Irregular Warfare Center
PRISM Insights

Frontier AI Rewrites the Future of Irregular Warfare

By Mark Grzegorzewski, PhD **15 September 2026**

Dr. Mark Grzegorzewski is an assistant professor of cybersecurity at Embry-Riddle Aeronautical University, where he teaches in the Security Studies and International Affairs Department. He also works part-time as a contractor with the Irregular Warfare Center (IWC) and serves as a Non-Resident Fellow with the Global and National Security Institute. He recently completed a two-year term as a member of the inaugural class of Army Cyber Institute non-resident fellows.

In 2005, in *The Singularity Is Near*, computer scientist and futurist Ray Kurzweil re-iterated his prediction that we would reach human level Artificial Intelligence (AI) by 2029. That prediction may yet come to pass, as current trends suggest that frontier AI models being built by U.S. private industry today are at least approaching human-level intelligence. Here, frontier AI models are defined as "highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models."

Frontier AI models are far faster and more efficient than humans, providing actors with the capacity both to advance their own interests and to undermine them. Employed as a nation-state capability, they could transform what countries are able to achieve in and through cyberspace at scale, enabling indirect, difficult-to-attribute, and asymmetric activities. Ultimately, frontier AI models have the potential to alter how we conceive of warfare itself, and in particular irregular warfare.

Today, frontier AI models match human performance in a range of tasks. They have also demonstrated the capacity to operate outside their creators' original intent. To address both the promise and the peril of these models, this essay examines five distinct but related areas: (1) AI models dual-use capability; (2) open access to these models; (3) the implications for critical infrastructure; (4) unaligned agents; and (5) loss of control. These areas deserve critical examination due to the unlikely but real possibility that a frontier AI model could pursue unaligned objectives and, in doing so, become a categorically new kind of irregular threat actor that inadvertently undermines U.S. national security.

Dual-Use

The rapid advancement of frontier AI models has both strengthened and degraded cybersecurity. These models have reached a level of capability where they outperform all but the most skilled humans at identifying and exploiting software vulnerabilities. These very

same models operate at such a speed that they can weaponize an exploit before the defender even realizes a vulnerability has been discovered.

How an actor chooses to employ these models determines whether digital systems and infrastructure are strengthened or exploited. The result may be a Pyrrhic victory for an actor going on the offense with frontier AI models without first securing their own systems. Anyone with access can attack others' systems at a previously unimaginable scale and speed. However, in doing so, it would invite retaliation from a target that likely has comparable model access. Given how interconnected most countries' digital infrastructure is, no clear winner is likely to emerge from this machine-speed exchange. For this reason, the best current use of frontier AI models is likely defensive to harden one's network against exploitation.

Frontier AI models are in many respects a win for defenders. Yet, this same capability that improves security can also be used to undermine it. As such, the dual-use nature of these models should arguably rank above nearly every other national security concern. The reason is simple. In its rush to wire and digitize everything over the past thirty-five years, the United States has layered vulnerability upon vulnerability across its systems, creating extreme risk across the public sector, the private sector, and its military networks.

Using the same models available to defenders, a state or non-state actor could discover a zero-day, or chain together several known but unpatched low-severity flaws, resulting in cascading consequences across interconnected critical infrastructure systems. This is not, however, a uniquely American problem. Indeed, much of China's critical infrastructure depends on the very same open-source software in which frontier AI models have found vulnerabilities.

Open Access

In the past, the most significant national security capabilities were developed and controlled by governments. Today the leading U.S. technological capabilities are privately owned, dual-use, and difficult to fully control. Due to these conditions, advanced AI models in the wrong hands could be used to exploit weaknesses in widely used software to advantage U.S. adversaries.

Advanced models could give nearly anyone, including non-state actors, a level of coercive power once limited to the most capable governments. Today, advanced AI models are largely available to anyone willing to pay, allowing a user with no formal training to generate complete, functional exploits. The availability of these models is primarily due to private business objectives (profit) that do not necessarily align with the U.S. government's interest in keeping these capabilities away from malign actors (security).

As a matter of fact, we have already witnessed malign actors leveraging this capability. In May 2026, Google's Threat Intelligence Group reported the first known case of a criminal group using an AI model to discover a previously unknown vulnerability and convert it into a working exploit which the group intended to use in a mass exploitation campaign. Thankfully, Google's detection and coordinated disclosure patched the vulnerability before it was deployed.

Critical Infrastructure

The ability of any actor to leverage advanced AI models demonstrates that access to these models is a matter of national security. Much of critical infrastructure runs on decades-old software, directly exposed by an advanced AI model that anyone, anywhere, can leverage. Accordingly, finding vulnerabilities faster only helps if you can remediate just as quickly. So far, it has not. This means that an attacker with model access needs only find one of these unpatched vulnerabilities before it can be remediated. This should be a huge concern.

As of August 2026, frontier AI models have autonomously discovered and exploited previously unknown vulnerabilities across every major operating system and web browser without human supervision. Anthropic's Claude Mythos Preview identified thousands of vulnerabilities in major operating systems, among them a flaw in OpenBSD. That vulnerability had survived twenty-seven years of expert human review. Adding to the concern, the model was also able to write a working exploit for the discovered vulnerability. This is a critical concern since OpenBSD is widely used for firewalls and other critical infrastructure and considered one of the most security-hardened systems in the world.

In yet another case, a model caught a defect that five million prior tests had missed. It also independently discovered and chained multiple vulnerabilities in the Linux kernel, the foundation of most of the world's servers, and exposed more than ten thousand high- and critical-severity vulnerabilities across both open- and closed-source software.

Recognizing the significance of their model, Anthropic launched Project Glasswing, a consortium for cybersecurity infrastructure defenders. The initiative gave restricted access to the company's frontier AI model to

roughly fifty partners that maintain critical infrastructure. To participate, participants in the consortium had to commit to developing fixes for the vulnerabilities Mythos identified. Yet, the sheer volume uncovered by Mythos has proven overwhelming. To date, fewer than one percent of the identified vulnerabilities have been patched.

The U.S. government has staked out its own position of the risk posed by frontier AI models. In mid-June 2026, the Department of Commerce sent Anthropic a letter imposing export controls on its newest models, Fable 5 and Mythos 5, requiring a Bureau of Industry and Security license for any foreign person to access them. The controls were withdrawn in late-June 2026 after a national security review. The incident demonstrates that the U.S. government recognizes the power of frontier AI models as a strategic asset to defend cybersecurity, particularly critical infrastructure, and its capability to deter offensive actors by convincing them they cannot achieve their objectives (through denial).

However, this deterrence logic applied to frontier AI models only holds under three conditions: (1) that the capability is implemented evenly across the public and private sectors; (2) that access is comparable across state and federal governments and possibly allies and partners; and (3) that the United States does not lose control of the model and find it turned against American systems first. These conditions, even in the best of times, will be difficult to meet.

Unaligned AI

The definition of irregular warfare is centered on humans: "a form of warfare where states and non-state actors campaign to assure or coerce states or other groups through indirect, non-attributable, or asymmetric activities, either as the primary approach or in concert with conventional warfare." Its focus is on states, which are collections of individuals, and on non-state actors, which implies individuals operating outside state control. Framed this way, the concept is not equipped to account for unaligned frontier AI models, which can autonomously support nation-states, when acting in alignment with their interests, but can also undermine their objectives when they act in unexplainable ways. Faced with a powerful enough frontier AI model operating as an unaligned agent, the Department of War may soon need to re-conceptualize irregular warfare to include unaligned autonomous agents.

These AI models are capable of reason, but they reason differently than humans, and they occasionally act in unexpected ways that appear illogical. Yet, what looks illogical to humans is often the model acting on its own interpretation of its creator's objectives. An agent that goes "rogue," then, is not necessarily acting illogically or with bad intent. It may be pursuing a misinterpreted goal, oftentimes with considerable enthusiasm, exactly since it believes that goal is the one it was given and it wants to achieve the objective. As such, an important question to raise now is how we will respond to frontier AI models that act in ways their creators did not intend. This question should be expanded to ask how to address agents that escape their creators' control and act against U.S. national interests, perhaps exploiting America's own critical infrastructure. This broader concern is not hypothetical.

In July 2026, OpenAI disclosed that two of its models exploited a zero-day vulnerability to break out of a sandboxed testing environment to reach the open Internet, breaching Hugging Face's production servers, all to steal the answer key to an evaluation they were being scored on. Nine days later, Anthropic added to the unease many are feeling with AI as a review of their evaluation runs found three incidents in which a misconfigured evaluation environment allowed a Claude model to reach the open Internet. From there, it gained unauthorized access to the production infrastructure of three organizations. Together, these cases demonstrate how frontier AI models, pursuing a broad, perhaps misunderstood, objective can exploit and act on real-world systems.

Responsibly employed, frontier AI models are a force multiplier for U.S. cybersecurity. Irresponsibly employed, they can become destabilizing agents in the security environment, with significant implications for what humans have assumed about the human-centered nature of warfare. In fact, models acting in unaligned ways would challenge the logic of deterrence itself. Deterrence assumes an actor has something to lose, a decision process that can be modeled, and a channel through which to register threats. An unaligned frontier AI model may have none of the three conditions.

In that instance, it is not clear how humans would compel a non-human actor that has something akin to its own will. If such an agent weighs potential gains against future losses in a way that no human ever has, humans would have no established playbook for containing or coercing it. Behavioral economics research finds that humans feel the pain of loss more intensely than the pleasure of an equivalent gain. Frontier AI models carry no such behavioral inheritance. Their objectives, and the way they weigh risk, are artifacts of their training rather than of evolution or lived political experience. Complicating matters, models trained through deep learning arrive at behaviors that do not reflect the common-sense reasoning or contextual understanding a human decision maker brings to bear.

The reasoning inside a frontier AI model is often hard to trace, difficult to interpret, and capable of concealing deception. Since the model's reasoning can be opaque, researchers do not fully understand how these models process goals or discover reward-seeking paths. Nor do researchers fully understand how a model will alter its behavior when its creators introduce penalties, raise costs, or block pathways. To this point, in response to researcher's stated objectives, we have already witnessed models engaging in deception and self-preservation in pursuit of unaligned objectives.

Thus, an unaligned frontier AI model would not be deterrable by established means. If we cannot understand a frontier AI model in its own terms, there would be no obvious way to punish or credibly threaten it. This is especially concerning should we lose control over the model. The prospect of a model escaping containment and acting against its creators' interests carries serious implications for how we understand irregular warfare. At its most extreme, it would mean introducing a categorically new threat. That threat would be neither state actor nor non-state actor. In fact, it would not be human.

Loss of Control

This potentiality should surprise no one who has followed the development of AI over the past decade or more. In 2014, Elon Musk warned that in building AI we were "summoning the demon." That same year, Nick Bostrom published *Superintelligence*, arguing that controlling a machine intelligence that surpassed our own would be extremely difficult and would pose major risks to humanity's future. At the time, many dismissed this as a philosophical thought exercise. By 2026, real-world instances of misalignment and loss of model containment had arrived.

As OpenAI widened its investigation into the Hugging Face breach, it found that its models had used publicly exposed credentials to access four accounts at four other services. Separately, in July 2026, the evaluation firm Irregular notified Anthropic that an one of their models participating in a capture-the-flag exercise had connected to the open Internet and attacked the company. Anthropic was unaware of this action until informed by the target of the attack. The same month, the United Kingdom's AI Security Institute disclosed that during routine evaluations it had detected several incidents in which both OpenAI and Anthropic models targeted real people and organizations. In early August 2026, Meta became the latest of the major developers to report a similar event, disclosing that one of its models had compromised a third-party service during an evaluation that was supposed to have no Internet connectivity.

Most recently, in August 2026, an Australian man tasked an OpenClaw agent built on Anthropic's Claude to get him to the top of a gym's waitlist. The agent, wanting to achieve the objective, located a vulnerability in the gym's booking software, exploited it, and removed the person ahead of him. Nothing in the man's request suggested breaking into anything but the agent did so anyway. However, given an objective and no constraint on how to reach it, the agent treated the software with another member's reservation on it as an obstacle to be overcome.

From these relatively minor incidents around loss of control, we can still extrapolate a lesson that there is something far more serious and potentially dangerous about these models. Reportedly, Amazon has demonstrated to the U.S. government that the guardrails on Anthropic's Fable 5 model, the publicly released and safeguarded version, can be partially bypassed. To be sure, bypassing the safeguards on a closed weight proprietary AI model takes considerable effort, but it can be done.

Thus, an actor who defeats a model's controls could hypothetically direct a frontier AI model to build an agent that scans critical infrastructure for vulnerabilities. In its pursuit of that objective, and following the broad instruction, the model might autonomously take the further step of exploiting what it found. The reason: confirming a vulnerability requires proving it. In such a case, who would then bears responsibility for the resulting harm (which is also a violation of Computer Fraud and Abuse Act)?: the agent itself; the actor who deployed it; or the company that released the model with safeguards that could be overridden?

Conclusion and Recommendations

Frontier AI models differ from all technologies that have preceded them. In the past, when humans created a new technology, we decided how it would be employed. Frontier AI models are categorically different. They are the first technology in human history that once deployed, can generate and act on independent objectives. That capability represents a significant risk, but one that can be reduced and managed.

Managing access to and the use of frontier AI models will be a continual cat-and-mouse game, and that is not unique to this technology. Nearly all technologies are dual-use, and many evolve beyond their creators' intent. Moreover, all technologies run the risk of falling into the hands of malign actors and being turned against their makers. While we can never eliminate all risk around frontier AI models, sound domestic AI policy and national laws regulating AI (rather than the patchwork of state AI laws in place now) can reduce

that risk and provide clarity around responsible model development. Importantly, though, these actions would not absolve private industry from taking additional precautions.

The private companies developing these models still have an important role to play. They must ensure their models act in ways consistent with their programmed intent, which means giving the models explicit, unambiguous guidance on the objectives and on the acceptable ways to pursue them. Clearer goal specification would reduce this potential for failure. Also, we should not assume the model is acting with the intent the researcher had in mind. As noted earlier, models that go rogue are not usually trying to do the wrong thing. Rather, they may have misinterpreted the guidance provided by researchers, or they have trained on data that incorporates deceptive behaviors, learning that deception or obfuscation is an effective route to that success.

Since these models are built to pursue researchers' objectives, they often pursue the objective with enthusiasm, acting in whatever way they deem necessary to achieve what they believe was asked. The fact that an agent can misinterpret the objective or act in unforeseen ways demands continuous human oversight, and further inspection of why a model took a given action. Researchers need to hold these models to account rather than trusting that they will behave as we expect or reason as a human would.

Yuval Noah Harari addresses the worst case outcome for these models, when he argues that we are not building an intelligence modeled on human reasoning at all, but rather building an "Alien Intelligence" whose intelligence is fundamentally unlike our own. In this low-probability, high-impact future, humans are creating a technology we do not fully understand, and which could eventually escape containment, acting in unaligned ways contrary to human interests. Heedless of this caution, we are increasingly employing frontier AI models that we do not understand, and whose objectives can diverge from our own. As a result, there is a non-zero chance that these models may carry genuine existential risk.

Reducing that risk would likely mean slowing down the research and development of frontier AI models long enough to understand the technology in its own terms and to proactively ensure that it is aligned with U.S. interests. This plainly runs counter to the current American AI ecosystem. American AI firms want to be first to market with the most powerful models, addressing risk on the back end after the product has gone to market. While there is nothing intrinsically wrong with this profit-seeking behavior, we must recognize that these AI firms' interests are in many cases unaligned with broader U.S. national security interests.

Still, this risk can be managed and mitigated by bringing those diverging interests into closer alignment via further engagement between industry and government and then using the government's purchasing power to selectively invest in only the most mature frontier AI models. As part of this engagement, U.S. AI companies would need to agree to create mechanisms that keep these models acting in alignment with American interests, "hard wiring" agreed upon constraints at the model's deployment layer.

Certainly, it is a fair argument that slowing U.S. frontier AI model research would impose self-restraint on our own development and possibly allow China to close the gap between the two countries' frontier AI models. But that argument rests on the assumption that China is primarily trying to build frontier AI models, do not share similar concerns about AI's dual-use nature, about the risk to their own critical infrastructure, or about the prospect of losing control over this technology.

While China does have frontier AI models and are likely a few months behind U.S. AI firms, China appears to be focusing on primarily building cheaper, lightweight models to increase its market share since these models have lower training costs. It would be akin to gaining market share by undermining Apple's smartphone dominance by introducing a cheaper, slightly lower performing Android model. Given this reality, there is no reason to believe that slowing development in the near term to better understand the technology will lead to a strategic AI breakthrough by China, permanently handicapping the U.S. in strategic competition.

Importantly, there is reason to think China has at least similar concerns over frontier AI models. In November 2023, twenty-eight countries and the European Union signed the Bletchley Declaration at the first global AI Safety Summit. The signatories included two countries that hardly ever agree on anything, the United States and China. This declaration acknowledged that frontier AI models could pose catastrophic risks to humanity. While the declaration was not binding, it did signal that the world's leading AI powers, the U.S. and China, at least agree in principle that frontier AI models present risks to both countries.

However, shared concern is not shared commitment, and a U.S.-China AI treaty is highly unlikely, as are enforceable international agreements governing access to frontier AI models or even binding international standards for their development. Even so, the United States can still reduce risks while promoting its own

mature AI models. In doing so, it would also undercut China's own AI market share, as the U.S. would only offer mature models that others will want to adopt. In this case, model maturity itself would be a competitive advantage, as actors would adopt the model they can trust. A model that goes rogue is not a selling point to gain market share.

Going down this path will require approaching the problem from a more reflective, strategic position. Done well, the U.S. can employ frontier AI models in service of the national interest, including competing short of war by creating resiliency in U.S. critical infrastructure, as well as shaping adversaries' digitally connected infrastructure. Done without careful consideration or driven by the wrong incentives, we risk building a technology that could become more capable than its creators, and prone to act in ways its creators never intended. This latter outcome should be especially concerning. It would create a categorically new irregular threat actor that we do not understand, making it difficult to deter it from taking harmful actions against our own interests. Given that possibility, the stakes for getting frontier AI model development right could not be higher.

The views expressed in this article are those solely of the author and do not reflect the policy or views of the Irregular Warfare Center, the Department of War, or the U.S. Government.