Trump plan to test AI models has a problem—US security teams were gutted by DOGE

Once covered models are defined, Nguyen then warned that the effectiveness of the safety testing will likely depend on whether AI firms are fully transparent and treat the process as a "genuine collaboration."

"Underneath the definitional problem sits an observability problem," Nguyen wrote. "The government cannot assess what it cannot see, and frontier capabilities are visible only to the labs that build them."

Ferren suggested that "the window for erecting proper cyber defenses to new AI models may also close quickly," and that even a well-designed government program may struggle to properly vet frontier models in such a short timeframe. "Even when well implemented, pre-deployment testing has limits," Ferren said, noting that Google's threat intelligence team has found state-aligned actors using frontier models to automate cyberattacks and "researchers have shown that Mythos-style vulnerability reasoning can be reproduced with open-weight systems."

So while AI may voluntarily submit to testing, they may be financially motivated to seek a rubber-stamp, rather than work with the government to test known frontier capabilities to their fullest extent.

"It will likely prove difficult to develop models that are incapable of malicious hacking yet remain commercially compelling," Ferren said.

He concluded that the EO “may yield short-term cybersecurity benefits,” but the “long-term effect” remains “unclear.”

Nguyen suggested the EO takes necessary steps to create “classified cyber benchmarking, voluntary prerelease evaluation, and coordinated vulnerability scanning” that “the national security community will need for decades” to “continuously evaluate systems that are probabilistic rather than deterministic, autonomous rather than directed, and whose capabilities change with every update.”

But the safety testing will have to evolve as fast as the technology does, Nguyen said, otherwise we risk assessing emerging models against “yesterday’s risks.”

That’s why, at its core, the process will depend on an honest exchange between stakeholders with deep technical expertise and confidential national security insights. It’s the only way to ensure the US focuses its energies on protecting the public from the most credible and consequential AI risks, rather than just providing “performative reassurances,” Nguyen wrote.

https://arstechnica.com/tech-policy/2026/06/trumps-ai-executive-order-may-not-prevent-dangerous-deployments/