Opinion | Anthropic’s Restraint Is a Terrifying Warning Sign

Thomas L. Friedman

Thomas L. Friedman

April 7, 2026

Credit...Vincent Forstenlechner/Connected Archives

Listen · 7:45 min

Normally right now I would be writing about the geopolitical implications of the war with Iran, and I am sure I will again soon. But I want to interrupt that thought to highlight a stunning advance in artificial intelligence — one that

arrived sooner than expected and that will have equally profound geopolitical implications.

The artificial intelligence company Anthropic announced Tuesday that it was releasing the newest generation of its large language model, dubbed Claude Mythos Preview, but to only a limited consortium of roughly 40 technology companies, including Google, Broadcom, Nvidia, Cisco, Palo Alto Networks, Apple, JPMorganChase, Amazon and Microsoft. Some of its competitors are among these partners because this new A.I. model represents a "step change" in performance that has some critically important positive and negative implications for cybersecurity and America's national security.

The good news is that Anthropic discovered in the process of developing Claude Mythos that the A.I. could not only write software code more easily and with greater complexity than any model currently available, but as a byproduct of that capability, it could also find vulnerabilities in virtually all of the world's most popular software systems more easily than before.

The bad news is that if this tool falls into the hands of bad actors, they could hack pretty much every major software system in the world, including all those made by the companies in the consortium.

This is not a publicity stunt. In the run-up to this announcement, representatives of leading tech companies have been in private conversation with the Trump administration about the implications for the security of the United States and all the other countries that use these now vulnerable software systems, technologists involved told me.

For good reason. As Anthropic said in a written statement on Tuesday, in just the past month, "Mythos Preview has already found thousands of high-severity vulnerabilities, including some in *every major operating system and web browser.* Given the rate of A.I. progress, it will not be long before such capabilities proliferate, potentially beyond actors who committed to deploying them safely. The fallout — economics, public safety and national security — could be severe.''

Project Glasswing, Anthropic's name for the consortium, is an undertaking to work with the biggest and most trusted tech companies and critical infrastructure providers, including banks, "to put these capabilities to work for defensive purposes," the company added, and to give the leading technology firms a head start in finding and patching those vulnerabilities.

"We do not plan to make Claude Mythos Preview generally available, but our eventual goal is to enable our users to safely deploy Mythos-class models at scale — for cybersecurity purposes, but also for the myriad other benefits that such highly capable models will bring," Anthropic said.

My translation: Holy cow! Superintelligent A.I. is arriving faster than anticipated, at least in this area. We knew it was getting amazingly good at enabling anyone, no matter how computer literate, to write software code. But even Anthropic reportedly did not anticipate that it would get this good, this fast, at finding ways to find and exploit flaws in existing code.

Anthropic said it found critical exposures in every major operating system and Web browser, many of which run power grids, waterworks, airline reservation systems, retailing networks, military systems and hospitals all over the world.

If this A.I. tool were, indeed, to become widely available, it would mean the ability to hack any major infrastructure system — a hard and expensive effort that was once essentially the province only of private-sector experts and intelligence organizations — will be available to every criminal actor, terrorist organization and country, no matter how small.

I'm really not being hyperbolic when I say that kids could deploy this by accident. Mom and Dad, get ready for:

"Honey, what did you do after school today?"

"Well, Mom, my friends and I took down the power grid. What's for dinner?"

That is why Anthropic is giving carefully controlled versions to key software providers so they can find and fix the vulnerabilities before the bad guys do — or your kids.

At moments like this I prefer to do a deep dive with my technology tutor, Craig Mundie, a former director of research

and strategy at Microsoft, a member of President Barack Obama's President's Council of Advisers on Science and Technology and an author, with Henry Kissinger and Eric Schmidt, of a book on A.I. called "Genesis."

In our view, no country in the world can solve this problem alone. The solution — this may shock people — must begin with the two A.I. superpowers, the U.S. and China. It is now urgent that they learn to collaborate to prevent bad actors from gaining access to this next level of cyber capability.

Such a powerful tool would threaten them both, leaving them exposed to criminal actors inside their countries and terrorist groups and other adversaries outside. It could easily become a greater threat to each country than the two countries are to each other.

Indeed, this is potentially as fundamental and significant a turning point as was the emergence of mutually assured destruction and the need for nuclear nonproliferation. The U.S. and China need to work together to protect themselves, as well as the rest of the world, from humans and autonomous A.I.s using this technology — a lot more than they need to worry about Russia.

This is so important and urgent that it should be a top subject on the agenda for the summit between Trump and President Xi Jinping in Beijing next month.

"What used to be the province of big countries, big militaries, big companies and big criminal organizations with big budgets — this ability to develop sophisticated cyberhacking operations — could become easily available to small actors," explained Mundie. "What we are about to see is nothing short of the complete democratization of cyberattack capabilities."

It means that responsible governments, in concert with the companies that build these A.I. tools and software infrastructure, need to do three things urgently, Mundie argues.

For starters, he says, we need to "carefully control the release of these new superintelligent models and make sure they only go to the most responsible governments and companies."

Then we need to use the time this buys us to distribute defensive tools to the good actors "so that the software that runs their key infrastructure can have all their flaws found and fixed before hackers inevitably get these tools one way or another." (By the way, the cost of fixing the vulnerabilities that are sure to be discovered in legacy software systems, like those of telecommunications companies, will be significant. Then multiply that across our whole industrial base.)

Finally, Mundie argues, we need to work with China and all responsible countries to build safe, protected working spaces, within all the key networks, both public and private, into which trusted companies and governments "can move all their critical services — so they will be protected against future hacking attacks."

It will be interesting to see what history remembers most about April 7, 2026 — the postponed U.S. release of bombs over Iran or the carefully controlled release of the Claude Mythos Preview by Anthropic and its technical allies.

Thomas L. Friedman is the foreign affairs Opinion columnist. He joined the paper in 1981 and has won three Pulitzer Prizes. He is the author of seven books, including "From Beirut to Jerusalem," which won the National Book Award. @tomfriedman • Facebook

A version of this article appears in print on April 9, 2026, Section A, Page 18 of the New York edition with the headline: Anthropic's Restraint Is Not a Publicity Stunt. Order Reprints | Today's Paper | Subscribe