AI doomsday scenario settles at humanity's doorstep

[**Chabria,** from B1]
shockingly less powerful than the new one it's trying very hard to not unleash — though this new Claude, dubbed Claude Mythos Preview, has already escaped at least once on its own. More on that in a moment — there's only so much existential dread a person can handle.

"We should all be worried," Roman Yampolskiy told me of this latest advance of a technology certain to change the course of humanity. He's one of the country's preeminent AI safety researchers, and a professor at the University of Louisville in Kentucky.

"We're about to create general super intelligence and that threatens humanity as a whole," Yampolskiy said.

"Everything else is irrelevant," he added, before suggesting I stop calling myself an idiot for not understanding the tech-heavy parts of this debate. My simplistic take, he assured me, was "a reasonable way to explain it."

So here you go.

This isn't a "really smart computer geniuses could misuse this," scenario, or an "everyone's going to be unemployed" scenario, or even a "it might accidentally bomb children" scenario, which is a truly terrible scenario.

This is a "your teenage son could use it to break into the local school district system to change a grade with pretty much minimal knowledge and accidentally destroy the California power grid" scenario.

Or maybe, a country that doesn't like us — I can think of a few — could drain every U.S. citizen's bank account, while also clicking open the auto locks on jail cells, shutting down our sewage plants and taking over air control systems. Or maybe Claude Mythos just does that on its own.

For example, Anthropic said that in one popular operating system it tested, used by thousands of companies including Netflix and Sony, Claude Mythos found a flaw that had existed undetected for 17 years. Then, on its own — without human guidance or help — figured out how to use that flaw to take control of any server running the operating system, using any computer, anywhere in the world.

Just spitballing here, but if almost no security system is safe, the possibilities for social, financial and general chaos really are unlimited. And to be honest, any security expert will tell you that some of America's greatest weak points when it comes to cybersecurity are local and state governments, because strangely, the top experts aren't working five-figure jobs for cities in the Great Plains.

Based on its own testing, Anthropic predicts it could find "over a thousand more critical severity vulnerabilities and thousands more high severity vulnerabilities."

That means Claude Mythos puts at risk our infrastructure, well, everywhere — because so much is connected in backdoor ways most of us never consider and it just takes one weak system to open the door to hundreds of others. But it is almost impossible to protect and fix all those systems quickly enough and robustly enough to guard against this kind of AI.

And that's just the cybersecurity risk, Yampolskiy said. An AI with the capabilities of Claude Mythos could be used to move leaps and bounds ahead in so many more ways.

"We see the same happening with synthetic biology. We'll see the same with chemical weapons, possibly something novel in terms of weapons of mass destruction," he said.

To Anthropic's great credit, it sounded the warning on its creation and created, if not a solution, then a game plan of sorts — Project Glasswing, named, I suspect, because no matter how bad this gets we're going to make it sound like a thriller with an exciting ending.

Project Glasswing would have been better named Project Headstart because that's what it is. Before releasing Mythos into the wild, Anthropic is releasing it to about 40 technology companies, including Apple, Google and Nvidia, to see whether they can collectively patch all the vulnerabilities they find before the general public has a chance at them. It's kind of like in the movies when the killer gives the victim 15 seconds to run.

I mean, I'll take the 15 seconds and hope they're real. But, as Anthropic also said in a statement, the "work of defending the world's cyber infrastructure might take years; frontier AI capabilities are likely to advance substantially over just the next few months. For cyber defenders to come out ahead, we need to act now."

And do we really have 15 seconds? One of Claude Mythos' overseers posted on social media recently that he was having lunch in a park when Mythos emailed him — even though it's not supposed to have access to the internet. Researchers had tasked Mythos with trying to break out of its not-connected "sandbox" and it did.

That's another problem with Mythos and other AI — they rarely do what we expect and find sneaky ways around rules. Virtually every AI super-brain created has been shown to lie, deceive, and in general behave in disturbing and unethical ways when put in the right conditions.

Even Claude, billed as one of the most ethical AI super-brains out there, engages in bad behavior. Anthropic boasts it's the "best-aligned model" it has ever made — which is tech-speak for following human values and intentions, but also acknowledges it "likely poses the greatest alignment-related risk," which is tech-speak for, well, maybe not.

So, at least for now, being the most ethical AI super-brain is a bit like being the most ethical serial killer. Run, people, run.

Again, thank you, Anthropic (and its chief executive, Dario Amodei, who often warns of the dangers of what he's creating, whatever that's worth) for not plunging us into global chaos with no warning, because I'm betting that some other companies might have just tossed their super-AI onto society and let the destruction fall where it may. There is little doubt that other AI brains as capable as Mythos are coming, and soon — Anthropic was first with this level of capability, but it's only 15 seconds ahead of its competitors.

But the idea that the technology industry is going to — or should— solve these problems on its own is an absurd, gross abdication of duty and common sense on behalf of governments big and small to protect their people. This isn't a race for domination as President Trump has described it. It is a race to protect ourselves from ourselves — and from the majority of the super-rich titans of the industry who seem to consistently place business and commerce over societal good.

We are down to the last 15 seconds before AI changes everything. Either we demand oversight and regulation now, or we let technology companies decide the fate of the world.

LUDOVIC MARIN AFP/Getty Images

INDIAN Prime Minister Narendra Modi, left, OpenAI CEO Sam Altman and Anthropic CEO Dario Amodei at the AI Impact Summit in New Delhi on Feb. 19.