[music]

Paul: When it comes to weaponized AI, you can forget Skynet, Tron and Hal 9,000. The AI villains plaguing commercial operations these days come in the forms of poison data sets, hijacked, AI models, and adversarial samples. The world of malevolent data is no longer just the stuff of science fiction, but a reality that cybersecurity experts, data scientists, and supply chain managers all need to be cognizant of.

Hello, I'm your host, Paul Thies. In this episode of *If/When*, we explored the topic of adversarial artificial intelligence with Dr. Jennifer Bloom, Senior Director and Data Scientist for Jacob CMS, Cyber Intelligence Business Unit, and Charles Ramsey Director and Data Scientist Jacob CMS, Cyber and Intelligence Business Unit. Well, Jennifer, and Charles, thank you both, so very much for joining me today.

I'm really looking forward to talking with y'all as we talk about artificial intelligence, and as we started putting this episode together, the fact that you could use artificial intelligence to attack other artificial intelligence, to me as a layman, it sounds like cutting-edge cybersecurity stuff, and it's really fascinating to see where the technology's going. I'm really looking forward to sitting down with both of you today and diving into this, so thank you both so much for joining me today.

Jennifer: Happy to be here.

Charles: Thank you, Paul.

Paul: All right. Just to jump in. Jennifer, my first question is for you, and can you describe for us what are some of the more common types of AI-empowered attacks that affect commercial operations?

Jennifer: I think that there are three types. One is AI model theft. One is adversarial samples, and one is training data poisoning. I'll go into a little bit of detail about each of those three, but AI model theft is pretty much when you hijack an AI model. You have a model that's trained and you embed it with some vulnerability, like a hardware ship or on a cloud network.

Then you have cybercriminals that can hack the systems or worse yet they can reverse engineer the machine learning models. Then you have what's called adversarial samples, which is where you essentially introduce mistakes into your AA model, so you've manipulated the data so that the model actually learns incorrectly. It incorrectly identifies things. One of the examples is you have a self-driving car, and instead of seeing a stop sign, it recognizes it instead as say a speed limit sign which as you can imagine, can have terrible horrible consequences.

Then finally we have the third one, which is training data poisoning as you're training the model, you introduce incorrect information, and so that's a bit different than what I mentioned before, but essentially you train the model incorrectly. Before, what I was saying was with the self-driving car, is that it knows what a stop sign is. You just introduced stuff to make it think it's something else.

In this case, you actually trained it so that it never sees a stop sign. It'll only see a speed limit sign, and the real danger with that is that you don't realize until it's too late, that this manipulation has taken place. It's really hard to find doubt where the mistakes are. It's really hard to undo, and so I think that's one of the harder things that's happening right now to fix.

Paul: Wow, that's fascinating. How does that happen? Is it like somebody hacks a system and introduces data that way or is it like social engineering? Does somebody on the inside poisons that data model all of the data pool, and I've seen some things where it's like, they'll take a picture or like a sound or something in the layer, like white noise over the sound, for instance, that's not detectable by human ears, but like the AI, it just flips the AI out, or they'll put something in the picture that your human eye could not see but it like totally messes the AI. What's the avenue of attack? Is it outside or is it like an inside job or is it tend to be both? How does that even work?

Jennifer: It's totally both. You can definitely have an insider and I would argue that the easiest way to do it is you have someone plug something to a computer a network, a USB port, and you can just introduce a malicious activity that way. The other avenue, of course, is when the user does it by accident, you clicked an email that you shouldn't have. You opened an attachment that you shouldn't have.

You went to a website that had malicious code and like anytime you get a text that's spam or they mimic it saying, I don't know if you've ever gotten a text that says it's from your bank. They need you to click this link, but it's not your bank, and so all of those are entry points. I'd say from both sides, both the user accidentally being tricked, so to speak, and then having a person on the inside using actual physical devices to introduce malicious content.

Paul: Now Charles, Jennifer brought up the idea of weaponizing self-driving cars, and the mind just can think of all kinds of different scenarios. Maybe you get in a car and autonomous vehicle and suddenly you're locked in and bad actors won't let you out until you agree to pay a ransom or something who knows. Let's talk about potential attacks that can be deployed against the common good such as weaponizing self-driving cars or attacking healthcare institutions. How likely are those kinds of episodes and how can we insulate against those?

Charles: Piggybacking off of Jen's great breakdown of the different types of AI. I think we would agree that there's so much good that AI can provide for the common good of those self-driving vehicles for healthcare institutes. I'll break down some thoughts on each one, but first, the colleague covered English, and this is potentially a simplified version of the breakdown that Jen provided.

There's a distinction between the attack itself between adversarial AI and weaponized AI, so this getting weaponized AI could include legal, autonomous weapon systems that we've all seen on like the show slaughter bots and YouTube videos. The idea is that weaponized AI attacks the actual system, but adversarial AI, AI systems can be confused by inputs to make bad decisions. An example was what Jen had talked about with AI poisoning or training poisoning and you mentioned it as well.

The focus is on the AI algorithms and the models themselves, and it may be used within weaponized AI attacks, but when you're going after the actual algorithms and models, so an AI model and attacking the training data in a training data poisoning example would be a single neural network. There are multiple layers and each different layer has potentially a different weakness that can be interjected by some malicious algorithm.

This could be totally undetected and it's in the noise to speak, but yet it impacts the overall model. Weaponized AI automates the attacks on the common good services mission, whereas the AI modeling attacks or the training data poisoning attack could be used for weaponized AI. Unfortunately, yes attacks are likely common, good services like transportation, self-driving cars that you had mentioned, smart cities, and these include security monitoring for good healthcare as an example.

I'm impressed with the availability and scale of AI tools that are used today for good but now we can imagine that AI autonomous attacks are happening. Taking a hacker, who's going after a healthcare system now automate that to scale from an AI attack that is just relentless and the frequency where we just can't keep up and then specifically in healthcare. An example that I give was actually an experiment is a body area network in healthcare.

Imagine an array of medical devices that sit on a medical-grade network and they share data from station to station. That includes the diagnostic, and this is a real growth area. Cyber security-wide survey makes it that we can see around 41 billion IoT devices by 2025.

Paul: Wow.

Charles: That attack surface is growing in the medical field. This is a sidebar, but as a proof of concept, we explored a heart rate monitor wireless system, or arm band and we started out as a joke, we intercepted the wireless signal from the heart rate monitor, and then we had a laptop nearby that shuffled to the music based on the beat per minute, that it was reading from a patient.

I looked it up, staying alive is 124 beats per minute. If you're running at 124 beats per minute, then we would shuffle that song. It was funny. You'd shuffle that song, and you'd hear staying alive in your playlist, but that's innocuous. It's more important, we were able to control the display and the control of the heartbeat shown on the screen. Imagine this a type of attack at scale to where AI is going in and changing all of the heart rate monitors or if impacting them simultaneously so that you can no longer trust the devices that are growing in scale you're no longer able to trust them due to these attacks.

Paul: Wow. It's of a new way of thinking about deep fakes in a sense too healthcare data and it's amazing the this idea of emerging technology escalating arms race because the attack surface with the proliferation of IoT devices and then with the advances in AI, it just seems black hat, white hat, it seems like it's just going to continue to exponentially grow and cybersecurity experts and data scientists and all these adjacent disciplines are going to have their work cut out for them.

Jennifer are taking this mindset and putting it maybe a little more subtly in terms of not necessarily just attacking the common good, but attacking supply chains. Maybe competitors are doing skunk works on each other's supply chains and stuff. Can you, Jennifer, describe some of dangers that adversarial machine learning and AI could present to an organization's supply chain.

Jennifer: It's a lot of things, but the main one is loss of personal information, not just not just the employees outside organizations, but of course air the organization's sensitive data itself. If you had trade secrets, there's a chance so that might be breached or compromised, but industries all around are suffering from just an explosion of cyber attacks especially on the software supply chain.

I think one of the most notorious instances in recent times is the of a hack involving solar winds where they breached a software network and they planted malicious code and it was dispensed to thousands of vendors and vendors customers. Once the update was installed, all the customers were affected and infected and they weren't to breach themselves. Another danger I found is that, especially with a lot of commercial vendors are using open source code and third parties to achieve their results for applications so your third party applications for your time sheets, for your emails, collaboration platforms, file sharing.

A lot of the times the end user has no insight into the security of your code that's being used and that's definitely a back door for adversarial attacks and the end user would just have no sense whatsoever that it was happening. One thing that's also I would mention in terms of supply chains is that it can definitely, Charles mentioned a lot of things in healthcare and I just want to piggyback on that, but in terms of fraud that's another thing that could definitely happen.

You have say, you're tracking prescriptions or you're tracking medical equipment a lot of that could be miscalculated. I say, in quotes, "the computer says you have this many, but you actually have that many," you can cause a lot of disharmony between not just healthcare, but any organizations if you had military equipment, any of those things can cause a lot of problems.

Especially if say you had soldiers on the ground that needed set equipment but you've introduced adversarial attacks and code that would make it so that they didn't get what they needed in the timeframe that they needed so it's not that we wouldn't notice that it was happening. It's just the question of that it'd be too late.

Paul: Now, Charles, how can an organization identify potential adversarial machine learning, AI pitfalls and their supply chains and then what steps do they need to take to mitigate those dangers?

Charles: Identifying potential adversarial, MIML and AI pitfalls. I love the example that Jen gave on solar winds and that's perfect example of attack on the supply chain to where early on introduced deep down into a module of a module potentially firm where I think the malware was called sunspot. I believe for solar winds and log for Jen is a similar example to where open source Java with malware embedded within it. From the Linux foundation when software is being developed it's compiled so to speak from source code and that software is used.

Jen also mentioned open source so imagine this is an open source module that you're used to using, but within that module you've got in a different copy and so your module is using a module, is using a module, is using a module on that last module so it's like a butterfly wing where that impacts the code maliciously so that's a pitfall. How can we look at each element within the supply chain and software and building a AI more to the point? How can we better do that? Couple of recommendations.

One, we can take what we've learned from the cyber practices and perform zero trust audits on open source modules so hash checking is not enough, which is what's being done today. We also use which I think would be interesting is use AI programs on the software themselves, so that they become assistant software code auditors for us. Then there's a thing from Linux foundation where they create and verify reproducible builds.

A build that can always produce the exact same outputs given the same inputs so that you verify the build before you put it into production. Then the second recommendation of course in AI itself, cyber training, but also the processes in software development workflow. If I'm trained to look for phishing attack as an example, but my processes are such that I'm supposed to click on a link to submit my code as just an example, then the process does not match the training.

Similarly with the AI, if I'm building a model and I have these particular processes, and one might break some type of cyber training that I've had then that itself causes an issue. Then also there's there's government entities that are working on recommendations for AI protection. The CISA, DHS Cyber and Infrastructure Security Agency. They have this concept up of a key building block.

What that is, is a software security baked in to these modules, which I think would go a long way to helping us so that we can now trust those building blocks of considered a software build materials is what I think they use, but the building block that we trust and there's of course, thousands in these building blocks we now trust our build and our AI models more. Then finally an automation of all the above so we continue to look for ways that AI can assist these proposed processes to mitigate potential dangers.

Paul: Excellent. Jennifer carrying that forward a little bit we're talking about AI as an attack technology, but picking up a little bit on what Charles was just saying there in terms of AI as a defense technology. Can you, Jennifer talk a little bit about how organizations could use machine learning algorithms to help identify weaknesses and potential adversarial conditions in their supply chains?

Jennifer: One thing that Charles mentioned was of course using AI to assist us like you just said statement and that's definitely a possible, the one thing I think that's very important is not so much thinking well, how can we beat them? Do we need to be faster? Do we need to be more efficient? Can stop these things from happening? I think what's most important is the techniques that are being used.

Questions that have been asked is how do we essentially plug up these adversarial back doors that can be entryway to maliciously affecting our code. One thing that I technique I really liked was called a mode connectivity and without boring every anyone too much but the idea is that you pretty much look at a, you have two models side by side and what it was originally designed to do was to help generalize them.

When they are trying to identify something they can work in harmony. That was what it was initially designed to do. The side effect of that is that they realized it's actually really perfect plug up backdoor. When they made it so that these two models are being generalized with each other and you don't minimize their accuracy and the developer has full control over what points to pick, like, how am I going to make these models relate to each other?

It's the developer doing that, that makes it very hard for a malicious actor to figure out what the developer is doing. Even if they did figure it out, they'd also have to have the clean example, so to speak so the clean data that the developer is working on so it's essentially two steps that can be preventative for malicious code being introduced. The malicious actor would have to figure out how to bypass that and they realize that it's actually really hard for them to do. That's one scenario, and then I'll just mention two others. They're also techniques so again, these are techniques to teach the AI, so it doesn't matter what their programed in or anything it's the techniques themselves that would be used. One is called auto-zoom, which essentially helps developers find black box adversarial, vulnerability so those would be in deep learning models and they do it with a lot less effort that's normally required. Then there's another one called hierarchical random switching, which was developed by some IBM AI researchers, which essentially makes things more random and more difficult for malicious actors to figure out.

Paul: Charles, can you talk a little bit about the idea of deceptive data, how it works, and how organizations can identify when deceptive data has been introduced into their training sets?

Charles: Yes, I think this is a call back to Jen's mention of poisoning the AI training model. From Dr. Charles Corey, one of our senior data scientists, he speaks to deceptive data as being deliberately bad so poisoning the AI model in sometimes subtle ways, sometimes not so subtle ways, but if the enough AI bots mislabel an image, then a person and his example, was Benedict Cumberbatch become something else but more of the points.

Mostly in AI training, one needs a lot of data to train that model so that it doesn't over fit in particular but what if the data itself is tainted or there exists blind spots in the training data set that lead to accurate models so that the model is sound, but still there's some known educations that exist in the distribution that can be exploited by malicious intent. Okay, so imagine blind spots within that data.

It's deceptive in that the omission of key data and intense prevent detection, for example. An example, facial recognition is getting better, right? Considered that vendors did not use to consider masks much prior to the pandemic, but now convolutional networks, CNN models must adapt to recognize faces even with masks. That was unintentional deceptive data that it wouldn't recognize faces with masks.

Now a new issue introduced in due to the pandemic so there are many other means that can circumvent AI training models and what are those means? What's the next mask that a CNN will need to encounter? AI can be developed to look for these weaknesses as a recommendation and preventing this type of deceptive data and so you have a AI mechanism that goes through and ensures the integrity of the data.

Paul: Then Jennifer, are there certain neural networks that are more susceptible to attack than others? For instance, are visual processors, more likely or more easily to be attacked than say, text-based or voice-based networks.

Jennifer: I'm going to say they're all equally able to be utilized or are susceptible to attack unfortunately, a lot of the adversarial machine learning that we've seen has overcome a lot of what we think would be showstopper. Visual voice text, which makes them, I think even more dangerous, and one thing that happened in Texas A&M was that they demonstrated that they could poison a machine learning model with just a few tiny patches of pixels and just a tiny bit of computing power in terms of for a visual process attack.

It doesn't require much these days unfortunately then you have a, what's called a, there's a technique called a TrojanNet, which creates a simple artificial neural network to detect a series of patches. That's something that the attackers can use and I wish I could have better news for which would be safer, but unfortunately they're all equally susceptible and same thing.

We mentioned earlier deep fakes that is something that you can mimic a person's voice. You can mimic what they look like. You can age a person's video or photo up, you can photo them down and you can also just create people from scratch, unfortunately, using deepfake technology. Yes, I would say AI and even with a text-based network, even that's not safe because if there's anything embedded like Charles was mentioning with anything open source you are prime for attack.

Paul: We'll shift our conversation just a little bit. Where do we go from here? This next question I have for both of you, I'll ask you Charles, and then I'll ask you, Jennifer, I'm going to ask you the same questions, but so Charles where do you see the future of adversarial AI going and what are some of the dangers that organizations might be on the watch for?

Charles: The future of adversarial AI, we can start, I think, with a quick thought experiment, if you will, in adversarial, AI and weaponized AI, what is terrorism? Right, so for both international domestic terrorism, the FBI states and defines it with violent criminal acts committed by individuals or groups so bodily harm or death. What are some other fears that we have that adversarial AI might impact?

Things like air, water, sleep safety, Maslows basic human needs. I didn't list them all, but so the international community has banned chemical and biological warfare after world war I yet it still exists. How might a criminal element use that against us at scale? Imagine adversarial AI, somehow manipulating this and using this. We've read about attempts and subway systems and mass mailing building control systems, which are becoming more and more integrated.

I mentioned earlier with the growth explosion of IoT with internet exponentially. Not talking about nest thermostats, talking about things like hospital process, complicated climate control systems, more or critical infrastructure from DHS CSA, so they oversee the protection of our critical infrastructure, but how might these be exploited using adversarial AI? AI can provide automation to these attacks and a more subtler point.

I don't know, I made up this term I hope it makes sense. Autonomy social engineering. We talked about social engineering, phising, baiting, dumpster diving, and you receive an invoice from your colleague and you click on it, and then more information is glean, but what if that were autonomous? Imagine adversarial AI to where you can't glean, you can't differentiate a machine from a person, and we're seeing attempts of this.

When you get fan calls or you get robots online, but imagine with the combination of deepfakes and the more accurate models and autonomous AI applied to social engineering, so that you become an exploited target. I see that in adversarial AI in the future, I also see this thing I made up called bias juking. Statistically, we tend to read more information that confirms to our thinking and juking is a slight diversion from these ideas that may persuade you in other ways.

AI slips into your normal stream of information that you receive either through the news or through friends or through other means, but then provides information and it would be scary if all of this were done via autonomous AI, that we couldn't recognize. An example, social media recruitment of ISIS is an example where they had propaganda pushed out and they were trying to get people to join ISIS. Now, imagine AI assisting in that, or actually taking over that campaign for a criminal or a nation state.

Paul: No. Maybe they are able to read social media sentiment and thing like that. Then it becomes targeted recruiting in a sense, it's like the nefarious version of target marketing they're able to reach those people most susceptible to joining their ideology, let's say, or to take action, then otherwise they might increases the effectiveness which is **[unintelligible 00:29:11]**

Charles: Yes, absolutely Paul. If you look at things like Neurolink, which is gathering more and more information around the human brain, some immediately jump to the Skynet scenario, right, where AI takes over the world. There's a fictional book by Douglas Richards where AI takes over a mind someone's mind so they figure out how to tactically communicate with this person.

That's far fetched granted, but consider an AI attack. That's more plausible like AI figuring out how to manipulate something as simple as sound audio wave. If there's Sonic weapons today, imagine a large scale noise attack that could pose a significant impact. That's another example of yes, there's recruitment idea that AI and autonomous social engineering this bias juking to where it just changes. It tries to persuade you just a little bit, but then also AI being used as an actual weapon.

Paul: Jennifer, my last question for you it's the same set the head for Charles, where do you see the future of adversarial AI going and what are some of the dangers that organizations might be on the watch for?

Jennifer: Charles already said a good chunk of them and I guess I'll just add a little bit, but in reference to say the deepfakes, in terms of, he mentioned using it as a social engineering platform to recruit for say terrorist organizations but also you can do the spread of misinformation. Like you said, tweaking things just a little bit I mentioned a fraud that could happen.

On a very small scale, of course, you can pretend that a dead person is still alive. You can keep saying that someone works for you and actually you can make up your own company and say that there are 200 people that work for you, but actually, none of them exist. It would require different organizations to be more cognizant of who they're going into business with because usually you just do.

What do you do? Do you usually just do a Google search on them? You'd hope to talk to somebody, but in this world of viewing the question is, is the person you see even real at that point if you can make all these deepfakes and you've never actually met any of these people in person. That I think is a scary notion on the fraud front but I would also like to say that, one thing I thought was fascinating was I was able to have the fortune experience to talk with someone who was in the entertainment industry.

He's a data scientist and a data engineer, AI expert. What they were doing was that they would map the patterns of people's facial expressions as they watched their TV shows. Based on certain expressions they were able to say obvious things like they had emotion or an emotional response. Someone asked, well, does that mean in the future could Netflix just buy if I had my camera turned on that Netflix could only show me TV series that they know I would like, that I would enjoy.

One thing that really struck me was he said the purpose isn't necessarily enjoyment it's that I know that it evoked a response negative or positive, and I'm using that powerful response to see if I can get your focus. I think that's something interesting to note that says you're on your computer and your camera, you think it's off, but it's actually on and malicious code is actually studying your expressions.

This goes beyond just looking at your social media feed and looking at the text and trying to do a send. This is actually looking to see, are your pupils dilating when you are looking at certain images, so it can get an even greater detailed understanding of what provokes a response negative or positive because both could be used maliciously depending.

Paul: Wow. We generate so much data in our daily living. It's like, it's just only a matter of time where everything we do is trackable and watchable and discoverable or whatever. That's pretty fascinating. It becomes like, what do you trust? That a whole idea of trust and the media and stuff like that obviously has been part of the national conversation for a couple of years now. Science fiction writers have probably been forecasting it for quite some time.

How do you thread the needle in terms of your day to day reality when so much information like you're alluding to Jen could be misinformation or you could just be being led in certain ways because things that are much more powerful and discernible than you are, are able to see how you react to things and adjust accordingly and then you extrapolate that ideas of the metaverse now **[unintelligible 00:34:15]**

Jennifer: I was going to mention that, yes.

Paul: This consensual universe, we're going to all participate in it's like, oh my goodness, it's a Pandora's box of like who knows. Anyway, I guess that's neither here nor there that's a whole other discussion. Obviously, we're just starting to really understand the dangers and the opportunities that all of this technology affords us. I really appreciate sitting down with both of you, Jennifer and Charles today to talk about this idea of adversarial AI.

Obviously, there's going to be a lot more to be said in the days ahead, but I really appreciate you both for your expertise and your insights. Thank you so much.

Jennifer: Thank you. This has been great.

Charles: Yes. Thank Paul. Thanks for the invite.

[music]

[00:35:24] [END OF AUDIO]