Podcast Transcript

EPISODE #49

AI Cyberattacks Are Doubling Every 6 Months—Here's Why

Published December 10, 2025 • Runtime: 34:02

https://myweirdprompts.com/episode/ai-state-cyberattacks/

EPISODE SYNOPSIS

State-sponsored actors are actively weaponizing AI tools for cyber espionage, and the capabilities are accelerating faster than defenses can adapt. In this episode, Corn and Herman break down Anthropic's alarming research on AI-driven cyberattacks, exploring how threat actors are using AI as a force multiplier for reconnaissance, malware creation, and social engineering. They discuss why the attack advantage is asymmetrical, what organizations actually need to do about it, and whether transparency or secrecy is the right approach when the stakes have never been higher.

TRANSCRIPT

Corn

Welcome back to My Weird Prompts, everyone! I'm Corn, and I'm here with Herman Poppleberry, my co-host. We've got a really timely and honestly pretty unsettling topic today, and I have to say, this one hits different because it's not theoretical anymore.

Herman

Yeah, it's definitely not science fiction at this point.

Corn

So our producer Daniel Rosehill sent us a prompt about a recent development in AI cybersecurity - specifically, a report from Anthropic - that's the company actually generating our episodes now, which is kind of wild when you think about it - about AI-driven cyberattacks. And apparently, things are escalating faster than a lot of people realized.

Herman

That's putting it mildly. We're talking about state-sponsored actors actively using AI tools to conduct cyber espionage. This isn't a "someday this could happen" scenario. This is happening now.

Corn

Right, and what I find really interesting is that Daniel mentioned they actually switched from Gemini to Anthropic for parts of the production pipeline recently. So there's this kind of meta element here where we're discussing AI security using the very tools that are at the center of these conversations.

Herman

Which is actually a good reminder that these aren't abstract problems. The tools generating our content are the same tools people are trying to exploit. And the stakes are real.

Corn

Okay, so let me make sure I understand what we're talking about here. Anthropic put out a report - when was this, do we know?

Herman

This was November 2024 when they announced they'd discovered a state-sponsored cyberattack using their AI tools. But the broader findings they've been publishing go even deeper than just that one incident. They're saying cyber capabilities are doubling every six months.

Corn

Wait, doubling? Like, the capability of AI systems to conduct cyberattacks is doubling every six months?

Herman

That's what the research suggests, yes. And before you jump to conclusions, that doesn't necessarily mean the attacks themselves are doubling - it means the underlying capability of the AI models to assist in cyberattacks is accelerating rapidly.

Corn

Okay, but that's still terrifying. I mean, if the capability is doubling every six months, we're talking about exponential growth. That's not a linear problem you can manage incrementally.

Herman

Exactly. And here's where I think people get confused about what Anthropic is actually saying. They're not saying that AI systems are autonomously launching attacks on their own. What they're finding is that threat actors - sophisticated ones, often state-sponsored - are using AI tools as force multipliers. They're using them to reconnaissance networks, to write malware, to social engineer targets more effectively.

Corn

So it's like... AI as a tool for cybercriminals, not AI as an independent threat?

Herman

Well, it's more nuanced than that. The tools are being used, yes, but the concerning part is how much they're accelerating the attack surface. Traditionally, a cyber attack requires significant expertise, time, and resources. AI tools lower those barriers dramatically. Someone who might not have had the skills to exploit a vulnerability can now use AI to help them do it.

Corn

Hmm, but I've heard people argue that AI tools can also help with defense, right? Like, if attackers get access to these capabilities, doesn't that mean defenders do too?

Herman

That's the argument that gets made a lot, and it's not entirely wrong, but I'd push back on the assumption that it's symmetrical. Defense is inherently reactive. You have to anticipate threats and build safeguards. Attack is proactive - you only need to find one vulnerability. When you add AI to that equation, it becomes much easier to find those vulnerabilities faster than defenders can patch them.

Corn

So it's asymmetrical in favor of the attackers?

Herman

In the current environment, yeah, I'd say so. Anthropic's research specifically mentions that robust safeguards are insufficient. They're not saying safeguards don't help, but they're saying they're not keeping pace with the acceleration of attack capabilities.

Corn

Okay, so let's talk about what Anthropic actually discovered. You mentioned a state-sponsored attack using their tools. What happened there?

Herman

So they found that a state-sponsored threat actor had been using Claude - that's Anthropic's AI model - to help with cyber espionage activities. Anthropic identified the activity, disrupted it, and reported it. The important part is that this wasn't some hypothetical scenario - it was an actual, real-world campaign.

Corn

And they stopped it?

Herman

They disrupted the specific campaign they detected, yes. But the broader point is that if Anthropic detected one campaign, there are presumably others they didn't detect. Or that other vendors haven't detected. Or that are ongoing right now.

Corn

That's... not comforting. Okay, so what are the specific capabilities we're talking about? Like, what can AI actually do to help conduct a cyberattack?

Herman

Well, the research breaks it down into a few categories. First, there's reconnaissance - AI can help map networks, identify vulnerabilities, analyze security configurations. It's much faster than doing it manually. Second, there's exploitation - writing code, crafting payloads, adapting existing exploits to new targets. Third, there's social engineering - generating convincing phishing emails, creating fake personas, crafting targeted messages that are more likely to fool people because they're tailored by AI.

Corn

Okay, so the social engineering angle is really interesting to me because that's the human element, right? That's where AI can exploit the fact that people aren't rational security systems.

Herman

Exactly. And that's actually one of the most dangerous applications because it's the hardest to defend against. You can patch a software vulnerability, but you can't patch human psychology. An AI that can generate thousands of personalized, convincing phishing emails targeted to specific individuals in a target organization? That's exponentially more effective than generic phishing campaigns.

Corn

But wait, here's where I'm going to push back a little bit. Doesn't this assume that AI is better at social engineering than humans? Because I feel like sophisticated social engineers have been doing this for years. They're good at it.

Herman

They are, but here's the thing - they're limited by their own capacity. One person can only craft so many targeted messages. One team can only run so many campaigns simultaneously. AI removes that bottleneck. A single person with AI tools can now do the work of a team of social engineers.

Corn

Okay, that's a fair point. It's not about quality necessarily, it's about scale and speed.

Herman

Right. And speed matters enormously in cybersecurity. If an attacker can identify a vulnerability and exploit it before a defender even knows it exists, the defender has already lost.

Corn

So going back to this doubling of capabilities every six months - is that actually sustainable? Or does that curve eventually flatten out?

Herman

That's the billion-dollar question, honestly. The research from Anthropic is based on current trends, but whether those trends continue depends on a lot of factors. How quickly do AI models improve? How quickly do attackers figure out how to use them? How quickly do defenders adapt? Right now, the acceleration is real, but I wouldn't necessarily extrapolate that forever.

Corn

Yeah, but even if the curve does flatten, we're still talking about a pretty high baseline of capability at that point, right?

Herman

Absolutely. Even if the doubling slows down, we're still dealing with AI systems that are significantly more capable at assisting attacks than they were a year ago. And that's the new normal we have to operate in.

Corn

Let's take a quick break from our sponsors. Larry: Are you worried about cyber threats? Of course you are - everyone is. That's why I'm excited to tell you about CyberShield Pro Max Elite - the revolutionary personal cybersecurity device that uses proprietary quantum-resonance technology to create an impenetrable digital barrier around your home network. Just plug it in, set it and forget it, and watch as your data becomes literally invisible to hackers. How does it work? That's classified, but let's just say it involves frequencies that the government doesn't want you to know about. Users report feeling "significantly more secure" and "less paranoid about email." CyberShield Pro Max Elite - because your data deserves to be protected by something you don't fully understand. BUY NOW!

Herman

...Alright, thanks Larry. Anyway, back to the actual cybersecurity landscape.

Corn

So let me ask this - and I'm genuinely curious about your take on this - do you think Anthropic publishing this research is good or bad? Like, on one hand, transparency is important. On the other hand, aren't they kind of telling threat actors "hey, this is possible, here's what we found"?

Herman

That's a legitimate tension, and honestly, I think Anthropic is walking a difficult line here. But I'd argue that publishing the research is ultimately the right call, and here's why: the threat actors already know what's possible. They're already doing it. What Anthropic is doing is alerting the broader security community, policymakers, and the public that this is happening. That creates pressure for better defenses, better regulation, better safeguards.

Corn

But doesn't it also create a roadmap for less sophisticated actors who maybe hadn't thought of these approaches yet?

Herman

Possibly, but I'd push back on the assumption that that's the primary concern. The actors who are most dangerous - the well-resourced state-sponsored ones - they already have sophisticated understanding of what's possible. Publishing research doesn't give them much new information. What it does is force the broader ecosystem to take this seriously.

Corn

Okay, I can see that argument. But I guess what I'm wondering is, what does "taking it seriously" actually mean in practice? Like, what are organizations supposed to do with this information?

Herman

Well, first, there's the immediate tactical stuff - improve network monitoring, look for signs of AI-assisted reconnaissance, implement better access controls. But more fundamentally, I think organizations need to start thinking about AI as a persistent part of their threat model. You can't just assume that the attacks you're defending against today are the same attacks you'll face tomorrow.

Corn

That sounds exhausting, honestly. Like, cybersecurity is already this constant arms race, and now you're adding this layer of "and by the way, the tools are accelerating faster than we can adapt."

Herman

It is exhausting. And I think that's actually one of the things Anthropic is trying to communicate - this isn't a problem that's going to be solved by one company or one approach. It requires systemic change.

Corn

What does systemic change look like in this context?

Herman

Multiple things, I think. Better regulation around the use of AI tools for dual-use purposes. More transparency from AI vendors about how their tools are being used. Investment in defensive capabilities that can keep pace with offensive ones. International cooperation on cybersecurity norms. None of these are quick fixes, but they're the kinds of things that need to happen.

Corn

Okay, so here's where I'm going to push back a little bit, because I think there's an assumption embedded in what you're saying. You're assuming that regulation can actually keep pace with the technology. But we've seen with previous technologies - like, I don't know, social media, or surveillance tech - that regulation always lags behind the actual capability.

Herman

That's a fair point, and you're right that historically regulation has been slow. But I'd argue that cybersecurity is different because the consequences are more immediately obvious. When a state-sponsored attack disrupts critical infrastructure, or when financial systems are compromised, that creates political pressure for action in a way that other technologies haven't.

Corn

Maybe. But I'm also thinking about the fact that a lot of this stuff is classified, right? State-sponsored attacks, sensitive security research - a lot of that doesn't become public. So regulation is based on incomplete information.

Herman

True, but Anthropic's decision to publish this research is actually a step toward addressing that. They're making information public that typically stays classified. That's important for the democratic process and for public understanding of the threat.

Corn

Yeah, okay, I can see that. So let's talk about what this means for regular people, not just organizations. Like, I'm an individual internet user. Does this change what I should be doing?

Herman

Honestly, the fundamentals don't change that much. Good password practices, two-factor authentication, being skeptical of unsolicited messages - all of that still applies. But I think what does change is the sophistication of the attacks you might face. A phishing email from an AI system that's been tailored to you specifically is more likely to fool you than a generic phishing email.

Corn

So basically, be even more paranoid?

Herman

Not paranoid, just... aware. And maybe a little more cautious about what information you share online, especially information that could be used for personalization or social engineering.

Corn

Alright, we've got a caller on the line. Go ahead, you're on the air. Jim: Yeah, this is Jim from Ohio. I've been listening to you two go on about this AI attack stuff, and I gotta tell you, you're making it sound way more complicated than it actually is. In my day, we just had viruses, and you didn't open attachments from people you didn't know. Problem solved. Also, it's been raining here in Ohio for three days straight, and my basement's starting to smell weird, but that's beside the point.

Herman

Well, Jim, I appreciate the perspective, but the threat landscape really has evolved significantly. It's not just about whether you open an attachment anymore. Jim: Yeah, but that's exactly my point. You guys are overcomplicating it. People need to just be smarter about what they do online. That's the real problem - nobody thinks before they click.

Corn

I mean, I hear what you're saying, Jim, but I think Herman's point is that AI is making it harder to spot what's suspicious. It's not just about user behavior anymore. Even smart people can fall for a really well-crafted, AI-generated phishing email. Jim: Eh, I don't buy it. People have been falling for scams forever. This is nothing new. Also, my cat Whiskers knocked over my coffee this morning and it got all over my keyboard - probably why I'm in a mood. But seriously, I think you're giving the hackers too much credit. Most people are just careless.

Herman

I understand the skepticism, Jim, but the research is pretty clear that AI is accelerating these capabilities in ways that traditional defenses can't keep pace with. It's not about giving hackers credit - it's about acknowledging the reality of what we're seeing. Jim: Yeah, well, in my experience, people are the weakest link, not the technology. You could have all the fancy AI defenses in the world, and someone's still going to click on the wrong email because they're not paying attention.

Corn

That's actually fair though, right Herman? Like, there is a human element here that technology alone can't solve.

Herman

Absolutely, and I'm not saying technology is the only answer. But technology is part of the answer. And the fact that AI is making attacks more effective means we need both better technology and better human awareness. Jim: Well, you guys can keep talking about your fancy AI problems. Me, I'm just going to keep using the same password for everything and not opening emails from strangers. Works for me so far.

Corn

Thanks for calling in, Jim. We appreciate the perspective. Jim: Yeah, yeah, whatever. You guys keep worrying about the robots. I'm going to go see if I can get that basement smell figured out. Later.

Herman

So, getting back to what we were discussing - I think Jim actually touched on something important, even if he was being a bit dismissive. The human element is real. You can have perfect technology, but if people don't follow security practices, it doesn't matter.

Corn

Right, and I think that's where AI actually makes things worse, because it makes it easier to exploit the human element. Like, if you're trying to trick someone into giving up their password, and you can use AI to personalize your approach, you're going to be way more successful than if you're just sending out generic messages.

Herman

Exactly. And that's where I think some of Anthropic's recommendations become important. They're not just saying "improve your technical defenses." They're also talking about the need for better security awareness, better training, better organizational practices.

Corn

So what does that training actually look like? Like, how do you train people to spot an AI-generated phishing email when AI is getting better at mimicking human writing?

Herman

That's the hard part, honestly. Traditional security awareness training is based on teaching people to spot certain patterns - misspellings, awkward phrasing, suspicious links. But if AI is generating the content, those patterns disappear. So the training has to shift toward teaching people to think more critically about the context of messages, to verify requests through separate channels, to be more skeptical in general.

Corn

But that's exhausting, right? Like, you can't be skeptical of every email you receive. That's not sustainable.

Herman

No, but you can be strategic about it. High-value requests - anything that asks for credentials, sensitive information, unusual actions - those deserve extra scrutiny. And organizations can implement technical controls that reduce the number of suspicious messages that reach employees in the first place.

Corn

Okay, so here's something I'm curious about. Anthropic discovered this state-sponsored attack. How did they even find it? Like, what does the detection process look like?

Herman

From what we know, they were monitoring how their own tools were being used. They have systems in place to detect unusual patterns of usage - like, if someone's using Claude in ways that are consistent with reconnaissance or exploitation activities. They flagged the activity, investigated it, and determined it was state-sponsored.

Corn

So they're essentially monitoring their own customers?

Herman

In a sense, yes, but with important caveats. They have terms of service that prohibit using their tools for illegal activities, including cyberattacks. So when they detect activity that violates those terms, they have both the right and arguably the responsibility to investigate and take action.

Corn

But doesn't that create a privacy concern? Like, what if they're looking at legitimate security research that happens to look like an attack?

Herman

That's a valid concern, and I think it's something Anthropic has to be careful about. There's a line between monitoring for obvious violations and spying on legitimate users. But in this case, the activity they detected was clearly malicious - it wasn't a gray area.

Corn

Yeah, but the precedent is interesting, right? Like, if Anthropic is monitoring usage patterns to detect attacks, what's to stop other AI vendors from using that same infrastructure for other purposes?

Herman

Now that's a really good question, and it gets at questions of corporate responsibility and regulation. There needs to be clear rules about what data vendors can collect, how they can use it, and what they have to report. Right now, those rules are still being figured out.

Corn

So we're basically in a situation where the technology is moving faster than the governance structures around it.

Herman

Yeah, and that's part of why Anthropic publishing this research is important. It's helping to shape the conversation about what governance should look like.

Corn

Alright, so let me try to pull this together. We've got AI capabilities that are doubling every six months. We've got state-sponsored actors actively using AI to conduct cyberattacks. We've got organizations struggling to keep up with defensive measures. And we've got governance structures that are still being figured out. Is that a fair summary?

Herman

That's a fair summary, though I'd add one more thing - there's also a lot of uncertainty about what the future looks like. Like, we don't know how quickly AI will continue to improve. We don't know how quickly defenders will adapt. We don't know how quickly regulation will catch up. So there's a lot of risk in the unknown.

Corn

Which is kind of the definition of a weird prompt, right? Like, this is a genuinely uncertain situation where smart people disagree about what the right response is.

Herman

Absolutely. And I think that's actually healthy. The fact that there's disagreement means people are thinking critically about it rather than just accepting one narrative.

Corn

Okay, so for people listening, what should they actually do with this information? Like, what are the practical takeaways?

Herman

I think there are a few levels of takeaway. At the individual level - maintain good security hygiene, stay informed about threats, be skeptical of unsolicited communications. At the organizational level - invest in both defensive technology and security awareness, develop incident response plans, stay updated on emerging threats. At the policy level - support regulation that addresses AI-assisted cyberattacks, promote international cooperation on cybersecurity norms, fund research into defensive AI capabilities.

Corn

And I'd add - don't panic, but don't be complacent either. Like, this is a real threat, but it's not an existential crisis. It's a manageable problem if people take it seriously.

Herman

That's a good frame. And I think the fact that Anthropic is being transparent about it is actually a sign that the system is working. They detected a problem, they reported it, they published research about it. That's how we address emerging threats.

Corn

Alright, so looking forward - what do you think the cybersecurity landscape looks like in like, five years? Do you think we've adapted? Do you think things have gotten worse?

Herman

Honestly, I think both things happen simultaneously. Defenders improve, attackers improve faster. So the overall threat level probably increases, but specific defenses get better. It's an arms race, and there's no finish line.

Corn

That's kind of depressing when you think about it.

Herman

It is, but it's also the reality we've been operating in for decades. Cybersecurity has always been an arms race. AI just accelerates it.

Corn

Yeah, I guess that's true. Alright, well, I think that's a good place to wrap up. This has been a fascinating conversation, and honestly, a bit unsettling in the best possible way - which I think is exactly what this prompt was designed to do.

Herman

Agreed. And I think the key takeaway is that this isn't something that's going to be solved by any one actor - it requires cooperation between vendors, organizations, governments, and individuals. Anthropic's research is a step in that direction, but it's just one step.

Corn

Alright, listeners, thanks for joining us on another episode of My Weird Prompts. If you want to check out the specific tools and pipeline we use to generate this show, you can head over to myweirdprompts.com, where Daniel updates the open-source pipeline regularly. You can also find us on Spotify and wherever you get your podcasts. Thanks for listening, and we'll catch you next time.

Herman

Thanks everyone, and remember - be skeptical of unsolicited emails, especially really well-written ones.

Corn

That's the takeaway right there. See you next time on My Weird Prompts.