin]: It's not me, it's not you. It's just fake.

[Siwei]: It's fake. And people use these synthetic faces on their social accounts, fake social accounts as profile images. There are also voices. I mean, you pick up the phone and somebody's voice, you'll hear some of this was maybe it's your friend or family member, but they're not really they're not the real person. It's created, again, by algorithms. And there are also videos. I think there are plenty of them online. You know, there there's a deep fake we know of Volodymyr Zelensky, the Ukrainian president, you know, asking the Ukrainian citizens to put down their weapon and surrender to the Russian. Fake. Right. And there is a fake Tom Cruise, a series of fakes, right? So I mean, there are many different forms and now ChatGPT can create very believable news stories. Those are, again, some a form of deep fake. So, I mean, if you think about this, I mean, every type of media or combination can be created using algorithms. Yeah.

[Kevin]: So, we tell, and I just did a podcast episode about this earlier this year I think. We tell people how to recognize threat actors, phony emails. We say look for language differences or generic terms or punctuation errors, grammatical errors. And what I'm hearing you say is deep fakes can be used to eliminate all of that. So what used to be the signs or indicators of fraud may no longer appear or in some assistance may not appear. So how can we identify a deep fake if it's specifically designed not to be detected?

[Siwei]: Well, there are actually a couple of ways... I want to say, you know, the very high level in the question, you know, we ask, well, the question of whether we can detect deep fakes, I think the what the way I like to approach the question is: under a reasonable budget limit, can we detect out deep fake because you know, if it gives someone infinite amount of time, effort, computational resources, you can... for sure somebody can make a deep fake. Eliminate everything that is, you know, kind of artifacts and nobody can tell that is a fake one, just gaze on that piece of media. So I think it's a right more there. But I say if we are talking about someone not spending much effort on that, there is an algorithm they just throw [garbled] data on the algorithm and then create generated faces, or voices or, you know, videos of faces... those are detectable with the current algorithms. And there are too main themes of detection. I mean, there are actually three themes. I would say the first one is the easy one is crowdsourcing, because you see a symbol, a single face, someone you know will say something that is kind of like unexpected. I mean, this is not this is a low attack where you go online, you pick up the phone, ask for that right person, you know, within a matter of seconds, you debunk the fake. If we say we just limited to the media itself, we do not go for other context. There are still ways to do that. First of all, all those models, even though they look like a very powerful... they create a very realistic, say, human faces. They're not... they do not understand the world, as I mentioned in the previous episode we had, the way those models learn about, you know, doing certain tasks is by taking a large amount of data and figuring out what are in the data. But the data, you know, piling all the data is a very inefficient way of teaching a model about the world. For instance, something very simple, a technique of simple example. We are doing this interview right now, right? So you are looking at me, I'm looking at you. We're both real people and if you look into my eyes, you can see that, you know, there is a light in my room. You can see the reflection. So my two eyes on my you know, on the you, my two eyes. And because I'm a real person, I'm looking at the screen. My two eyes are roughly looking at the same thing. So the two reflections should be the same. That's just a simple matter of a simple fact from the physical world, from the, you know, the physics and geometry, but if you look at some AI generated faces, you look into the eyes, the reflections are different. So one eye, almost like a it's like one eye looking at one thing, the other. I look at the difference and that's simply not possible. And this is a thing like, you know, the models are knowing all of the subtleties of facial skins and facial hair. I can now make that little physical, physical fact, right? I call it the Achilles heels for the generative AI models. And this is the reason we can detect them. I mean human by just eyeballing all those images, we can find a lot of those artifacts. The my favorite example are teeth... computer algorithms aren't do not generate very well, you know, very good teeth especially in the videos. Yeah. They sometimes are missing something or inconsistent so looking for those. But we are we goal is not just like you know humans looking at that we developing algorithms based on all this intuition for instance we have an algorithm based on the intuition I just mentioned comparing the reflections of the eyes and tell us how likely this is a deep fake. So that's one method. The other method is fundamentally those images are created from a different source, different path, right? Like, like the image you're looking at me is really created by the camera, taking my presence required this time as space and save it as a form of image. While all those images created by algorithms, they come from a different world, not in this physical world. So statistically they look very differently. So if we have the right way to look at the algorithm, I make the analogy like x-ray. So if somebody we want to see through human body, we want to understand if there's anything wrong inside the body, you know, just by looking at the body, we cannot get any of those information. But if we change the way we probing the image, we can see a lot more information we otherwise... we couldn't see. So it's the same idea. We developed algorithms almost like you can think about this as analogy of x-rays to human body. We look at the signals from a different perspective and many of the artifacts otherwise will not be visible to us, can now become visible. And then comparing that we can see, yeah, this has come from a different source than that ...then the camera. So these are always in words evolving. We already have a tool and this is supported by the National Science Foundation and ARPA, which is the research branch of the DOD, and it actually worked pretty reliably, I will say, for anything that [garbled] of the model, although we have very good chance of catching them. Yeah, right.

[Kevin]: So one question I have, Siwei, is what is there any beneficial use to this technology, for example, you know, that the automobile causes, you know, tens of thousands of deaths every year, but there's a salutary purpose to the automobile.

[Siwei]: Yes.

[Kevin]: The same can be said for all sorts of things. How did the deep fake technology arise? And was it solely for the purpose of nefarious or criminal activity, or was it some...what was the beneficial purpose and how did it...

[Siwei]: That's what. Yeah, that's a good question. That's why, you know, when I actually give talks, I don't like to use the word "deep fake" or I only mention it because when we say, "deep fake," it looks like it carries this negative feeling, this technology, I call them generative AI technology. They are abundant beneficial uses of it. I can give you a few examples. For instance, for movie industry and advertisements, you don't need the actor to be there. I mean, just think about how much cost and effort that will save for us and let alone, you know, we you talk about an actor or actress who are no longer live...

[Siwei]: Exactly. Like there's a new episode of Star Wars. We have the Princess Leia and presenting her, you know, in her younger age. That was a...

[Kevin]: Better example in mind. Yeah.

[Siwei]: But the movie and entertainment industry is one thing I think there are also really beneficial use of deep fakes. My favorite example is a Canadian company doing this audio deep fakes. So, you know, we have patients, stroke patients. After stroke their language, their language ability got affected and they can't now really articulate their words as they used to be. And that caused a lot of stress on their caregivers or family, you know, their children, because they simply cannot express themselves well. So there is a Canadian company developed this algorithm developed this technology based on exactly almost exactly the same technology behind the deepfake audio is to learn their current articulation voice patterns and normalize their ...or sound their voices as they were normal. And this is based on an algorithm. And you can imagine how much of benefit, how much stress that can reduce from the caregivers. So I think there are definitely good use of those technology. And that's why, you know, we cannot just kill the technology for the bad side effects.

[Kevin]: Yeah, so there are and we've got a little bit of time left. So there are there are benefits and there are risks to the use of this generative technology. Where do you see it happen? Where do you see it going? How do you see it playing out over the next few years? As I can imagine, when we talk about deep fakes, I'm thinking of the scenario you mentioned. I'm thinking, you know, Volodymyr Zelensky is telling people to lay down arms or someone fakes the president declaring or not president...the president announcing a nuclear attack on a nation in order to induce a third world war. Where... how will this play out, do you think, as you sit here today?

[Siwei]: Well, I think there are a couple of things when you realize that this is a fastly developing field. So everything is very dynamic there. There are two particular aspects of this dynamic nature that, you know, we need to pay attention to. Number one is we need to think about the human users. Humans are very versatile. So deep fakes, I like... maybe this is not a very good analogy, but I can compare it with the virus, like COVID 19. So the first line of deep fakes, you know, is struck a few people who are unaware of that. But as we see more and more of them, we started to develop some immune... our immune system will get better. We start to find a way to tell a real from fake... our the ability of telling real from fake will also develop. So eventually we... I think the humans will ...we'll figure out the users will figure all the way of course with the help of researchers, government and media we'll figure all the way to you know better to have a better ability to discern what are deep fakes and what are real. The other side is ... but the other side of this dynamic competition is, you know, we work in the media forensics area. We try to expose deep fakes. There is a competition between us and those people who are making deep fakes. And that's just a reality, you know, the algorithm I'm developing today, may be effective for deep fakes now... there'll be people taking, you know, they read my paper to understand what we're doing, to confuse the algorithm, make them, you know, even more powerful. So I think one thing is, number one, as a researcher, it's my responsibility to continue catching. You're keeping up the pace and developing more and more effective detection algorithms. On the other hand, I think the public, the government need to put a lot more attention paid, a lot more attention to these countermeasures, because just compared to the resources we've got and those people who are, you know, making generative AI models is a hugely imbalanced war. If we talk about in terms of the investment of money, we're on the losing side and it's a huge gap between, you know, the investment there. So I think, you know, increasing awareness of the general public, increasing investments into the countermeasure research area, those are the two most effective ways. On the other hand, I think government, the regulation could also play a role, but not in the way of, you know, banning the technology behind this or, you know, stipulating say, you know, you cannot you can do a certain thing. You cannot do that. I think the government can. What a government can do is make sure that all of the media created, that the two providers say open the I have open AI ChatGPT or, you know, Stable Diffusion for their image journey to put on watermark on their choose. So anything come out of there is we'll have a clear indication to the to the viewer that these are not real these are created by model the same way as, you know, we watch a movie, it will say "based on a real story," but it's not a real story we should do that, you know, to make sure that everybody is aware of that. Yeah. That they'll be my answer to this.

[Kevin]: You know, I appreciate that, Siwei. There is a lot to talk about here, but we have to leave it there for now. But I want to have you come back and talk more sometime about deep fakes and particularly how we can spot them and we can enact rules that may make it more difficult for the threat actors to confuse us into fraud and identity theft and other harms. But absolutely. Thank you. Thank you so much for coming on. I really appreciate it.

[Siwei]: It's my pleasure. Thank you so much for having me. Yes.

[Kevin]: Professor Siwei Lyu of the State University of New York at Buffalo, professor of computer science and engineering. Our thanks to him and our thanks to you for joining us for this episode. We'll be back soon with another one.

[Kevin]: The *Cyber Sip* podcast is available on barclaydamon.com, YouTube, LinkedIn, Apple Podcasts, Spotify, and Google Podcasts. Like, follow, share, and continue to listen.

This material is for informational purposes only and does not constitute legal advice or legal opinion. No attorney-client relationship has been established or implied. Thanks for listening.

Barclay Damon Live *podcast transcripts and captions are automatically generated through artificial intelligence, and the texts may not have been thoroughly reviewed. The authoritative record of* Barclay Damon Live *programming is the audio file.*

Thanks for listening.