Lee Hutchinson:	00:00	This episode is sponsored by Darktrace, the world's leading AI company for cyber defense and creator of autonomous response technology. From subtle insider threats to machine-speed ransomware, cyber attacks will inflict more than $1 trillion in damages during this year alone, wreaking havoc before security teams have time to investigate. By using artificial intelligence, Darktrace learns while on the job to distinguish friend from foe and when it senses an attack, the AI fights back against the bad guys within two seconds. It's time to supercharge your security stack. Start a free trial at www.darktrace.com/trial.
Sean Gallagher:	00:38	This is Sean Gallagher, IT editor of Ars Technica and welcome to our third and final podcast in our series on artificial intelligence. In our last episode, we talked with experts about how AI technology might be used to defend against data theft by human beings. This time, we're looking into the world of adversarial AI, where artificial intelligence is used to defeat other systems either by hacking through their defenses or deceiving them with input that makes them give unexpected results. Like making a facial recognition system mistake me for Brad Pitt, for example.
Sean Gallagher:	01:10	Back in 2016, the Defense Advanced Research Projects Agency held the finale of its Cyber Grand Challenge, a contest in which teams from companies and universities built automated systems designed to win a hacking challenge known as capture the flag. Seven competitors made it to the finals and for a day, their systems tried to score points on each other by finding bugs in software, patching them on their own servers and attacking them on competitors. Dr. David Brumley, CEO of the security company ForAllSecure and director of Carnegie Mellon University's CyLab put together the winning team. Ars deputy editor Lee Hutchinson and I spoke with Dr. Brumley about that and about how far AI has progressed as a cyber defense and cyber weapon.
Sean Gallagher:	01:54	Joining us now is Dr. David Brumley, professor at Carnegie Mellon and also of ForAllSecure. Thanks for joining us, David.
David Brumley:	02:03	I'm happy to be here today, Sean.
Sean Gallagher:	02:05	So we met briefly at the DARPA Grand Cyber Challenge about three years ago.
David Brumley:	02:11	Yeah, we met at the DARPA Cyber Grand Challenge. It was an amazing contest. DARPA had put out a challenge in 2014 that said, essentially, can we build fully autonomous cybersecurity

systems? Gave people two years to think about and develop, and then had a grand challenge, a big bake-off at DEFCON, and we ended up winning.

David Brumley: 02:31 So we built a system called Mayhem. It ended up winning. There was really three parts to it. The first part was offense, where we found new zero-days in applications we were given. The second part was defense, where we would automatically rewrite binaries when we found a flaw and patch them on the fly. And then the third part was a strategy engine where the goal is to win. So we played Mayhem against the best computers and we won and then we played against the best humans and we did pretty well over the three-day contest. We started out very high in the rankings, then slowly over time, we kind of eked down a little bit. The interesting thing that we took away was that Mayhem was really good at systematic reasoning to find flaws. It was kind of like a computer chess engine.

Sean Gallagher: 03:15 Right.

David Brumley: 03:15 But when you look at humans and their abilities to find new vulnerabilities, they have this thing computers will never have. They have creativity. And so we were able to find and exploit things that humans never did, because we could reason really deep in the program, but humans always had the insight, the other players at DEFCON.

David Brumley: 03:32 The other thing that the computer was really good at was defense. Mayhem was much better than humans at coming up with a patch on the fly and making sure that patch not just improved the system's security, but also met performance and functionality objectives. Where we ended up using more advanced algorithms was in the strategy engine and it was kind of cool. Like I said, our goal was to win and I don't think of Cyber as about creating safe systems. I think of them about winning against the attacker. And so we really built a strategy engine around that concept about how do we win the game? How do we move faster than people?

Sean Gallagher: 04:08 So with that behind you, how has what you did with Mayhem evolved and where do you see it going from a standpoint of defending and protecting and, in some cases, going after adversaries?

David Brumley: 04:23 Yeah, so we wrote a paper based on CGC and one way to think of it is, suppose you're a really small country. You don't have a lot of cybersecurity expertise. What could you do to win against the United States. Right? So you don't have as many computer

security experts. And one of the things that we found was if you got really good at just watching your network and stealing exploits, so every time the US tried to exploit you, you use that exploit against someone else, you could actually do really well in the contest. You could, in some cases, even win. And it comes back to if the US found a brand-new, let's say it found a new zero-day in Windows 10 and it used against this really small country, well, that small country may not have very many Windows 10 devices. So the utility the US gets is a little bit, but the US itself runs a lot of Windows 10 and so that country can start attacking us with that same exploit we gave them.

David Brumley: 05:18 And so we started looking at this question of when you look at computers, it's different than weapons systems. It's different than, you know, mutually assured destruction comes out of nuclear theory, but the same idea applied where you could get into this stalemate where if I'm a country and I have a lot of new exploits coming out, but my own citizens are vulnerable, I'll choose not to use them because if I use them, well, whoever I use them against may copy them and use them against me. So that was kind of interesting. It's very mad-like, right?

Sean Gallagher: 05:50 Right. So out of all of this, what have you learned about the whole realm of adversarial AI in terms of its potential as a tool to be used against us in your research?

David Brumley: 06:02 Oh, yeah. So adversarial AI, well, first, for people not familiar, artificial intelligence research was always developed under this idea of a benign environment, right? Like you're going to learn user click rate patterns for ads and there's not really an adversary. And these systems just break when you throw an adversary in it who's actually going to be malicious. So what we've found so far, and I guess where I've seen the most success is using adversarial AI for offense. Even in the CGC, one place we use machine learning is we would generate chaff traffic. So we would analyze all the exploits we knew about and then we'd use that to generate chaff traffic. And it was so effective that even the commenters during CGC were talking about, "Mayhem wasn't able to land an exploit. Their system must be malfunctioning." And that wasn't the case. We were just launching chaff to waste time against our adversaries because if they're analyzing that, they're not analyzing the real exploits, so those can slip underneath the radar.

Sean Gallagher: 06:58 So do you see AI playing a role in helping to improve software quality in the long-term so that we can improve the basic hygiene of the systems we're protecting going forward?

David Brumley: 07:14 Well, I think, yeah, you have to be smart about it. So one of the places that we saw is even if you're not going to take a computer-generated patch, AI can do better at evaluating vendor patches and determine whether to field them or not. Because if you look at a lot of the exploits out there - and one of the people from the Tailored Access Operations group, Rob Joyce, even give a talk about it - a lot of attackers are not using brand-new exploits, they're using well-known exploits. And the problem is people just aren't getting those patches applied to their system quick enough.

Sean Gallagher: 07:44 Right.

David Brumley: 07:45 And I think if we can remove that barrier, like if I told you instead of just, "Hey, install this update, it's available." If I could tell you, "If you install this update, there's a 99.9% probability that you'll have no performance impact, no functional impact, it'll improve security and all your friends have applied it and it hasn't been a problem." I think you'd immediately say, "Yeah, I should just apply that." And so that's the sort of thing that AI can help us do, is really systematically evaluate how quickly we should roll out patches so that there's not that human delay of someone clicking okay or that human delay of someone in enterprise running it in pre-prod before prod, because until production is patched, you're vulnerable.

Sean Gallagher: 08:28 Right. So this is something that I've been paying a lot of attention to because of the fact that there's so many organizations out there that simply don't have the internal staff to do that sort of work, to do evaluation of what needs to be patched or can't even do a security audit, for example.

Lee Hutchinson: 08:47 Well, and in some ways, doesn't that just move your attack surface?

David Brumley: 08:50 I mean, that's a good point, Lee. People worry about putting all your eggs in one basket and if you have an AI system that essentially is your network defense, sort of automatically fielding patches, what happens if that gets compromised? The attacker has huge power then. I'm less worried about that. In computer security, a fundamental idea is that we have a trusted computing base. The idea is that it's okay to put all your eggs in one basket, because then you just have to worry about that one basket. You just have to worry about it. And if you can make that trusted computing based really, really small, it's easy to verify that it's correct and can't be penetrated.

David Brumley: 09:27 So I think today, the way I look at it is we have this problem of all the computers and all the systems are undefended and we don't know how to roll out patches. If I just had to worry about the security of one system, the auto updater, I think my personal opinion is that's better. But I think the other point is it doesn't have to be all or nothing. You can have degrees where humans can be in the loop when they're available. So when I give talks about Cyber Grand Challenge, I know what people love about it. The tech crowd loves the technology, they love the idea of automatic exploit generation, automatic patching. But if I go talk to a sys-o, really, in the back of his head, he's thinking, "Man, I can't hire enough experts. I want this so that I can augment my human workforce."

Sean Gallagher: 10:10 So to spin back towards the offense and defense sides of things, having been through the research you've done and the Grand Cyber Challenge and all of that, how do you feel this reflects on the future environment we face in terms of dealing with a cyber warfare? Given that a lot of this technology, while it's fairly complex now, is getting easier to use, easier to deploy, and you do have this vast number of potentially exploitable flaws in systems that can just be browsed across the internet by people.

David Brumley: 10:49 I think when you look at AI and you look at autonomy and what we're doing with Mayhem, it's really transformational, meaning people have to think different about how to incorporate this. It's going to make their life easier. I'd also comment, right, it's not just a black box you plug into the network and you don't have to do anything. You do have to build processes around that.

Sean Gallagher: 11:07 All right, so it's going to change how you think organizationally as well?

David Brumley: 11:12 Yeah, absolutely. It's transformational at all levels, so we have to think about things like, okay, I can put in a computer system that can provide high probability whether or not I should field this patch. I'm a big believer in using the autonomous systems, having done this stuff in practice myself, both for offense and defense and I don't think offense is, by the way, just about country A breaking into country B or a hacker breaking into your system. I think it in terms of red teaming your own network to make sure it's safe. I think what computers can do is they can start freeing people from the more mundane tasks, like applying a patch is pretty mundane.

Lee Hutchinson: 11:49 So I keep coming back to this red team, blue team AI thing, which is absolutely fascinating to me, and I know that's sort of

		the core that this entire section is about on adversarial AI. Is there a point where humans are removed from this loop not because of convenience or because we have better things to do, but because the threat and attack and defense landscape sort of gets beyond what we can deal with with reaction times and creativity because the machines have trained themselves into a level of attack and defense that we're not adapted to deal with?
David Brumley:	12:21	Well, I mean, if I was going to guess, like into the future, I think that we will reach that level where the execution of plans is left to computers because they can react so much faster. On the back end, you still need the humans coming up with brand-new insights. So at the end of the day, these attack algorithms like AI, they're optimization functions. So in CGC, we had a scoreboard and we were optimizing to win that particular game. One of the questions always that a human's going to need to answer is, what is the actual utility I'm trying to achieve? How do I think through all the different effects out there and whether or not they matter to me? And once you think through that, you can let a computer do that.
David Brumley:	12:59	As an example, I talked about attackers can reuse exploits and most people immediately think of that as revenge, like if the US launches an attack against Russia, they can copy the bits off the wire and use it against us. That's not what the theory says. What the theory says is now Russia has that and for example, they could use it just against our allies. Maybe Russia and their utility function says that they should attack Australia. And so you have these pretty complicated reasoning game engine things going on where computers are just going to be better about reasoning through everything, but you still need the human to really pop above all that and say, what is the mission I'm trying to achieve? And also, what are our values? What is our utility function?
Lee Hutchinson:	13:40	In your position, do you see potentially that there is somewhat of a international worldwide agreement on the behavior and ethics in AI or do you think this is still getting sort of an area where the US feels one way and other countries potentially feel very different?
David Brumley:	13:55	I mean, I think it's, people all feel different about this sort of stuff. I don't have any particular insight, but from what I see, it seems like the US, for example, for a long time didn't want to do any research in what used to be called PSYOPs, but this idea of persuasion. And then we see Russia doing this in the 2016 election. I mean, clearly Russia has actually effectively

conducted a campaign and they probably learned far more than we would on how to carry out these sorts of campaigns.

Lee Hutchinson: 14:26 Yeah, absolutely.

David Brumley: 14:27 This is something I do think about where it does seem, if you look at it in broad strokes, people like China will break in and steal IP because in China, it's okay for the government to hand over all IP to industry. That's just how their society works. In the US, we have a very clean separation. The intelligence community can never hand over IP to any business to profit off of. That's just in our ethics. And so I do wonder how these things are going to play out on the global stage.

Sean Gallagher: 14:54 Well, at least we know where China got that airplane from.

David Brumley: 14:59 Yeah, I mean it really does speak, it's exactly what we showed in Cyber Grand Challenge and some of our other analysis. Literally, if you went to DEFCON, the world's most elite hackers, and all you did is steal other people's exploits, you got exploited once, but you could use it against everyone else instantaneously, you could always come in second and sometimes that's good enough.

Sean Gallagher: 15:21 There's a lot to be said for second place. I want to thank you, David, for joining us. This has been a great conversation and look forward to following what you do in the future.

David Brumley: 15:32 Yeah. Thanks for having me. I mean, these are topics important to me and it's really great to hear people are interested in hearing about them.

Lee Hutchinson: 15:40 There's a battle happening right now for the world's most sensitive data and cyber criminals are gaining ground. Their sophisticated attacks are scanning for the slightest cracks in the digital perimeter: an employee falling for a phishing email, a cloud application left up without a firewall, or even a smart refrigerator using a default password. Once they get inside, it's only a matter of minutes before your data is encrypted, stolen, or erased entirely. At this point, for most organizations, it's game over. Darktrace has changed that game for thousands of smart cities, international nonprofits, and Fortune 500 companies. With the first ever AI-powered autonomous response technology, Darktrace instantly neutralizes in-progress cyber attacks that are already inside the enterprise, containing the threat without interrupting your normal workflow. Autonomous response is on guard 24/7, on the weekends and

on holidays, intelligently defending your data on your behalf. The reality is that the next automated attack will strike too fast for humans to mount a defense, but with Darktrace, the machine is fighting back. Find out how on darktrace.com.

Sean Gallagher: 16:47 Artificial intelligence can be used to defeat other artificial intelligence in ways that don't exactly meet our definition of hacking today. Lujo Bauer, a professor of electrical and computer engineering at the Institute for Software Research at Carnegie Mellon, has researched ways to use AI to defeat technologies such as facial recognition, making the image processing algorithms believe that one person is actually another just by wearing a specially crafted eyeglass frame. Lee and I spoke with Professor Bauer about his work and about how one AI can fool another.

Sean Gallagher: 17:18 So welcome, Lujo Bauer, who's a professor of electrical and computer engineering and of computer science at Carnegie Mellon. Thanks for joining us today.

Lujo Bauer: 17:28 Thanks very much for having me.

Sean Gallagher: 17:30 So we wanted to talk with you about adversarial AI and I've read through some of the papers you've written on the topic, but to begin with, I wanted to ask you if you could sort of give a definition of what adversarial AI is.

Lujo Bauer: 17:44 Sure. So there are a couple of different things that people might mean when they say adversarial AI. The broad definition is any circumstance where the adversary, the bad guy is trying to use AI against us, which could mean the bad guy using AI to create an attack or it might mean the bad guy misusing our AI so that it behaves in an unexpected way.

Lujo Bauer: 18:08 The narrower definition of adversarial AI is that often the way we use AI, or more specifically, machine learning is that we train a machine learning algorithm such that later we show it various inputs and it does something in response, it gives us some sort of answer in response. We show it pictures of faces, it tells us who these faces are. And now adversarial machine learning, in this case, is when an adversary comes up with a way to create an input to the machine learning algorithm such that if you or I looked at the input it would look normal to us, it would look just like a picture of me or a picture of you, but the machine learning algorithm which otherwise works perfectly would recognize it as a picture of somebody else.

Sean Gallagher:	18:52	Okay. Right. Now, I understand that you did something where you printed glasses frames that changed how facial recognition recognized you.
Lujo Bauer:	19:02	Right. So one of the spaces that I was working in with respect to adversarial AI is, this is based off of face recognition. So people got quite excited about this several years ago when machine learning algorithms for face recognition started doing better than humans could at recognizing people's faces. And the same was true for object recognition. So algorithms that would look at pictures and tell you what's in the picture, they started outperforming humans. And at about that time also, people noticed that these algorithms had this weird behavior, that even though they worked incredibly well under normal circumstances, if you wanted to, you could tweak some images so that they would look pretty much normal to you or me, but they would get recognized as having a different object in them. And this was clearly weird behavior, but it wasn't really clear whether it was dangerous behavior because it wasn't clear when an adversary, an attacker might have so much fine control over what the input to one of these machine learning algorithms was.
Lujo Bauer:	20:03	And so what we had set out to do is try to understand for more realistic settings where you might actually use these machine learning algorithms, in these settings, does the attacker still have enough power to cause the algorithm to misbehave? So we were thinking of settings, for example, when I'm trying to log into my computer via Windows Hello, or maybe I'm passing through an airport and there is face recognition being run on the surveillance tapes that take video of me. In a circumstance like that, the adversary has fairly limited control over the image. They can't change arbitrary parts of the background of the image, but they can maybe change their own appearance in specific ways.
Lujo Bauer:	20:43	And so we set out to see, well, if the only thing the adversary can do, can change their own appearance, for example, by putting on a pair of specially designed eyeglasses, is that enough that he might be able to fool a a well-designed machine learning algorithm, a well-trained machine learning algorithm? And it turned out that it was. My students were able to come up with an algorithm of their own that would design a pair of eyeglasses such that if I wore that pair of eyeglasses, I could impersonate somebody of my choosing.

Lee Hutchinson:	21:15	And this works by focusing change, I guess, with the eyeglasses on the specific points that a typical facial recognition machine learning algorithm would key on in order to differentiate faces?
Lujo Bauer:	21:27	Yeah, so that's actually a really interesting question. So it turns out with a lot of these modern machine learning algorithms, it's not really that the algorithm has very specific features that keys on. It's that when you're training the algorithm, you just give it lots of images labeled, in this case, labeled with people's names and yet you let the algorithm figure out what it thinks is important to key off of. And so you, as the person who trained the algorithm, you don't really know. These algorithms then have this weird property that even though under normal circumstances, they might not pay much attention to eyeglasses at all, there's something about this particular image with these particular eyeglasses which causes the area of the face that's covered by eyeglasses to be important to the algorithm.
Lee Hutchinson:	22:15	This kind of thing is absolutely fascinating to me because it speaks to the common misunderstandings that folks like me have about the nature of machine learning. When you talk about facial recognition or whatever as a function, you as a human think that facial recognition means facial recognition. You're not actually recognizing faces in the heuristic, neurological way that a human recognizes faces with something like this. It happens in a way that, I think when it gets represented in the media, often you hear it referred to as like, "Nobody's really sure what the computer knows or what the computer is keying off of, what this algorithm is." It's looking at something, but it's looking at things that it has determined as the differentiators rather than the things that you and I would think of. Would you say that's a good characterization?
Lujo Bauer:	23:01	Absolutely. And specifically it is paying attention to the specific data set that you used during the training time when you trained the algorithm, the specific features that were important for that data set. And when you and I recognize people, we recognize them not just based on maybe having seen a person for a few minutes 10 days ago. We also do it based on decades of experience knowing what humans look like from different angles and all of that kind of background is absent from one of these machine learning algorithms.
Sean Gallagher:	23:34	We see these kinds of behaviors in things like facial recognition where, when the training set is tilted in one way or another. For example, there was a big issue with African-Americans not being recognized by facial recognition or being identified as someone

else because the training set was smaller or because of the way the algorithm was developed, it was biased against them.

Lujo Bauer: 23:55 Yes, absolutely. And certainly I've seen several examples of faces of different skin tone being differently, successfully recognized. Another quite famous example is of, I think this was a algorithm, was being trained to to differentiate between wolves and dogs or a particular species of wolf and dog and it turned out to work incredibly, except after the fact, people realized that what the algorithm had actually learned to do is distinguish between whether there was a snowy background in the picture or not. All of the training images for one of the animals had snow in the background.

Sean Gallagher: 24:39 So with the algorithm that your students developed, and this is sort of black box hacking, I guess, since you don't know what the algorithm is, necessarily, in the facial recognition system, do they use a machine learning algorithm to try and measure the outputs until they got the desired output in terms of generating the eyeglasses frame design or something like that?

Lujo Bauer: 25:02 Yeah. The space that we normally work in is a white box space. Namely, we assume that the attacker has access to the algorithm that they are attacking, because in practice this often is the realistic worst case scenario.

Sean Gallagher: 25:17 Right.

Lujo Bauer: 25:17 The attacker manages just to bribe somebody or something along those lines. But it also turns out that not having access is not that unsurmountable an issue for the attacker for a couple of reasons. One is that the attacker might simply be able to train their own face recognition algorithm and now they have an algorithm in their possession that they can attack in a white box setting. And there are research results that indicate that if you have two algorithms that are trained about equivalently well to do the same task, then attacks that are successful against one algorithm are likely to transfer pretty well to the other algorithm, too. So this is a way in which, even though an adversary might not have complete access to the algorithm that they're actually attacking, they effectively train a substantive algorithm that they have full access to and then they create an attack that they use against the system that they don't have access to.

Sean Gallagher: 26:14 So the OpenAI Institute put out a deepfake text, a machine language tool that we did a test of a while back and it seems like that might be something that someone could use to, with the

right training, target specific marketplaces or rating systems, for example, generating fake Amazon reviews and things like that to post about particular products. Do you see that as being a possibility as far as somebody figuring out what the rules are on filtering reviews and writing an AI to generate positive or negative reviews for a specific product?

Lujo Bauer: 26:54 Sure. And I think there are two aspects to this. I think one is attackers who might try to evade an algorithm that's put into place to detect something like fake reviews and certainly it's possible that the machine learning could help them do that. The other aspect is that often we distinguish between benign behavior and malicious behavior is by deciding whether the behavior seems like it's reasonable or whether it comes from a person that we know. So for example, phishing emails are much more infective if they are tailored to our circumstances. If I get a phishing email that seems to be from somebody that I talked to today and it mentions in the subject line something that was relevant to my work day today, then I'm very likely to believe that email, at least for a long enough to open it, and maybe then, if the text inside it is not riddled with grammatical errors and doesn't sound totally different from the person in question, maybe I'm likely to click on whatever link is in that email.

Sean Gallagher: 27:55 Do you see any particular type of machine learning being any more vulnerable to this sort of an attack than any others? Most of the machine learning types we're talking about here are guided AI or they're just given training sets and it's unguided. Is there a difference between the two in terms of how vulnerable they are to this sort of attack?

Lujo Bauer: 28:17 Right. And just to be precise here, there's AI and there's machine learning. And when we talk about AI, this is really the collection of algorithms that make computers behave in ways that we normally ascribe to intelligence. The definition is kind of interesting in that it changes over time because we get used to what computers can do and we no longer ascribe certain behaviors to intelligence. So what we think of as AI today might not be the same as what we think of as AI tomorrow. Machine learning is a subset of that, and when we talk about machine learning, we normally talk about these more statistical algorithms that learn from data and so they are trained to do a particular task in response to some input. Now, the part of the AI that is not in machine learning, maybe often those algorithms are less susceptible to any of these attacks that we've been talking about in part because these algorithms that are in AI but not in machine learning, they are designed by humans and humans supply the intelligence behind the scenes, so to speak-

Sean Gallagher:	29:23	Right.
Lujo Bauer:	29:24	... versus these machine learning algorithms that we just train on data. They're the ones that we are much less likely to understand why they're doing whatever they're doing and it's much easier, perhaps, for their training data to not be complete enough for the behavior of the algorithm to always be reasonable.
Sean Gallagher:	29:45	So from the standpoint of where we are now and where we could be in a decade or so, is there anything that can be done in terms of how we think about applying these different technologies to make them more resistant to malicious or accidental misuse or manipulation?
Lujo Bauer:	30:06	I think so. I think there are quite a few things that we could be doing. I think one of the challenges is that we're facing this problem at a point when the pace of innovation is very high and the pace of going from the lab to the real world is very fast and so it's maybe a little bit more challenging now to put the brakes on than than it might've been 10 years ago, but other things that we might need to do is think carefully about how we are training these systems. Where's our training data coming from? Are we sure that the training data is accurate? Are we sure that it's representative of the domain that we're interested in? Are we sure that the attacker can't control some subset of the training data?
Lujo Bauer:	30:51	And then in building the algorithms themselves, maybe we need to be careful that we don't expect too much of them. Maybe we are willing to trade off some accuracy under benign circumstances or maybe better robustness under adversarial circumstances. Maybe when we are tempted to use one of these algorithms, maybe we use two different ones and trust their answer only if they agree. Various things like that. But certainly if we compare both the state of the art and the state of practice in engineering AI systems with the state of practice in engineering standard software systems, we are lacking in the AI space in a good understanding of metrics and processes that will improve the likelihood that somebody can deliver a safe product. So this is something that's still very much evolving and being worked on.
Lee Hutchinson:	31:48	It almost sounds like there's a whole new untapped area of industrial espionage or real espionage waiting to be exploited, of poisoning a competitor or an enemy's learning algorithms or poisoning the learning database that algorithms are trained on.

Lujo Bauer:	32:05	That very well could be the case. I would almost be surprised if something like that hasn't happened already.
Sean Gallagher:	32:13	So the Department of Defense DARPA has been trying to push forward this thing called explainable AI where the algorithms that are created are in some way human-readable so that you can understand what's going on inside of them. Is that something you see as being practicable with some of these applications, where you can actually look at an algorithm generated by a machine learning environment and say, "Okay, yeah, that's what it's doing and this is where the problem with it is."?
Lujo Bauer:	32:44	Yeah, I think it definitely has the potential to hold. Actually, I have a colleague, Matt Fredrikson, who's also a professor here who works on explainable AI and we've worked together a little bit in that we have given him data from our eyeglasses attacks to see if his algorithms can look at how the face recognition algorithm is working and try to devise an explanation for when it is saying this person is so and so, devise an explanation for why the algorithm is reaching that decision. In fact, I've seen examples from his research where his algorithm is able to detect that the face recognition algorithm is making the decision it's making because of the eyeglasses. And so if you are able to get this kind of information to, say, a TSA agent who is examining the decisions of an algorithm as it's recognizing people who are walking through an airport, then the agent might be able to say, "Oh, well, this particular identification was made on the basis of eyeglasses and therefore it's an untrustworthy identification. I'm not going to believe it."
Sean Gallagher:	33:59	So it sounds like you've got a lot of potential research targets to hit with your students going forward. Is there a particular area you're going after right now?
Lujo Bauer:	34:06	What we're trying to work on now is approaches for making these algorithms more robust. We have a bunch of stuff in the works, but it's still a little bit too early to tell what will pan out.
Lee Hutchinson:	34:20	Lujo, let me ask you, is there anything really that you think we haven't gotten to that you'd like to bring up in this context?
Lujo Bauer:	34:28	Thanks for asking. You know, maybe the only thing that I want to make sure I convey is that I don't want it to come across as being against AI or machine learning, because these are amazing technologies that really can bring about a huge amount of good. The way that I hope that my comments and my experience can be interpreted is in helping people recognize

that there are risks that we should be acknowledging. And this is the same as with any new technology, particularly the ones that seem like they solve many, many problems. There's usually some flies in the ointment that mean that we have to take additional steps before we can get the benefits without incurring too much risk.

Lee Hutchinson: 35:16 Okay. Well, Lujo, thank you very, very much for making the time. We really do appreciate it and we're glad to have spoken with you today.

Lujo Bauer: 35:23 Absolutely. Thanks very much for having me.

Sean Gallagher: 35:26 While there haven't been any known cases of adversarial AI attacks in a while, the security researchers are already looking into ways to defend against them. To get an idea of the work that's being done, we talked to Max Heinemeyer, director of threat hunting at Darktrace.

Sean Gallagher: 35:41 Now joining us is Max Heinemeyer, the director of threat hunting at Darktrace. Thanks for joining us for this podcast.

Max Heinemeyer: 35:48 Thank you very much.

Sean Gallagher: 35:50 So the topic at hand is adversarial AI and we've spoken with some experts in the field. Wanted to get your take on how Darktrace sees the development of adversarial AI as a threat vector for going after the security software that's out there. How does Darktrace see that as a growing threat, if at all, right now?

Max Heinemeyer: 36:14 That's a great question. Thank you so much for that. So when we talk about adversarial AI, I think of two things, actually. One, I think about [protecting 00:36:22] AI systems. So trying to exploit the new attack surface that AI systems provide. The other thing I'll think about when I hear the term adversarial AI is thinking about how AI is going to be used in cyber attacks. That sounds like a subtle difference, but it makes all the world's difference, actually, because one is attacking AI systems and trying to fool them, maybe facial recognition systems, maybe deepfakes, maybe cyber security solutions using AI. The other one, which we find much more interesting at Darktrace, actually, is how AI was used by the bad guys. By black hats, by attackers, by red teams right now or in the near future.

Max Heinemeyer: 37:01 Of course, it's incredibly important to look at what we call adversarial AI, which is trying to break machine learning

systems, to make sure our own machine learning system using unsupervised machine learning is bulletproof and as secure as possible from that perspective in a tech vector. But we also, because we live and breathe the threat landscape, we have over 3000 customers seeing live attacks every single day, see how the threat landscape changes, it's very important and dear to our heart to think about the next paradigm shift. And that is what we call census AI. That means using AI in cyber attacks.

Sean Gallagher: 37:34 So have you seen any indication that AI is being used or is that still something that you feel is [inaudible 00:37:42]?

Max Heinemeyer: 37:43 So this is a really good question. We don't have hard proof, but if you think about it, another objective of using AI for attacks, offensive AI, is to remove attribution and make detection harder. So by the very nature of this, it gets harder to detect these things or attribute them to actual AI. How do we know if [inaudible 00:38:01] tweet is created by human or by an AI if it's not distinguishable from a human because it's so perfect and blends in? However, we can see that it makes perfect sense for the attackers to do it because there's much more open source infrastructure available training. You can just go to any open university and start learning machine learning things or just use the results that have been pre-trained.

Max Heinemeyer: 38:23 Also, going on a slight tangent here, [inaudible 00:38:26] has just started talking about Operation Glowing Symphony. I don't know if you've come across this or read it, but it's all on the internet, and they disclosed how US Cyber Command has basically hacked ISIS in 2016 and they discussed their methods in doing so. It's a very fascinating topic, but the key takeaway here for our discussion is that before the US Cyber Command hit the big red button to start hacking ISIS, of course it is years of research, but they scripted everything to the last bit and byte, so they heavily relied on full automation as far as it goes. They didn't just have the operators hack into the ISIS networks, they had everything mapped out and scripted. It's all well-documented on the internet already, but being that, I don't want to say an APT group, but a very skilled hacking group with a lot of resources which some nations might define as an APT, US Cyber Command is scripting, automating everything they can. I don't see why they wouldn't use, or any other very advanced hacking entity, even better ways of automating things, which is using AI.

Sean Gallagher: 39:27 That's interesting. That gets into a little bit of the realm of some of the discussions we had around the Cyber Challenge and the idea of building game theory into AI models to sort of make

decisions about what types of techniques to use at any given time and when to deploy a certain set of scripts and when not to. I would imagine that given enough time and given enough monetary incentive to go after targets, that would be something that an attacker can learn relatively easily based on open source information.

Max Heinemeyer: 40:11 Absolutely. You touched upon another great example here. Everything we talked about so far, for us, for Darktrace, it's the short-term perspective on offensive AI. Things that are probably happening right now or will start happening in the near future. Another example here, since you mentioned the monetary incentive, is if we think to the ends of the attack life cycle, let's say data exfiltratation. And normally when you run an operation and you exfiltrated gigabytes of data as a hacking group, you have to sift through the data and see how can you monetize it, if you can use it to ransom your victim, if you can extort them, if you can upload data or sell it to a competitor.

Max Heinemeyer: 40:45 So to give you another very example of how AI can speed this up right now, I don't know if we are familiar, everybody on this podcast, with the hack that happened in Germany earlier this year in January, but basically a script kiddie, a German script kiddie hacked a lot of German politicians and artists and dropped their private data like private Dropbox data, private emails on the internet. And the interesting bit here is that A, that's what's possible to do by a script kiddie to hack hundreds of politicians, but more interestingly, the script kiddie pointed out in his parting messages things like, "Look at this politician. He's clearly got an [inaudible 00:41:20] adult pictures on his private Dropbox."

Max Heinemeyer: 41:23 And the interesting bit is that the script kiddie did manual labor, going through the data dump and sift out these images. What if they would have used, instead of manual work and manual analysis, pushed all these images through something like the already existing open source Yahoo NSFW, not safe for work, neural network, where you put something in and out comes adult images, violent content, and all the things we want to [inaudible 00:41:49] on social media so it doesn't get shown to users? So what I'm saying is weaponizing the existing research projects, but this time for monetizing the data dump.

Sean Gallagher: 41:58 So what can be done to counter this sort of an attack in terms of hardening of the defense on the defender's side and and is there a way to use artificial intelligence and machine learning to identify these types of attacks as they emerge?

Max Heinemeyer:	42:15	I would say yes. Of course, people still need to do the basics, need to have their protective skin, so to speak, their basic cyber hygiene. But we think it's ridiculous to think you can defend against the machine with a human. So how can people defend themselves if these attacks come around where you can't use signatures or a set of rules to detect them? They can just change on the fly and adapt to your environment. So while the basics are still important, I think we need to embrace new technology more. And, I mean, I'm biased. I work for an AI security vendor, obviously, but I see it every day. I see how we can catch ransomware that has never been seen before just using a neural activity behaviors and detecting anomalies. And it's super powerful, so it's right to use, correct to use, I think AI is one of the few things we can use in security to go forward if we anticipate the rise of offensive AI.
Sean Gallagher:	43:09	One of the biggest concerns I have about offensive AI from the standpoint of a defender is that, looking at the way cyber attacks in general have been going over the past few years as organizations tighten up their security around malware and ransom and things like that is that it's gotten more hands on keyboard, more what you call living off the land, and that's the sort of thing that if you have a backend that is automated using AI to deal with whatever the complexities of the environment you're going after are, could be extremely dangerous in terms of how quickly an adversary could use an automated tool based on AI to rapidly move around within the network and gain access to data.
Max Heinemeyer:	43:58	Exactly. The best attackers out there look like your admins. They looked like your developers, right? They blend in. And how do they blend in these days? Well, they sit there and listen. They hack you and then they wait and run screen grabs and understand how you work. Now, why should they do this? This could easily be done by an AI. The MITRE framework, I'm sure everybody here knows MITRE ATT&CK, the attack matrix?
Sean Gallagher:	44:21	Right. Right.
Max Heinemeyer:	44:22	But the MITRE organization does, also, a lot of other good research and they did some research into CALDERA, which is an autonomous decision agent that can make, based on uncertainty and imperfect information, decisions. For example, for lateral movement. It's jumped onto a computer, it listens for things, and it understands what the best pathway forward is to hack into another computer to maybe finally get the main admin cracked. I don't think it's as powerful as I make it sound, but that's the concept in the machine learning they do research

on. Again, the open source projects, all of these things are already there.

Max Heinemeyer: 44:55 We also see that there are some interesting plugins going around for things like the post-exploitation framework Empire Powershell where the initial infection takes place and then the implant doesn't ping back, doesn't [inaudible 00:45:10], but it sits there and learns the traffic to blend in, to configure the implant, the malware, to blend in with the network. So instead of saying statically, "Always connect back to my server on DigitalOcean to this .PHP website every 30 minutes," the implant is going to sit on the hacked computer, doesn't ping back, but listen and understand if you sent specific user rights normally. if you go to YouTube, if you use Dropbox, if you go to .JSP websites, .PHP websites, and then configure the payload so that it blends in perfectly with the environment. And like you said, the best attackers do this stuff manually, but if they start scaling up and automating this, they can increase their return and invest in their tech range massively.

Sean Gallagher: 45:52 One of the biggest security issues companies have is with the human beings and their behavior and what they do with data. You see a more interactive AI being used to, say, carry out more complex spear phishing attacks or going after a broader number of targets and having close to real-time interaction with human targets to get them to give up things like sending a wire transfer to an account or something like that?

Max Heinemeyer: 46:24 Yeah, you could imagine that, right? There's a lot of AI chatbot software out there. If you go to any website these days and you wait four minutes, that's going to be a chat window popping up and saying, "Hello, my name is Jack. Can I talk to you?" And it's never Jack in the first place. It's always an AI system that can answer some basic questions.

Sean Gallagher: 46:42 Right.

Max Heinemeyer: 46:42 And you could easily think how this could be used by the bad guys to initiate at scale again. Right? A human can do this, of course, but if you don't want to talk with one company, but 500 at the same time, how we can use these chatbots maliciously to target 500 companies or more at the same time on LinkedIn like, "Hey, I saw your message there. Can we have a chat?" "I don't know you. Who are you again?" "Oh, well, we talked around this topic last week at a conference." "Oh, okay. I see." And just have the chatbot do first 10% of the interaction until they divulge or give out their email address or maybe their phone number, and then go from there. So again, it's about

scaling up and alternating the tasks that are more cognitive-intense that normally are done by humans. So AI is going to try to replicate what humans do, basically.

Max Heinemeyer: 47:27 And just on this topic, I find it madness that most of the industry in security is trying with the same thing and the same approach we had for the last 30 years. Because, look around, [inaudible 00:47:39] good enough. Even really good companies with big budgets and great security teams are [inaudible 00:47:43], not meaning professionals at a specific company, can get attacked by a single person with mental health issues. So if that's one of the big cases out there, how can anybody else like the small bakery chain around the corner or the taxi company next door, even dare to think they can defend themselves? And then I look across the shop floor at big conferences and I see people talk about we need more signatures, we need more pen testing [inaudible 00:48:08], we need more security awareness training. Yeah right, because this has gotten us so far, right?

Max Heinemeyer: 48:12 So I think even talking about the weakest link, humans here, a lot of people think we need to train them, we need to make them aware. But let's be honest, there's always going to be a [inaudible 00:48:22] of people who don't want to be trained, who can't be trained, forget things. Even we security experts are susceptible to fall to phishing lures. So we need to shift our mindset here and think radically different. And this is what we tried to do with Darktrace. For example, when we think about spear phishing, I used to be an ESCO hacker and penetration testing lead for central Europe and I tested various companies and once I started testing the Darktrace email defenses, I'm not trying to toot my own horn here, but since then, even a [inaudible 00:48:54] is a vector. So what I'm trying to say is instead of trying to go in with the same approach that has failed for the last 30 years, try something new. Try to let the machine do the heavy lifting. I think you catch my drift here. Right?

Sean Gallagher: 49:05 So what can people do now to harden themselves, so to speak, against these types of threats? From a company perspective, from an individual perspective, what can people do today to make themselves ready for these things as they emerge?

Max Heinemeyer: 49:25 From a company perspective, embrace new tech. And no, I'm not saying throw the baby out with the bath water and forget all your old tech, but look at the new things. Look at [inaudible 00:49:33] companies and disruptive approaches because we enter the new era now and there are new things that are going to be game-changers like [inaudible 00:49:40] machine learning and AI. So for companies, take a look at new tech and what they

		can deliver. Test them, see if they actually do what they promise and run trials with them if you think they do what they promise.
Max Heinemeyer:	49:51	From an individual's perspective, there's a few quick and easy wins. Don't reuse your passwords. Or use something like Password Safe, if you know what that is. Go ask a friend who's more computer-savvy to help you install one. Use strong passwords, which could be as easy, length beats complexity, so it could be long but easy to remember. Just looking around my room and it could be something like, "Window built foam bloom 33," and I might misspell window so it's in no dictionary. So I've got a nice mnemonic phrase which I can easily remember and it's a strong, long password. And keep your software up to date and be aware. So I know I didn't address any of the AI threats we've been talking about, but we may not forget that there are many, many sites out there that don't use AI yet which a lot of people fall victim to. So if they follow these simple things, keep long strong passwords, don't use it twice, use multi-factor authentication if you can. Very popular these days. Incredibly important. And keep your software up to date and be aware, then you do better than 99% of most people out there and you will be a very hard target.
Sean Gallagher:	50:58	and considering how some of these threats are emerging, they may just as well use some of these low-hanging type attacks just in an automated session anyway, so defending against those sorts of things now can protect you from more advanced threats later.
Max Heinemeyer:	51:12	I couldn't have said it better.
Sean Gallagher:	51:14	Okay, great. Max, thank you very much for your time on this, and this has been really informative and I think it helps us flesh out this topic a bit more from the perspective of what our listeners have to worry about right now.
Max Heinemeyer:	51:31	Thank you so much for having me on the show. I really appreciate it.
Sean Gallagher:	51:34	That's it for this episode and for our mini-series of podcasts on AI. We hope you've come away with some things to think about and for continued coverage on these topics, be sure to keep reading ARS Technica. Until we do this again, take care and we'll see you in the comment section.

Lee Hutchinson:	51:48	Once again, this episode was sponsored by Darktrace, the world's leader in AI cyber-defense. With more than 3000 organizations relying on its AI technology around the globe, Darktrace is transforming security from the inside out. Start your 30 day free trial by visiting darktrace.com/trial.