ALM | LAW.COM

New York Law Journal

TECHNOLOGY TODAY

ARTIFICIAL INTELLIGENCE

When Models Are so Capable They Can't Be Released

By Katherine B. Forrest
June 30, 2026

After more than a year of a totally hands-off approach to federal regulation, on June 2, the White House issued an executive order outlining a voluntary 30-day process for pre-reviewing very high capability AI models.

If an AI model meets certain yet-to-be-developed and perhaps never-to-be-publicly-released benchmarks (because they will be classified), they would be designated "covered frontier models". Once designated, the developer would collaborate with the Federal Government and "select trusted partners" to "promote secure innovation and strengthen the cybersecurity of critical infrastructure."

Three days later, on June 5, the White House issued a National Security Presidential Memorandum (NSPM) 11, on the topic of "Artificial Intelligence and the National Security Enterprise." NSPM 11 states both that the federal government should not get in the way of innovation, while also stating that there are use cases that can "protect our warfighters during peacetime and on the battlefield, enable precise operations that minimize harm to civilians, and ensure the United States continues to maintain technical overmatch against our adversaries and strategic competitors."


Katherine B. Forrest

NSPM 11 states that it is the policy of the current administration to "accelerate the development and use of AI for national security applications."

So where did all this come from and does it signal increased federal regulation of AI generally?

Well the answer to the first question is easier than the answer to the second. There's a backstory to the Executive Order and NSPM 11 that starts with the reality that the capabilities of AI models has been increasing at a rapid pace. But a second part that has to do with increased

KATHERINE B. FORREST *is a partner at New York law firm Paul, Weiss, Rifkind, Wharton & Garrison LLP.*

use of autonomous AI in many domains of security and warfare worldwide.

Let's focus on a publicly accessible part of the backstory to the Executive Order. Here's what the public knows: in our real, and not sci-fi world, in mid-April, Anthropic publicly stated that it determined one of its models (or family of models) too powerful to release to the general public. This was a first, and was a wake-up call for many in government and the private sphere.

We will walk through a short version of the public history of "Claude Mythos Preview".

At the end of March a news article in Fortune leaked the existence of powerful new family of models being developed by Anthropic called "Mythos." On April 7, Anthropic made its first public statements about Claude Mythos Preview—describing it as a step-change in AI performance, intelligence and cyber capabilities.

About a week later, Anthropic said it decided to withhold the model from public release because, in Anthropic's view, the model demonstrated unparalleled ability to discover weaknesses, or in cyber-security-speak, vulnerabilities, in computer code. Anthropic and others expressed concern that the model had the potential to expose critical infrastructure to cyber-attack.

In fact, according to Anthropic's testing, the model had found vulnerabilities in all operating systems (that's right, "all") and all browsers. Anthropic reported that in one test of its capabilities it was asked to try and escape a secure environment (referred to as a sandbox); it did.

And it then emailed one of its developers to tell him so. Anthropic characterized Mythos Preview's capabilities as going beyond finding "bugs" or errors in code; according to Anthropic, the model was able to discover and reason through vulnerabilities across different parts of software—sometimes just a weakness that it could work on, and break through.

The system card for Claude Mythos Preview is worth a read—and available on the web. It describes all of this and more in some detail. But what happened next was unprecedented. Anthropic said it provided the model to a select and limited set of trusted entities that together controlled a significant majority of the critical infrastructure in the United States: banking, software, communications companies, among others.

Project Glasswing was born. An initial group of 20 was expanded to 150 entities. During that same period of time, Anthropic said it worked with the Federal Government because of Claude Mythos's ability to be weaponized by bad actors against U.S. infrastructure.

Some of what happened next is not crystal clear but the events are as follows: on June 2, the White House issued its Executive Order outlined above; on June 5, NSPM 11 was released, on June 9, Anthropic released a modified version of Claude Mythos Preview called Claude Fable 5 and Claude Mythos 5. Anthropic said it had added certain guardrails and mitigations to the Preview version of the model.

As of June 9, Claude Mythos 5 was released to at least some of the original Project Glasswing partners. Anthropic announced that it also gave access to about 150 more organizations in 15 countries. General customers of Claude Enterprise were not given access to Claude Mythos 5 but instead were given access to Claude Fable 5.

Anthropic described Fable 5 as having additional protections built into it that prevented certain queries, for instance, from being answered. If a query crossed a line, it was processed by Claude Opus 4.8 automatically (though this could be interrupted and the query would just not be answered).

Based on Anthropic's description of those mitigations, it appeared, for just a moment, that there was a way to allow safe public access to some of the powerful capabilities of the Mythos family of models. But that didn't last long.

On June 12-just three days after the public release—the U.S. Commerce Department issued a directive (under an export control framework) requiring Anthropic to suspend access to foreign nationals to both Fable 5 and Claude Mythos 5, citing national security concerns. Given the difficulties in determining who is a foreign national and who is not, Anthropic suspended access to the models generally.

There have been news reports that a company using the models, and the federal government, had discovered a way around some of the safeguards. But for now, the Mythos model family has been pulled from the public.

So where does this leave us? In terms of regulation, there is a fabric of executive orders, the NSPM 11, export control directives from the Department of Commerce based on national security concerns. There is no overarching set of federal regulations.

As a practical matter for companies using various AI models, no action is needed. But for developers, some uncertainty has been injected into the clarity of model release. The actions with regard to Claude Mythos 5 and Fable 5 occurred after release. Ex post actions will not always be successfully executed, particularly with open source models or if the secret sauce of a model (sometimes it can be the model weights and parameters) are otherwise known.

One thing we do know is that as powerful as the Mythos family is, it is unlikely to be a unicorn. There are or will be other models either in development, that have been or will be released, that will have similar capabilities.

That's just based on the history of how AI models have progressed to date: someone may get there first, but others catch up. And we have little insight into what other countries, particularly China, may have now or are on the cusp of having.

For sure, within a short period of time we will somehow learn of another developer whose model has achieved some of the same benchmarks. We don't know what happens after that.