SDS PODCAST

EPISODE 1002: FABLE 5: THE FULL STORY FROM CAPABILITIES TO DRAMA

Jon Krohn: 00:00 This is episode number 1002 on Anthropics Babel five. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. Today's topic is Claude Fable five, a model that Anthropic released last week and that's notable not just for being powerful, but for being a brand new kind of release for the company. And as if that weren't enough of a story on its own, the whole thing took a dramatic turn when the US government forced Anthropic to pull the model entirely only three days after it launched. So this episode is really two stories braided together, what Fable five is and why it's already been yanked off the shelves. Let me start by explaining what the model is because you need that to understand why its disappearance is such a big deal. I should also state here at the outset that despite Anthropic being a sponsor of this podcast, including of this very episode, Anthropic doesn't have any editorial influence on what topics we cover on the podcast or how we cover them.

00:58 Everything in this episode is my genuine research backed perspective. Okay. That aside, here we go into the meat of the episode. For a while now Anthropic has organized its models into named tiers. The lowest tier was Haiku at the small, fast and cheap end, then saw it in the capable middle and Opus was at the powerful expensive top. But earlier this year, the company introduced a tier that sits above Opus, which it calls Methos class. The first Methos Class model called Claude Methos Preview came out back in April and as I detailed in a dedicated episode of this podcast, number 990, Methos wasn't something you or I could simply go and use online. Anthropic deemed it too dangerous for general release and instead handed methos to a small group of cyber defenders and critical infrastructure providers through a program called Project Glass Wing. The concern specifically was that a model this good at finding and exploiting software vulnerabilities could do enormous damage in the wrong hands.

02:00 So that's the backdrop. Methos class capabilities have existed for a couple of months, but they were locked away. What changed last week is that Anthropic figured out how to put a version of that technology in front of the general public and that version is Fable five. The nomenclature here is important and helps us understand both models. Fable five and its lockdown sibling Methos five are the same underlying model. The only difference is the safeguards. Anthropic even chose the name deliberately. Fable comes from the Latin fabula, meaning that which is told, which is closely related to the Greek word methos. Same route, same model, different guardrails. Methos five goes to the trusted glass wing partners with the cyber safeguards lifted. Fable five goes to everyone else with those safeguards firmly switched on. Now, let's talk about what Fable Five can actually do because the headline is that it's a big step up.

02:57 Anthropic says it's state-of-the-art on nearly every benchmark of AI capability they tested with especially strong results in software engineering, knowledge, work, vision, and scientific research. And here's the detail I find most interesting. The longer and more complex the task, the larger Fables lead over Anthropic's other models becomes. This isn't a model that's marginally better at quick answers. It's a model built to stay coherent across long multi-step jobs. A few concrete examples bring this to life. On the software engineering side, Stripe tested Fable five on a 50 million line Ruby code base and had to perform a code base wide migration in a single day. Work that the company estimates would've taken a full team of engineers more than two months by hand. On vision, earlier Claude models needed elaborate helper tools to play the video game Pokemon Fire Red and they still struggled. Fable F beat the game using nothing but raw screenshots and a minimal harness.

03:55 And on memory, when Anthropic let the model play the deck building game slay the spire with access to

persistent file-based notes, that model boosted its performance three times more than it boosted the previous Opus model. The model is in effect getting better at taking notes for itself and acting on them later. The price reflects the capability tier. Fable five costs $10 per million input tokens and $50 per million output tokens. That's roughly twice the cost of Anthropic's Opus 4.8 model so it isn't cheap, but it is also less than half the price of the original Methos preview. So while we're getting more capability, the cost per capability trend is still moving in the right direction following that expectation that we can expect any capabilities we have today to cost about 1% of what they cost today two years from now. So you can expect a 99% cost reduction on these state-of-the-art capabilities by two years from now, which is, yeah, that's the trend we've been seeing for years and I expect to continue for years in the future.

04:58 All right, that's capabilities and price. Let's move on now to the safeguards. This is the part that makes Fable5 unusual and it's worth dwelling on because it's where Anthropic safety posture becomes very visible to the everyday user. Fable five ships with a set of separate AI systems called classifiers that watch for queries in three sensitive areas, cybersecurity, biology and chemistry, and something called distillation, which is when someone tries to extract a model's capabilities to train a competing model. Anthropic has been vocal about specific Chinese labs, for example, about doing this kind of distillation. Anyway, when a classifier flags a request in one of these areas, Fable5 doesn't refuse outright. Instead, the request quietly falls back to Opus 4.8, which is still a very capable model and the user is told that this has happened. These safeguards are, in my experience, extremely stringent. For example, I tried to get Fable5 to proofread the script for this very episode and instead Anthropic defaulted to Opus 4.8 for the task.

05:59 Why exactly? Claude doesn't say, but perhaps me simply mentioning the word cybersecurity in my script flagged an aggressive tripwire. Anthropic themselves have been candid that they tune their classifiers conservatively, which is a polite way of saying that they err on the side of caution and will sometimes flag perfectly harmless requests. By their own figures, the safeguards trigger in fewer than 5% of sessions, which means more than 95% of the time you're getting the full unfiltered fable experience. But that remaining few percent has already produced some grumbling with early users reporting that innocuous prompts occasionally get bounced to the weaker model, just like I complained about a minute ago. Anthropic says reducing those false positives is a priority, that's good news, and that they'll refine the classifiers after launch. Now, why go through all this trouble? Well, it's because the same capabilities that make Methos class models so valuable are so- called dual use.

06:58 A query that helps a security professional, a cybersecurity professional hardened a system could in different hands help an attacker break into one. A biology query that advances gene therapy research could with malicious intent point towards something far more dangerous like developing a bioweapon. Anthropics bet is that with strong enough guardrails, they can deliver the benefits broadly while keeping the worst risks contained. And they ran an external bug bounty program with over a thousand hours of testing that they report turned up no universal way around the safeguards. Others claimed otherwise, however, so let's dig into that next. Fable F arrived just days after Anthropic publicly urged the major AI labs to agree on what it described as a coordinated brake pedal for frontier development, warning that systems are advancing fast enough that they may soon be capable of improving themselves with little human input. Such recursive self-improvement is what I'm expecting episode number 1004 of this podcast to be on in a week's time.

07:59 So sit tight for that. But Anthropic releasing their most powerful public model in the same week that they give such a warning on recursive self-improvement is a striking juxtaposition and it captures the tension Anthropic is openly trying to navigate, ship the capability but ship it carefully. And that tension is exactly what brings us to the wild news from last Friday because it turns out Anthropic's careful safety first launch didn't save the model at all. On Friday evening, US time Anthropic received a directive from the federal US government ordering it to immediately shut off worldwide access to both Fable five and Methos five, citing national securities concerns. The company confirmed it received the order at 5:21 PM Eastern on Friday the 12th and by a couple of hours later, users who'd been happily working with Fable found it simply gone, replaced by an error message in the chat.

08:54 Three days. That's how long the most capable model ever offered to the public lasted before it was pulled. Here's the mechanism because it's a strange one. The directive was technically an export control action. Commerce Secretary Howard Lutnick sent a letter to Anthropic CEO, Dario Amade, stating that Fable five and Methos five would be subject to export controls, meaning they couldn't be accessed by any foreign national, whether outside the US or inside it. Now, you might think Anthropic could simply block foreign users and keep the model running for everyone else, but the order reached so broadly covering foreign nationals everywhere, including Anthropic's own foreign born employees. So if you were born in Canada and then started working for Anthropic in California but hadn't become a US citizen, that order would block you from being able to access the system. So the company decided the only practical way to comply was to switch the models off for absolutely everyone worldwide.

09:54 So an export control order nominally aimed at foreign access ended up taking the model away from domestic

users too. Importantly, every other Claude model is unaffected, this is specifically about the Methos class tier. Why did the government do this? This is where it gets contentious and where you can hear Anthropic's frustration in their public response. The company's understanding is that the underlying concern is a claimed jailbreak of Fable F, a method that another company reportedly demonstrated for getting around the model's cyber safeguards and this alarmed the government. The Wall Street Journal published that the reporting company was Amazon and that it was Amazon CEO Andy Jassy himself that relayed word of the Fable five jail break to the White House. Anthropic in response strongly disputes how serious that finding actually is. They described the technique as a narrow non-universal jailbreak that essentially amounts to asking the model to read a specific code base and fix the software flawed fines.

10:53 And they point out that the vulnerabilities surface this way were previously known minor ones that other public models, including OpenAI's GPT 5.5 can find just as easily and that legitimate security professionals turn up every single day. In other words, Anthropic's argument is that the capability being restricted isn't unique to Fable and isn't the doomsday scenario it's being treated as. There's an irony here that's worth sitting with for a moment. Just days before all this, Anthropic was the company publicly warning that Frontier AI is getting dangerously powerful and urging the industry to coordinate on slowing down. Some critics had even accused them of over-hyping the danger of their own models as a kind of marketing and then the government took those very safety warnings in effect at their word and pulled the plug arguably faster and harder than Anthropic might have ever wanted. Their cautious messaging may have, in a sense, backfired.

11:46 It's also not the first clash between this company in Washington. Earlier this year, the Department of Defense

labeled Anthropic a supply chain risk, a designation usually reserved for foreign adversaries and a lawsuit over that is still active. There's reporting that the US administration had previously tried and failed to stop the fable launch altogether so the export control route may have finally been a lever that worked. The timing could be considered awkward for Anthropic on the business side because as I discussed on CBS News last week, Anthropic recently filed confidentially for an IPO, reportedly disclosing a revenue run rate of about $47 billion in evaluation near a trillion dollars, which is something like an 80X revenue increase relative to the previous year and yeah, also brings a valuation potentially ahead of where OpenAI was, but having your flagship model pulled by your own government in the run up to going public is not perhaps the headline any company wants.

12:50 Anthropic has complied with the order while making it very clear it believes the government got this one wrong and there's already legal wrangling underway with the judge having allowed certain agencies to keep using Anthropic's tools while proceedings continue. So this is far from settled, at least at the time of me recording this. There's also that old adage that all news is good news where perhaps there is an upside to this ban. Anthropic's models being so powerful that the government bans them. This seems to imply that Anthropic is the lab that's actually at the frontier while their chief rivals OpenAI and Google are racing to catch up. Maybe that's not bad for Anthropic's brand or valuation after all. I guess we'll see. Now, where does this leave you, dear listener, as a practical matter. At the time of recording this episode, Fable Five and Methos Five are off the table entirely for everyone.

13:42 If you'd started building on Fable through the Claud API or on a subscription plan, your sessions will now fall back to your default model or to Opus 4.8, which as I mentioned earlier, is itself a very strong model and was

the most capable public option right up until Fable launched. When Fable comes back and in what form now depends on how this fight between Anthropic and the government plays out. So that's the saga of Fable F. The first time Methos class capabilities were made available to the general public wrapped in a deeply cautious, a deliberately cautious set of safeguards priced as a genuine premium tier and then pulled by the government just three days after launch in one of the fastest reversals the AI industry has ever seen. I've personally not lived through anything like it. It's a remarkable case study in just how entangled frontier AI, national security and commercial pressure have become and it's a story that clearly isn't over.

14:37 We'll be watching closely to see whether Fable returns and what conditions come attached if it does. I, for one, can't wait to get back to using its capabilities. I was having a blast. Well, back to Opus for me. Keep building anyway, listener, and keep watching this space for updates. 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, I just hope you'll kep 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.