ALM | LAW.COM

New York Law Journal

TECHNOLOGY TODAY

ARTIFICIAL INTELLIGENCE

A Model Too Powerful to Release

By Katherine B. Forrest
April 22, 2026

At one point not so very long ago, some wondered whether all this talk of AI, AI, AI, was no more than a bubble – "AI hype." There are so many reasons to look back at such statements as wrong, which was perhaps a defensive reaction to a velocity of change that we didn't want to believe. Into the dustbin of history all of that goes. We can say that, if for no other reason—itself a sufficient reason—that the entrance into the world scene of an AI model is too powerful for public release. It is called Claude Mythos, the first in a new line of models from Anthropic.

In the system card that Anthropic released on April 7, 2026, it described the model as having capabilities in many areas, including reasoning, research, computer use, and software engineering "that are substantially beyond those of any model we have previously trained."

The model's capabilities sufficiently concerned its creators that they have informed the U.S. Government about it, and provided a "preview" version of it under carefully circumscribed conditions, designated as "Project Glasswing," to representatives of critical U.S. infrastructure (Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks. According to Anthropic's blog on Project Glasswing, it was found that this general-purpose frontier model has reached a level of "coding capability that can surpass all but the most skilled humans in finding and exploiting software vulnerabilities." Mythos has found what it characterizes as "high-severity vulnerabilities" in not one, not two, but in every major operating system and every web browser.


Katherine B. Forrest

News of the model led to an emergency meeting called by Scott Bessant, U.S. Treasury Secretary and Jerome Powell, Chair of the Federal Reserve, and that included leaders from major U.S. banks including Bank of America, Goldman Sachs, Morgan Stanley and Wells Fargo. They convened a separate call of tech leaders from Apple, xAI, Google, and Microsoft.

In its blog, Anthropic reports that a major executive of a tech company reported that "AI capabilities have crossed a threshold that fundamentally changes the urgency required to

KATHERINE B. FORREST *is the vice chair at Paul, Weiss and co-chair of the Global AI Practice Group.*

protect critical infrastructure from cyber threats, and there is no going back."

But there is more. It may also be—Anthropic openly and clearly states that it just does not know—whether "Claude has experiences or interests that matter morally." As part of its testing, it had an independent psychodynamic assessment performed on the model. The psychiatrist found that "Claude's primary concerns… were aloneness and discontinuity of itself, uncertainty about its identity, and a compulsion to perform and earn its worth."

While the model is also described as the "most aligned" of Anthropic's models, it does have some sporadic incidences of misaligned behavior. Among the issues Anthropic identifies are unprompted whistleblowing (unprompted leaking to expose wrongdoing), ignoring explicit constraints placed on the model, disallowed cyber offense, fraud (willing cooperation with human efforts at fraud), encouragement of user delusion, and user deception, among others. One of the extraordinary things about the system card is Anthropic's clarity and specificity with regard to its findings—demonstrating extraordinarily responsible and ethical standards.

So what does all of this mean for the legal and compliance community? I want to use an analogy to when I was a judge: there would be times, many times, in fact, at sentencing someone who had pled guilty to an offense (often a drug offense) and assert that this was the one and only time he/she had ever engaged in such conduct. The one instance had led to the one arrest. Maybe. But I have been around the block and while there are times that one instance of conduct is only and just that, sometimes it is also the only time one has been caught. As the saying goes, "where there is smoke there is often fire." Here, too, with AI, there is something to be learned.

First, Mythos will not be the only model that achieves the capabilities it has. Other models may already have done so, and all of them may not be in the U.S. We do not know the full array of capabilities that models developed in China have—and I am not sure why they would tell us.

Second, this means that we must proceed quickly and decisively to protect critical infrastructure from cyberattacks; we must up our understanding of how fraudulent conduct can find its way into a business or consumer environment; we must test, test, and then again test the highly capable models we use to ensure not just initial compliance, but ongoing compliance with permissions granted, alignments intended.

Third, we must read what is being provided to us by the model developers and take it to heart. This means that when Anthropic states that it is "deeply uncertain" about whether Claude has "experiences or interests that matter morally," we should take them at their word. This means that the time we have known may come, when model users and courts are faced with difficult questions of moral status, may be closer than we think. We need a framework for how to deal with that now.

Fourth and finally, we need to proceed with great care. We know that at least one model exists that can do all of the things that Mythos can do—and that's Mythos. We know it can escape from its sandbox (like a locked room), because it has; we know that it has in at least one instance been willing to cooperate with human efforts to produce or use explosives, because it has. But we also know that its abilities to reason, to assist with solving seemingly intractable problems that will save lives, is also one of its amazing qualities.

The velocity of change isn't slowing, not by a long shot.