SDS PODCAST

EPISODE 990: INSIDE MYTHOS: ANTHROPIC'S LOCKED-DOWN FRONTIER MODEL

Jon Krohn:	00:00	This is episode number 990 on Methos. Welcome back to the Super Data Science podcast. I'm your host, John Crone. Today's topic is Anthropic's Methos, a Frontier AI model so capable at finding software vulnerabilities that Anthropic has decided not to release it to the general public. We'll talk about all the reasons why they might have done that later on in the episode, but that is their main claim. Of course, it's been a few weeks since Methos came out, but now that the dust is settled, we can do some detailed reporting on it. And I also have lots of tips at the end of the episode on what you can do to avoid having lots of security vulnerabilities in all of that code that you're pushing, especially if it's code that you're pushing using a code gen tool like Codex from OpenAI or Claude Code from Anthropic themselves.
	00:56	Anyway, let's cast our minds back now to 2019. OpenAI had just finished training a then new large language model called GPT-2, and in a move that drew a fair amount of mockery at the time, the lab declared that it was too dangerous to release. The research director at OpenAI back then was Dario Amade, and Dario insisted that the world needed time to prepare. GPT2 ended up getting released later that year anyway. In a sequence, a far more powerful models has of course been deployed since and Armageddon has not been unleashed. Now, seven years on, Dario, these days, the CEO of OpenAI's biggest rival, Anthropic, is sounding the alarm again. On April 7th of this year, he declared that the new addition to Anthropic's Claude family code named Methos is too powerful to be made widely available at this time. And this time, having looked at the technical evidence, well, maybe he's right.
	01:55	Let's start with what Methos actually is because the technical story here is striking. Claude Methos preview to provide its full name is a general purpose frontier model, not a specialized cybersecurity tool. Anthropic is explicit

that they did not train it to find software vulnerabilities. The model's hacking abilities emerged as a downstream consequence of broad improvements in code understanding, reasoning, and agentic autonomy. On SWE bench, for example, a popular benchmark of real world software engineering problems, the top performing publicly available models are pushing up against an 80% accuracy. According to Anthropic, Mythos Preview hits 94%. That's a big jump from 80. Essentially, half of the problems that couldn't be solved previously are now solved. It's on the cybersecurity numbers, however, where things get really spicy. Here's a concrete comparison. Anthropic's previous frontier model Opus 4.6 was already capable. I was recently at a conference at Tulane University in New Orleans where I met John Dickerson, the CEO of Mozilla.ai, with Motsilla, of course, being the makers of the open source Firefox browser that is famously secure.

03:03 When researchers ran Opus 4.6 against a set of 147 Firefox vulnerabilities, it produced working Shell exploits two times across several hundred attempts. They then ran the same test with Methos preview. Methos succeeded 181 times. That's nearly a 100x increase relative to OPUS 4.6. As another cybersecurity example, taking Anthropic's internal benchmark of around 7,000 entry points across open source repositories from the OSS Fuzz Corpus. Sonnet 4.6 and Opus 4.6 each reached what's called tier five, meaning complete control flow hijack, wherein the attacker takes over what code the program executes next. So Sonnet 4.6 and Opus 4.6 reached that really dangerous tier exactly once. Methos achieved that same tier five kind of attack on 10 separate fully patched targets. That is a 10X difference. These 100X and 10X deltas on cybersecurity vulnerability discovery, that's consistent with a generational leap, not an incremental improvement. In the few weeks before the announcement of Methos, Anthropic ran methos against real world critical software using a remarkably simple agentic

scaffold, an isolated container with the source code, Claude Code wrapping the model, and a prompt instructing it to find a security vulnerability.

04:33 The model reads the code, hypothesizes vulnerabilities, runs the project to confirm them, attaches debuggers as needed, and produces either a negative result or a working proof of concept. With this setup, Methos identified thousands of zero day vulnerabilities across every major operating system and every major web browser. Modzilla, for example, working with an early version of Methos on Firefox patched 271 vulnerabilities in a single software release. For context, all of 2025. Saw Modsilla addressed just 73 high severity Firefox vulnerabilities. So that's about four times the annual figure in a single AI-driven sweep. To go one level deeper on confidence in these findings of 198 findings that Anthropics had manually reviewed by professional security contractors, 89% of those received the exact same severity rating from the contractors as the model had self-assigned, and 98% were within one severity level. This means that this isn't a case of a model hallucinating bugs.

05:37 The contractors agreed with Methos' risk assessments at near human expert reliability. So that's the technical picture. Now, let's talk about the rollout because the marketing strategy is just as interesting as the model. Rather than launching Methos publicly, Anthropic launched something called Project GlassWing, an industry consortium with launch partners that include Apple, Google, Microsoft, AWS, NVIDIA, JP Morgan Chase, the Linux Foundation, Cisco, CrowdStrike, Broadcom, and Palo Alto Networks. So yeah, a lot of big names in this big consortium. The idea is that these defenders get to use Mythos Preview to scan and harden their code bases before the capability proliferates. Anthropic has committed up to a hundred million dollars in usage credits across the program, plus $4 million in

direct donations to open source security organizations. But there's a commercial logic here too, isn't there? Mythos preview is priced at $25 per million input tokens and $125 per million output tokens.

06:34 That's roughly five times the price of Opus 4.6, suggesting Mythos is genuinely far more compute hungry and Anthropic has been rationing capacity on Claude already. Exclusivity also makes Mythos harder to distill. That is harder for rival labs. Anthropic has previously named three Chinese labs in particular to use Methos's outputs to train cheaper imitator models. And keeping the model gated nudges the enterprise customers toward anthropic native tooling like Claude code instead of model agnostic products. So while the safety framing is real and well justified, this rollout is also great business and the media sensation around a model too powerful to be released, made for great marketing indeed. OpenAI, not to miss out on a media splash followed suit days later with a similar too powerful to be released announcement regarding a version of GPT 5.4. Now, not everyone is convinced this is a watershed moment, mind you.

07:34 Bruce Schneider, one of the most prominent figures in computer security, he's been a public voice in the field for roughly 30 years and is widely respected, has noted that the security firm aisle was reportedly able to replicate some of Anthropic's findings using older, cheaper public models. The point being, finding vulnerabilities and weaponizing them are different things and current AI may help defenders more than attackers for now. There's also a geopolitical dimension to all this. Project Glass Wing could effectively neutralize zero-day vulnerabilities that the US government has historically hoarded for offensive cyber operations. Defense Secretary Pete Hagseth, who labeled Anthropic a supply chain risk earlier this year, is unlikely to be thrilled by that. And open-way labs, particularly those in China, will likely produce comparable capabilities within months. A recent EPOC AI

analysis estimates the average capability lag between proprietary and open-weight frontier models at just three months, though it can extend to between five and 22 months on certain benchmarks.

08:31 So the defender headstart that Glass Wing buys is real, but it's not unlimited. So here's what I take away as an AI practitioner from all this. We are entering an era where automated vulnerability discovery is no longer bottlenecked by scarce human expertise, and where the gap between finding a bug and writing a working exploit has collapsed from months to minutes. If you build, deploy, or maintain any kind of software at scale, your patching pipeline now needs to operate at machine speed. The defenders who modernize their tooling first, automated scanning, AI assisted code review, runtime behavioral enforcement, those are the ones who'll come out ahead. And if you're working on AI capabilities themselves, Methos is a remarkable illustration of how dangerous capabilities can emerge as side effects of general improvements rather than from targeted training. Now, if you're one of the many listeners generating tons of code through GenAI tools like ClaudeCode or Codex, you're going to want to be extra careful because you probably are creating applications with way more vulnerabilities than time and expert hardened software like Mozilla's Firefox.

09:37 We've talked about tools like CodeRapit on the show before back in episode number 927 for automating the review of pull requests. You can also use CloudCode or Codex themselves to run a dedicated security step after generating your code for you. And there is a big industry of AI native security tools being specifically built for the AI native era, including Socket, and/or Labs and Semgrep. I've got links to all of those for you in the show notes. All right, that's the end of today's episode. If you enjoyed it or know someone who might, consider sharing this episode with them. Leave a review of the show on your favorite

podcasting platform or YouTube. Tag me in a LinkedIn post with your thoughts. And if you aren't already, be sure to subscribe to the show. Most importantly, however, we hope you'll just keep on listening. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.