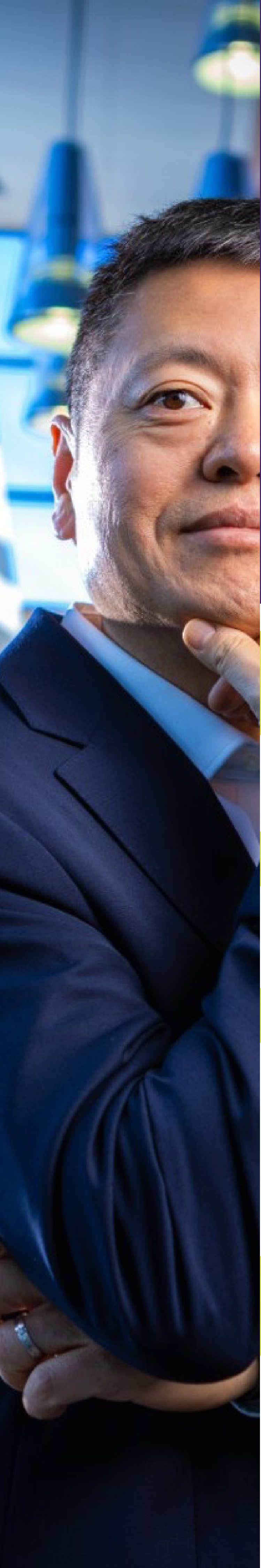

ScienceNatters

*In this episode of Science Natters, **Professor Siwei Lyu** from **University at Buffalo The State University of New York**, talks to us about his research in the field of generative AI and media forensics, which focuses on the detection and prevention of deepfake media.*

In this conversation, we speak about how we can protect ourselves from being tricked by fake images and videos that we might see online, the potential dangers of these deepfakes and why it is so important for all of us to be aware of these dangers.

Contact information: siweilyu@buffalo.edu

Reference: Sun, C., AlBadawy, E., Davison, T.F., Robinson, S.R., Chang, MC., Lyu, S. (2024). Using Vocoder Artifacts For Audio Deepfakes Detection. In: Nowroozi, E., Kallas, K., Jolfaei, A. (eds) Adversarial Multimedia Forensics. Advances in Information Security, vol 104. Springer, Cham.

doi: 10.1007/978-3-031-49803-9_11

Funder: PI, SaTC: CORE: Small: Combating AI Synthesized Media Beyond Detection, National Science Foundation, Project SaTC-2153112, 2022-2025, $616,000.

00:54 – Hello Siwei, welcome to Science Natters. I normally start with three quick-fire questions. The first one is what's your favourite meal to cook?

Siwei: That's interesting, I didn't expect that question. I like to cook Indian food. According to my kids, I do a fantastic beef curry stew!

If you could have a superpower, what superpower would you have?

Siwei: I would like to be super focused on things – that's really important for me, and it's also a feasible superpower.

If you could be an animal, what animal would you be and why?

Siwei: I think I would probably want to be a hawk because you can fly up high, see the situation better and choose your moment to go down. And that's the way I do research. I go around the field and understand the general situation. Whenever I identify some opportunities, I will go directly to them, so a hawk is a good model for me.

02:18 Could you give us a hawk's eye view of your research, a very brief overview of what a deepfake is and the technique that you're developing to detect them.

Siwei: I do research in this area called multi-media forensics or some people call it digital media forensics The idea there is to use algorithms to expose manipulated or synthesised multimedia content. This includes audio, images and videos. Since 2017, we have seen a rapid development of AI-generated synthetic content – deepfakes. My research has been shifting focus to the mitigation of deepfakes using AI technologies.

To expand, deepfakes are audiovisual multimedia created or edited with the help of artificial intelligence algorithms. Now, we detect deepfakes by learning, by understanding this whole process, how the generative AI models are creating this media. One of the important ideas is that even though they (AI algorithms) learn, they're learning without actually understanding. So, a lot of times, the models are able to create something that looks real at a glance but, if you look deeper, there are many important regularities of our physical world or our bodies that do not look right.

I'll give you a simple example. Many of the AI algorithms can create a human face and human body to the finest level of detail. But they make some mistakes, like a hand with six fingers, eyes that are not looking in the same location. Or in physics – you see someone standing in the sun without a shadow. These are the ways we try to detect deepfakes. Supplementary to that approach, we also use data driven approaches. The idea there is I just give them (the algorithms) tons of data at this time, some of them are real, some fake. Then, I start with an algorithm with a blank slate and the algorithm will start making decisions at random. But what we do is almost the same thing as when we train a dog or teach a little kid to learn certain skills: every time the model makes a correct decision, we'll give an incentive to the

algorithm. So basically, we tell the algorithm "next time try to repeat what you did last time". If they did something wrong, we'll tell them, "No, this is not right, you shouldn't do this. Next time, try to avoid making similar mistakes".

So doing this, with time, with enough numbers of data, the algorithms will do the same as a cat or a kid – they will learn. They will grow their abilities by this repetitive interaction with the data, with our feedback. That data driven algorithm can actually figure out and they can make decisions, so we combine the more semantic-based detection and this data driven detection together to develop more effective detection algorithms.

05:44 Thank you. That was a really clear explanation of how the two different methods work. There's something I was thinking of in the lead up to this podcast. I saw something on the news this week about a fake telephone call that was going around with Joe Biden telling democrats not to vote in the primaries. It made me think that this year there's going to be loads of elections all over the world, billions of people voting. Some our listeners might not be old enough not to vote yet, some of them might be, but what can they do when they're on social media or online to protect themselves from deepfakes and misinformation? How can they make sure that what they're reading is accurate?

Siwei: First of all, I'm glad you used the example of Biden's voice. I was actually helping some reporters authenticate that voice, and we used our algorithm to detect it as an AI generated voice. So, if you search for that news story, my name will pop up. We're in the first line of defence of this AI generated media.

Coming back to your original question, what is the impact of deepfakes on elections? I think that is truly a serious issue. I think up to this point, it's very likely that it could influence opinions and cause some damage. I think the line of defence for this one is the news media working on this promptly so whenever something shows up, they immediately do some fact checking or find some people to help them verify the authenticity of the media. So, relying on reputable media channels for updates on elections is very helpful. Another step of quick fact checking; just do a quick Google search to confirm that at least there's a different opinion about this. I think that's important. And I think underneath all this, is having an elevated level of awareness, knowing that these days we have the capacity of using generative AI models to create audio-visual media like this and they could sound and look real. If we know that, next time we see something, at least we'll have this little red flag saying 'maybe this is wrong, maybe this is not real.' If we have that level of protection, then I think we'll be free of this problem at a much better level.

Just being more aware and taking a more active part in what you're reading. If something sounds wrong or surprising, then check it out somewhere else.

Siwei: Exactly.

08:52 One of your hopes for the future is to train the next generation of researchers in the field and help your students come on. Why do you see that as such an important aspect of your work?

Siwei: I think exactly because of the situation of deep fakes. The generative AI technology behind deepfakes is growing so fast, and we need more people trying to play the defensive role and able to apply technology, knowledge and skills to protect other people from attacks using deepfakes.

We have been talking about elections. Those are at the national level or a political level, but there are also other ways that people can become victims of scams driven by AI models and deepfakes, like personal scams, financial scams. Somebody you know calls you and says, "I need money. Loan me some money and send it to this account" or worse, for teenagers or kids, somebody may impersonate their parents and ask them to meet at certain places and that could lead to some disastrous situations. That's why we need to protect everyone from this problem.

That's the social impact of this. In terms of the research or the scientific angle, this is also a very fascinating research area. So, what is the best thing we can do? On one side, you learn all this at once, computer science, AI, mathematics, programming, and on the other hand, you play some detective work. If there's a piece of deepfake online, I have this urge to figure out what is wrong with it because it's like an ultimate puzzle, and as a researcher I can use what I learned to combat this problem.

10:51 It sounds like the perfect balance of doing something good for society and the people around you, but I can also tell you enjoy what you're doing and you're curious.