NARRATOR: Deep learning, part of artificial intelligence, is being used to create fake videos that look and sound like the real thing. But could this same artificial intelligence be used to turn the tables on these nefarious acts? In this edition of Intersections: The RIT Podcast, Professor Matthew Wright, director of RIT's Center for Cybersecurity Research, talks with John Sohrawardi, a Ph.D. student in the Golisano College of Computing and Information Sciences, about software they are creating that uses artificial intelligence to help journalists detect and root out such deep fake videos.

JOHN: The first deepfake I saw was the Key and Peele one, the Obama one.

MATTHEW: Right. Me too.

JOHN: That was meant to raise awareness about deepfakes and it did.

MATTHEW: For sure.

JOHN: It made us do research on it.

MATTHEW: That's the one where you have Jordan Peele, the actor, and Obama. So, it starts out with Obama's face, and he's talking about how people can be made to say or appear as if they're saying anything at any time. Then the reveal is it's actually Jordan Peele doing all the talking, and Obama's face is being manipulated to match what it is that he's saying. And that's all largely being done by AI, primarily, with the deepfake technology. Yeah, that's a real eye-opener when you see that for the first time. You're like, "You mean a computer did this?" Then the potential is so obvious because they used Obama. You see some of these other ones now that they put Nicholas Cage's face into some movie that he wasn't in, and that's funny. But it's when you see actual world leaders, and they could be made to say anything, and now people are going to probably believe it because it's video. It's not just a photo that could be Photoshopped. It's video. Video should be hard to manipulate this way. But no longer.

JOHN: So, deepfakes, if you break it down, are deep-learning-based fake videos. It's face swaps, mostly, or face reanimations where you can create a video of a person saying anything you want pretty much. And it uses deep-learning technology and it's mostly automated without you doing too much to make it work. It started off with something good and positive. It could cut down movie budgets, it could be in museums. I think there is already a museum that uses it. So, they reanimated Dalí. You go to a painting and Dalí, himself, a holographic image of Dalí, is there. He's describing the painting and he can even take a selfie with you.

MATTHEW: So, John, you mentioned how deep learning is involved in deep fakes. Maybe you could tell us a little bit more about what the deep learning is that's going on in making these videos?

JOHN: Deep learning comes off machine learning and it basically learns the patterns of anything that it's given input of. For deepfakes, the more traditional deepfakes, it learns

more intricate positioning of the faces. It learns what the eyes, the nose, everything looks like for one person. And it learns the same for another person. And then it's able to filter it out and then somehow it does magic. [laughter]

MATTHEW: Well, it's learning a representation of this person.

JOHN: Yes.

MATTHEW: And then it's learning a representation of that person. Then it's able to, essentially, swap parts of the representations. It's like a numerical vector. This is the essence of Obama in a video. And then you have this is the essence of Jordan Peele in this particular video or here's what he's doing in this scene in this video. And then you're able to take what it is that Jordan Peele is doing and then transfer it on to this other representation that's this is what Obama is in video and then re-expand that back out into actual working video. Buy, yeah, it definitely does seem like magic. And in terms of where cybersecurity fits in, why we're interested in it, the way I see it is that we have these issues with deep learning as this up-and-coming technology. And cybersecurity is always the application of security thinking to any computing technology whether that was people looking at networks or operating systems or now we have deep learning. What are the security implications and risks associated with deep learning? One of these is deepfakes. And people can be generating these deepfake videos and you've got the need to detect that those are actually fake and not real videos. Or to understand more about what is in the fake or why someone is putting a fake out. Those become important parts of the news, potentially. And then there's other aspects of deep learning that are also interesting from a security perspective about why these networks, which are so good at understanding so much about what we used to think is something only people could do and not computers. And now they are so good at image classification, for example, that they beat humans at this. But if you can just tweak the image a little bit, you can completely fool these systems. And these sort of surprising results about deep learning that come out if you're looking for things from an adversarial perspective. What could go wrong?

JOHN: You can give a computer an image of a truck, and in a normal case it will recognize and classify the image as a truck. If you alter it a little bit, all of a sudden it will become an ostrich. Although you see it as a truck, the computer sees it as an ostrich.

MATTHEW: And the amount of manipulation is so small. It's clear that the algorithm is really fragile. That means it's really dangerous because as soon as you put that into a self-driving car, bam, you have accidents. You put that into a tool that's going to classify malware to put this back into the cybersecurity context. You have bad guys creating malware, and they can manipulate their malware a little bit and it tricks the classifier in to thinking that's not malware. I don't know what that is, probably just regular software and it's fine. There's a lot of possibilities for attackers to take advantage of these weakness if we rely too heavily on these deep learning models.

JOHN: This is the cybersecurity space of it. We now have to protect the computers' deep learning algorithms.

MATTHEW: So, John, why don't you tell us a little bit about the detection mechanisms that you and the other students have been building for this deepfake project?

JOHN: We've been approaching the problem by fighting deep learning with deep learning in a sense. The more obvious way of going about it is learning what the fake ones look like – the fake faces in the deepfake's case. What the fake faces look like in different types of fakes and then the original ones. And then you just try to learn what separates them. That just involves building a huge data set of fakes and a huge data set of reals and then using a different deep learning network to tell them apart. But that's also a weakness because somebody else could just build a better fake and learn from whatever you built to detect it.

MATTHEW: So, how have we been at least trying to combat that to a degree?

JOHN: One way of combating that is building different types of methods to detect the fakes. So we design different types of detectors – audio detectors and synchronization detectors. Because one of the weaknesses of deepfake creation methods is a deepfake algorithm will look at each frame and think of it as an individual frame. It doesn't realize that it is a whole video. So it will swap out faces from each frame and then just join it up back into a video. Yes, you could do some smoothing afterward, but it will still be a video created frame by frame with swapped out faces. And that, we think, is one of the weaknesses of deepfake methods, at least right now. One of our methods just targets that weakness and tries to detect whether the video is a deepfake or not a deepfake. And we have been doing quite well at that for now at least. Until they make better ones, and we have to improve our methods. It's really fun for me because I get to work in different fields for this. I'm working on both the user study side where I get to interview journalists because we're building a tool to help journalists detect deepfakes first so it doesn't get into the news. We find out what their current verification process is, what they want or expect from a tool, what their performance expectations are and interface expectations. And I also build the interface, so I do a lot of coding. And I also work on the deep learning models to detect the accuracy of deepfakes. But we also have a big research group and other people working on audio detection models and video detection models. So, yeah, it's been fun.

MATTHEW: In terms of this research, I think it's a really great example of the type of research that we do at RIT. It's interdisciplinary because it brings in the user side – we also have a journalism professor who is part of the project – and, of course, the technical side and bringing those all together under this umbrella. Because the Global Cybersecurity Institute is about solving cybersecurity problems using all the techniques in the toolbox. So, not just traditional computer science skills, but all the different skills that are needed to really address these problems because they are interdisciplinary problems as well. And then another thing that I think is really reflective about RIT in this project is that it is a real world project. We're not developing theoretical tools to do this

detection. We're building an actual tool that we are designing to work with actual journalists. And we have been out there interviewing journalists, making sure that their needs are being met by the design of the tool. And we're working toward having a really robust and reliable and usable tool in their hands in the next few months. That's something that I'm really excited about. That we're making an impact on the world doing something as we prepare for this upcoming election and the possibilities of deepfakes coming out during the election season. That we are doing something to help journalists be prepared for that, to me that's really the exciting and rewarding part of the work.

NARRATOR: Thanks for listening to 'Intersections: The RIT Podcast,' a production of RIT Marketing and Communications. New episodes debut on the first and third Thursdays of each month. Subscribe to 'Intersections' on iTunes and TuneIn, or visit us at www.soundcloud.com/rittigers. For more about our university, visit www.rit.edu.