April 2026

ICIT

The Deferral Trap

Compounding Risk and AI Adoption Governance

David Mussington, Ph.D., CISSP, DDN QTE
ICIT Fellow, Co-Chair, ICIT FCEB Resilience Center
Professor of the Practice, University of Maryland School of Public Policy

www.icitech.org

Table of Contents

About ICIT

The Institute for Critical Infrastructure Technology (ICIT) is a nonprofit, nonpartisan, 501(c)3 think tank with the mission of modernizing, securing, and making resilient critical infrastructure that provides for people's foundational needs. ICIT takes no institutional positions on policy matters. Rather than advocate, ICIT is dedicated to being a resource for the organizations and communities that share our mission. The views and opinions expressed in this essay are solely those of the author(s) and do not necessarily reflect the official policy or position of ICIT. Any assumptions made within the analysis are not reflective of the position of any entity other than the author(s). To learn more, please visit www.icitech.org

CyberRisk ALLIANCE | Thank you to our Strategic Partner CyberRisk Alliance | cyberriskalliance.com

Introduction

On April 7, 2026, Anthropic took an unusual step. It published a 244-page system card for a model it had no intention of releasing to the public. Claude Mythos Preview — described in internal company documents as the most capable model it had ever developed — had demonstrated cybersecurity capabilities so advanced that Anthropic concluded broad release was not responsible. Instead, it launched Project Glasswing: a gated initiative providing access to a curated set of technology partners, with the explicit goal of letting defenders get ahead of the model's offensive potential before equivalent capabilities proliferated to actors less committed to responsible deployment.

Anthropic's frontier red team found that Mythos Preview was capable of identifying and then exploiting zero-day vulnerabilities in every major operating system and every major web browser — with many vulnerabilities ten or twenty years old, and with no human involved in either discovery or exploitation after an initial prompt.[1] In one documented case, Mythos Preview fully autonomously identified and exploited a 17-year-old remote code execution vulnerability in FreeBSD, triaged as CVE-2026-4747, that allows anyone to gain root on a machine running NFS — beginning from an unauthenticated position anywhere on the internet.[1] Critically, Anthropic did not explicitly train Mythos Preview to have these capabilities. They emerged as a downstream consequence of general improvements in code, reasoning, and autonomy — the same improvements that make the model substantially more effective at patching vulnerabilities also make it substantially more effective at exploiting them.[1]

The Mythos moment is significant for reasons that extend beyond its immediate cybersecurity implications. It establishes, with unusual clarity, that AI capability has crossed a threshold in the offensive domain — and that the defensive and governance infrastructure available to most institutions has not kept pace. That gap is not a projected risk. It is a present condition.

This paper argues that the gap has a specific and underappreciated cause: the systematic underweighting of deferral costs in institutional AI risk calculations. Organizations and governance bodies that have treated AI skepticism as the default responsible posture have implicitly assumed a stable baseline — one in which restraint preserves optionality without accumulating cost. That assumption is wrong. The baseline is not stable. And in an environment where adversarial AI capability is advancing on timelines that do not pause for institutional deliberation, skepticism toward AI adoption does not reduce risk. It compounds it.

The argument proceeds in five sections. Section 1 examines the logical structure of the inversion problem — why caution, in a deteriorating threat environment, is not cost-free. Section 2 grounds that claim empirically, drawing on Volt Typhoon, Salt Typhoon, Iranian ICS operations, and Mythos as four reference points that together characterize the current baseline. Section 3 develops the compounding risk mechanism — the specific ways in which deferral accumulates risk asymmetrically over time. Section 4 describes what governed adoption looks like in practice, establishing that the choice between capability and accountability is not forced. Section 5 draws out the policy implications for institutions operating in the current environment.

Section 1: The Inversion Problem

Caution toward powerful and rapidly evolving technologies is rational. Artificial intelligence presents genuine governance challenges: opacity in model behavior, difficulty attributing outputs, risks of miscalibrated deployment in high-stakes contexts, and institutional capacity gaps that make oversight difficult to sustain. Organizations and policymakers who approach AI adoption carefully, who insist on accountability frameworks before deployment, and who resist pressure to adopt capabilities faster than governance structures can absorb them, are not being obstructionist. They are being responsible.

The argument here is not against caution. It is against the assumption that caution is cost-free.

Every risk management posture is evaluated against a baseline. The implicit baseline in most institutional AI skepticism is a stable status quo — one in which deliberate, measured adoption preserves optionality while avoiding the harms associated with premature deployment. That baseline assumption deserves scrutiny, because the environment in which these decisions are being made is not stable. It is deteriorating.

Critical infrastructure sectors face sustained, patient, and increasingly sophisticated adversarial pressure. Governance expertise is insufficient relative to the pace of technological change. And the frontier of AI capability — including offensive applications — is advancing on a timeline that does not pause for institutional deliberation. In that environment, restraint has a cost that compounds over time. Each period of deferred adoption is a period in which the gap between adversarial capability and defensive capacity widens.

The question this paper poses is not whether caution is warranted. It is whether caution, in a heightened risk environment, is itself risk-free. The answer, we argue, is that it is not — and that recognizing this changes what a genuinely conservative AI governance posture looks like.

Section 2: What the Baseline Actually Looks Like

Abstract arguments about deteriorating threat environments require empirical grounding. The case that the status quo is not a safe baseline rests on specific, documented developments — not projected risks, but observed ones. Four reference points illustrate the trajectory.

Volt Typhoon. Beginning no later than 2021 and continuing through at least 2024, a People's Republic of China-affiliated threat actor conducted systematic, patient intrusion into United States critical infrastructure networks — communications, energy, water, transportation. A joint advisory from CISA, NSA, and FBI assessed that these PRC state-sponsored actors were seeking to pre-position themselves on IT networks for disruptive or destructive cyberattacks against U.S. critical infrastructure in the event of a major crisis or conflict, and that the group had maintained access and footholds within some victim IT environments for at least five years.[2] The operational signature was notable for what it was not: U.S. authoring agencies assessed with high confidence that Volt Typhoon actors were pre-positioning themselves on IT networks to enable lateral movement to OT assets to disrupt functions across multiple critical infrastructure sectors — behavior assessed as inconsistent with traditional cyber espionage or intelligence gathering operations.[3] This is infrastructure compromise as strategic latency — capability held in reserve, to be activated on adversarial timelines, not defensive ones.

Salt Typhoon. Where Volt Typhoon targeted operational infrastructure, Salt Typhoon targeted the communications layer itself. Salt Typhoon breached major telecom carriers in a global, multi-year espionage operation that targeted the phone conversations of key American officials, with at least 600 organizations notified by the FBI that the hackers had interest in their systems, according to FBI Cyber Division Assistant Director Brett Leatherman.[4] The intrusion reached into the most sensitive layer of domestic communications infrastructure: Salt Typhoon breached America's lawful intercept systems that house wiretap requests used by law enforcement to surveil suspected criminals and spies — systems that telecoms are required to engineer under the Communications Assistance for Law Enforcement Act of 1994.[5] The Chinese campaign extended beyond the telecom industry into transportation and military infrastructure networks, with the data stolen providing Chinese intelligence services the capability to identify and track targets' communications and movements globally.[4] By August 2025, FBI Assistant Director Leatherman confirmed that Salt Typhoon had broken into companies in more than 80 countries, revealing for the first time the full global scale of the campaign — with AT&T, Verizon, and Lumen among the previously confirmed U.S. victims.[6] Senator Mark Warner, then-Chairman of the Senate Intelligence Committee, characterized the operation as the worst telecom hack in the nation's history. As late as December 2024, CISA's executive assistant director for cybersecurity stated that U.S. officials could not say with certainty that the adversary had been evicted, because the full scope of what they were doing remained unknown.[7]

Iranian ICS Operations. PRC actors are not the only state-sponsored threat operating against U.S. critical infrastructure at operational scale. Iranian government-affiliated actors, specifically IRGC-linked groups, have conducted sustained campaigns against U.S. operational technology in the water, energy, and government sectors. Beginning in November 2023, IRGC-affiliated cyber actors accessed multiple U.S.-based water and wastewater facilities by compromising internet-accessible Unitronics PLC devices, with the campaign extending through January 2024 and targeting devices across multiple critical infrastructure sectors including water, energy, and government services.[8] These were not reconnaissance operations. As a result of the exploitation, organizations from multiple U.S. critical infrastructure sectors reported disruptions including configuration wiping, software-based mechanical sensor tampering, and disruption of human machine interfaces — resulting in operational disruption and financial loss.[9]

The campaign has continued to evolve. A subsequent CISA advisory designated AA26-097A, issued April 7, 2026, assessed that Iranian-affiliated APT actors were conducting exploitation activity targeting internet-facing operational technology devices across government services, water and wastewater systems, and energy sectors, with the authoring agencies assessing the group's intent as causing disruptive effects within the United States.[10] The timing of that advisory — published on the same day as the Mythos Preview announcement — is itself instructive. The Iranian threat to physical infrastructure and the AI-enabled offensive capability frontier are not sequential developments. They are concurrent conditions, operating simultaneously against the same institutional baseline.

Claude Mythos Preview. The fourth reference point is of a different character from the preceding three. Project Glasswing brings together Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks as launch partners, along with over 40 additional organizations that build or maintain critical software infrastructure.[11] The initiative was launched because, as the Mythos Preview System Card states directly, the model demonstrated "a striking leap in cyber capabilities relative to prior models, including the ability to autonomously discover and exploit zero-day vulnerabilities in major operating systems and web browsers" — capabilities that "could, if broadly available, also accelerate offensive exploitation given their inherently dual-use nature."[12] The results of independent evaluation were operationally significant: across roughly a thousand open source repositories, Mythos Preview achieved full control flow hijack on ten separate, fully patched targets — a tier of severity that previous generation models reached only once each across the same corpus.[1]

What the System Card makes explicit — and what distinguishes Mythos from a simple capability demonstration — is the alignment risk framing. Anthropic describes Mythos Preview as simultaneously "the best-aligned of any model that we have trained to date by essentially all available measures" and the model that "likely poses the greatest alignment-related risk of any model we have released to date."[12] The mechanism is instructive: a more capable model, deployed with greater autonomy in higher-stakes contexts, can produce more consequential failures even if its per-action alignment is superior. Applied to the governance question, this means that the arrival of Mythos-class capability is not simply a threat vector problem. It is a governance architecture problem — one that cannot be resolved by restricting access alone, because the underlying capability trajectory will continue regardless of any single deployment decision.

Mythos matters for this argument not primarily as a threat vector — though that dimension is real — but as a capability marker. Anthropic has already privately warned senior government officials that Mythos makes large-scale cyberattacks significantly more likely this year.[13] The question of whether adversarial actors, state-sponsored or otherwise, are developing or acquiring equivalent capability is not one that can be answered with confidence in the negative. The more defensible assumption, given the Volt Typhoon, Salt Typhoon, and Iranian ICS campaign records, is that capability development is ongoing on timelines that are not visible to defenders.

Taken together, these four data points describe a baseline that is already compromised at the infrastructure level by Chinese pre-positioning, already penetrated at the communications layer by multi-year telecom espionage, already subject to active Iranian disruption campaigns against physical control systems, and now facing an AI-enabled offensive capability frontier advancing faster than defensive and governance institutions are currently positioned to match. This is the environment in which AI skepticism operates as a default posture. It is not a stable baseline. It is a deteriorating one — on multiple simultaneous fronts.

Section 3: How Skepticism Compounds Risk

The threat environment described in the preceding section does not pause for institutional deliberation. That asymmetry is the core of the compounding risk problem.

When an organization, agency, or governance body defers AI adoption — whether for legitimate reasons of accountability, capacity, or oversight readiness — the deferral does not occur in a static environment. It occurs against a backdrop of adversarial actors who face no equivalent institutional friction. State-sponsored threat actors are not convening working groups on responsible AI deployment. They are integrating capable systems into operational workflows on timelines driven by strategic advantage, not governance maturity. The result is a capability gap that widens not because defenders are moving backward, but because the asymmetry of adoption pace is itself a form of relative decline.

This is the mechanism by which skepticism compounds risk. It is not that caution is wrong in principle. It is that caution has a duration cost that accumulates asymmetrically in a competitive environment. Each institutional cycle spent deliberating over adoption frameworks is a cycle in which adversarial capability continues to develop, infrastructure compromise deepens, and the analytical gap between what defenders can assess and what adversaries can execute grows wider. There is a secondary compounding dynamic that operates at the expertise level. Effective AI governance — the kind that produces accountable, auditable, human-teaming deployment rather than uncritical automation — requires practitioners who understand both the technology and the domain well enough to design meaningful oversight. That expertise does not exist at scale. It is developed through engagement, through building and operating governed AI systems, through the iterative process of learning where these tools fail and how failure modes can be detected and corrected. Institutions that defer adoption also defer the development of that expertise. When they eventually do adopt — as competitive pressure will ultimately require — they will do so without the institutional knowledge base that makes governed adoption possible. The choice to wait does not preserve the option of careful adoption later. It erodes the capacity to execute it.

A third dynamic operates at the policy and standards level. Governance frameworks for AI in high-stakes domains — critical infrastructure protection, national security analysis, cyber defense — are being written now, by actors who are engaged with these systems. Institutions that remain on the sidelines during this period are not avoiding the governance problem. They are ceding influence over how it gets resolved. The accountability standards, deployment frameworks, and oversight architectures that will shape AI use in consequential domains for the next decade are being developed in spaces that require active participation to influence. Skeptical non-participation is not a neutral posture with respect to that process. It is an abdication of institutional voice that could otherwise shape outcomes.

None of this argues for uncritical or ungoverned deployment. The risks associated with poorly governed AI adoption in high-stakes contexts are real and should not be minimized. A system that produces fast, fluent, and wrong outputs — without the epistemic governance architecture to detect and correct error — can cause serious harm. The argument is not that speed of adoption trumps quality of governance. It is that the costs of deferral are not zero, that they compound over time, and that they are systematically undercounted in institutional risk calculations that treat the status quo as a safe default.

The genuine risk management question is therefore not adoption versus restraint. It is what governed adoption looks like, how quickly it can be built, and what the cost of each additional period of delay actually is. That reframing — from whether to how and at what pace — is what a compounding risk analysis requires.

Section 4: The Governed Adoption Alternative

The compounding risk argument is vulnerable to a predictable objection: that it proves too much. If the costs of deferral are real and accumulating, the implied conclusion might seem to be that speed of adoption is the governing criterion — that institutions should move fast and accept governance deficits as the price of competitive relevance. That conclusion would be wrong, and it is not the one this paper advances.

The relevant alternative to skepticism is not uncritical deployment. It is governed adoption — deliberate integration of AI capability within an accountability architecture that makes outputs auditable, failure modes detectable, and human judgment genuinely authoritative over consequential decisions. The distinction matters because it reframes the policy question. The choice is not between safety and capability. It is between two different approaches to managing risk: one that locates risk primarily in the act of adoption, and one that distributes risk management across the design of the system doing the adopting.

The case for governed adoption is not purely theoretical. Practitioners and researchers working at the intersection of AI capability and high-stakes analytical domains have demonstrated that epistemically governed human-AI teaming is achievable with currently available components — not as a projected future state, but as a working operational architecture. The core design principles that make this possible are worth articulating, because they establish that the choice institutions face is not forced.

Human authority by design, not by policy. Governed adoption architectures do not treat human oversight as a compliance requirement layered onto an otherwise autonomous system. They build human judgment into the analytical workflow as a structural precondition for output generation. Before synthesis occurs, the system requires documented human contribution — a substantive analytical act that cannot be bypassed or compressed. This is the mechanism by which human authority remains genuine rather than nominal. The AI component accelerates ingestion, surfaces connections, and generates structured outputs. The human component validates, challenges, and takes responsibility for conclusions. Neither substitutes for the other.

Provenance at the claim level. Governed adoption requires that every analytical output carry a traceable evidentiary chain — from the source document or data point, through the analytical steps that produced the conclusion, to the human judgment that authorized it. This is not primarily a transparency requirement for external audiences, though it serves that function. It is an internal epistemic discipline that makes error detectable and correctable before it propagates. A system that can produce fast, fluent, and wrong outputs without internal mechanisms for detecting wrongness is not a governed system. It is an accelerated liability.

Data sovereignty as an architectural constraint. High-stakes analytical and security domains routinely involve sensitive materials that cannot be externalized to commercial inference infrastructure without unacceptable risk. Governed adoption architectures accommodate this by supporting local inference capability for sensitive workloads, ensuring that analytical augmentation does not require trading away control over the materials being analyzed. This is technically achievable now, with available open-weight models and modest on-premises hardware investment.

Audit independence. Governed adoption requires that the record of analytical activity — what was ingested, what was synthesized, what human judgments were made, and when — be maintained in a layer that is independent of the operational system. This ensures that the accountability record survives system updates, personnel changes, and operational failures. It also creates the evidentiary foundation for after-action review, institutional learning, and external oversight when required.

These principles are not aspirational. They describe architectures that practitioners are building and operating today. Their existence is sufficient to establish that the policy choice institutions face is not between dangerous capability and safe restraint. It is between governed adoption — analytically powerful, epistemically disciplined, human-teaming in its operational logic — and ungoverned deferral. In a deteriorating threat environment, the latter is the more dangerous option.

A recent contribution to this debate from IBM frames the governance question primarily as one of openness versus opacity at the model level — arguing that open-source development is the appropriate response to frontier AI capability because scrutiny produces resilience.[14] That argument has merit as far as it goes. But openness of the base model does not guarantee epistemic governance of the deployment. Anthropic's own red team observes that gaining practice with current models through appropriate scaffolds and procedures is valuable preparation for when Mythos-class capabilities become more broadly available — and that it takes time for people to learn and adopt these tools effectively.[1] A poorly governed open-source deployment compounds risk as readily as a poorly governed closed one. The governance question that matters most is not what the model is, but how institutions use it — and whether the architecture surrounding its use keeps human judgment genuinely authoritative over consequential outputs.

Section 5: Policy Implications

The compounding risk argument, if accepted, has direct implications for how institutions approach AI governance — not as an abstract policy question but as an operational and organizational challenge with measurable costs attached to delay. Five implications are worth drawing out explicitly.

Reframe the default posture. The most immediate implication is institutional. Governance bodies, oversight agencies, and organizational leadership that have treated AI skepticism as the default responsible posture need to reexamine that framing against the threat environment described in this paper. This does not mean abandoning caution. It means recognizing that caution has a duration cost, that the cost is not evenly distributed across the risk landscape, and that risk calculations which treat the status quo as a safe baseline are systematically miscalibrated. The appropriate default posture is not adoption or restraint. It is governed adoption on an accelerated timeline, with accountability architecture built in from the outset rather than retrofitted under pressure.

Invest in epistemic governance infrastructure now. The four principles outlined in the preceding section — human authority by design, provenance at the claim level, data sovereignty as an architectural constraint, and audit independence — are not self-implementing. They require deliberate investment in technical architecture, organizational process, and human expertise. That investment takes time to mature. Institutions that begin building epistemic governance infrastructure now will be positioned to deploy capable AI systems accountably when competitive and operational pressure makes deployment unavoidable. Institutions that defer that investment will face a choice between ungoverned deployment and continued deferral, neither of which is acceptable. The time to build the governance architecture is before the capability pressure becomes acute, not after.

Treat expertise development as a strategic priority. The expertise required to govern AI in high-stakes domains — to design meaningful oversight, detect failure modes, evaluate outputs with appropriate skepticism, and maintain genuine human authority over consequential decisions — does not exist at scale and cannot be developed without engagement. Academic programs, professional development pipelines, and institutional hiring strategies that treat AI governance expertise as a secondary credential are operating on assumptions that no longer hold. In a threat environment where adversarial AI capability is advancing rapidly, the shortage of practitioners who can govern AI deployments responsibly is itself a national security problem. Addressing it requires treating expertise development with the same urgency applied to capability acquisition.

Engage the standards process actively. The governance frameworks, accountability standards, and oversight architectures that will define AI use in critical infrastructure and national security contexts for the next decade are being developed now — in interagency processes, international standards bodies, sector-specific working groups, and allied coordination forums. Institutions that are not actively participating in those processes are not avoiding the governance problem. They are ensuring that the frameworks that emerge reflect the priorities and assumptions of those who did participate. For departments, agencies, and organizations with equities in how AI governance standards develop, active engagement is not optional. It is a form of risk management.

Recalibrate the relationship between offensive awareness and defensive investment. Anthropic's red team documented that Mythos Preview wrote exploits in hours that expert penetration testers said would have taken them weeks to develop, and that the process of turning public vulnerability identifiers into functional exploits — historically taking skilled researchers days to weeks per bug — now happens much faster, cheaper, and without human intervention.[1] That compression of the exploit development timeline changes the defensive calculus fundamentally. The red team's assessment is direct: language models that can automatically identify and exploit security vulnerabilities at large scale could upend the tenuous security equilibrium that has held for the past two decades, and the capabilities future models bring will ultimately require a much broader, ground-up reimagining of computer security as a field.[1] Institutions that frame AI investment as an enhancement to existing defensive posture are underestimating the shift. In an environment where adversarial AI capability is already at or approaching the Mythos threshold, AI-enabled defense is the minimum viable posture, not an upgrade to it.

Taken together, these implications point toward a governance agenda that is urgent without being reckless, ambitious without being naive. The compounding risk argument does not require institutions to abandon accountability in favor of speed. It requires them to recognize that accountability and speed are not in opposition — that building epistemic governance infrastructure is itself the accelerant that makes capable AI deployment safe enough to sustain. The alternative — waiting until governance frameworks are perfect before adopting the capabilities that a deteriorating threat environment demands — is not caution. It is a choice to fall further behind on a timeline that adversaries are not observing.

Conclusions

On the Firefox 147 benchmark, Mythos Preview developed working exploits 181 times compared to just 2 for the previous generation model — a 90-fold improvement in exploit development capability in the span of months.[1] That is not an incremental development. It is a threshold crossing, and it arrived faster than most governance timelines anticipated.

The governance response to threshold crossings of this kind has historically been cautious — to slow down, to study, to defer deployment until accountability frameworks are adequate. That instinct is not irrational. But it depends on a condition that no longer obtains: that the environment being managed is stable enough that deferral preserves rather than erodes the option of safe adoption later.

Anthropic has already privately warned senior government officials that Mythos makes large-scale cyberattacks significantly more likely this year.[13] The adversarial capability frontier is not waiting for governance frameworks to mature.

The compounding risk argument is ultimately a claim about opportunity costs in a competitive environment. Every institutional cycle spent deliberating over AI adoption frameworks is a cycle in which adversarial capability continues to develop, governance expertise fails to accumulate, and influence over the standards process accrues to others. These costs are real, they are measurable in retrospect, and they are systematically undercounted by risk calculations that treat the status quo as a safe baseline.

The conservative position — properly understood, against the actual threat environment — is accelerated governed adoption: deliberate, auditable, human-teaming in its operational logic, and built on accountability architecture that makes outputs trustworthy enough for institutional use. That is a demanding standard. It is also an achievable one. The tools, the architectural principles, and the practitioner knowledge required to meet it exist today.

What is missing is not capability. It is the institutional will to recognize that restraint, in this environment, is itself a risk posture — and not the safer one.

Bibliography

References

1. Carlini, Nicholas, et al. Assessing Claude Mythos Preview's Cybersecurity Capabilities. Anthropic Frontier Red Team. red.anthropic.com/2026/mythos-preview/. April 7, 2026.
2. CISA, NSA, FBI et al. PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure. Joint Cybersecurity Advisory AA24-038A. February 7, 2024. cisa.gov/news-events/cybersecurity-advisories/aa24-038a.
3. CISA, NSA, FBI et al. PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure. Joint Cybersecurity Advisory AA24-038A. PDF version. media.defense.gov. February 7, 2024.
4. Nextgov/FCW. "Salt Typhoon Hackers Targeted Over 80 Countries, FBI Says." August 27, 2025. Citing statements by FBI Cyber Division Assistant Director Brett Leatherman. nextgov.com/cybersecurity/2025/08/salt-typhoon-hackers-targeted-over-80-countries-fbi-says/407719/.
5. Nextgov/FCW. "Hundreds of Organizations Were Notified of Potential Salt Typhoon Compromise." December 2024. nextgov.com/cybersecurity/2024/12/hundreds-organizations-were-notified-potential-salt-typhoon-compromise/.
6. TechCrunch. "FBI Says China's Salt Typhoon Hacked at Least 200 US Companies." August 27, 2025. Citing FBI Assistant Director Brett Leatherman statements to The Washington Post. techcrunch.com/2025/08/27/fbi-says-chinas-salt-typhoon-hacked-at-least-200-us-companies/.
7. Dark Reading. "CISA Issues Guidance to Telecom Sector on Salt Typhoon." December 2024. Citing CISA Executive Assistant Director Jeff Greene. darkreading.com/cyberattacks-data-breaches/cisa-issue-guidance-telecoms-salt-typhoon-threat.
8. CISA, FBI, NSA, EPA, INCD, CCCS, NCSC-UK. IRGC-Affiliated Cyber Actors Exploit PLCs in Multiple Sectors, Including U.S. Water and Wastewater Systems Facilities. Joint Cybersecurity Advisory AA23-335A (updated December 18, 2024). cisa.gov/news-events/cybersecurity-advisories/aa23-335a.

9. EPA, FBI, CISA, NSA. Joint Advisory on Iranian-Affiliated Cyber Threats to U.S. Water Systems. December 2023. epa.gov/newsreleases/epa-fbi-cisa-nsa-issue-joint-cybersecurity-advisory-water-system-regarding-iranian.
10. CISA, FBI, NSA, EPA, DOE, U.S. Cyber Command. Iranian-Affiliated Cyber Actors Exploit Programmable Logic Controllers Across U.S. Critical Infrastructure. Joint Cybersecurity Advisory AA26-097A. April 7, 2026. cisa.gov/news-events/cybersecurity-advisories/aa26-097a.
11. Anthropic. Project Glasswing Initiative. anthropic.com/project/glasswing. April 7, 2026.
12. Anthropic. Claude Mythos Preview System Card. 244 pp. April 7, 2026. Key citations: release decision rationale, p. 12; alignment risk framing ("best-aligned / greatest alignment risk"), p. 53.
13. Fortune. "Anthropic is Giving Some Firms Early Access to Claude Mythos to Bolster Cybersecurity Defenses." April 7, 2026. fortune.com/2026/04/07/anthropic-claude-mythos-model-project-glasswing-cybersecurity/.
14. Thomas, Rob. "Open Source, After Mythos." IBM Newsroom. April 9, 2026. newsroom.ibm.com/2026-04-09-Open-Source,-After-Mythos.
15. CISA, NSA, FBI et al. Countering Chinese State-Sponsored Actors Compromise of Networks Worldwide to Feed Global Espionage Systems. Joint Cybersecurity Advisory AA25-239A. August 27, 2025. cisa.gov/news-events/cybersecurity-advisories/aa25-239a.
16. CISA, NSA, FBI et al. People's Republic of China State-Sponsored Cyber Activity: Actions for Critical Infrastructure Leaders. Joint Fact Sheet. March 19, 2024. cisa.gov/news-events/alerts/2024/03/19/cisa-and-partners-release-joint-fact-sheet-leaders-prc-sponsored-volt-typhoon-cyber-activity.

Dr. David Mussington

Fellow, Institute for Critical Infrastructure Technology (ICIT), Co-Chair, ICIT's Center for FCEB Resilience, Professor of the Practice at the University of Maryland's School of Public Policy

Dr. David Mussington is a Fellow of the Institute for Critical Infrastructure Technology (ICIT) and Co-Chair of ICIT's Center for FCEB Resilience. Additionally, he is a Professor of the Practice at the University of Maryland's School of Public Policy. Prior to rejoining UMD in January of 2025, David served as the Executive Assistant Director for Infrastructure at the Cybersecurity and Infrastructure Agency, U.S. Department of Homeland Security. At CISA, David was one of three presidentially appointed officials charged with implementing the nation's critical infrastructure security and resilience strategies and plans across 16 critical infrastructures. He also led interagency efforts on counter- and anti- terrorism efforts, playing a leading role in reducing the risks of domestic targeted violence, school safety, and physical infrastructure security standards. He was also a founding member of CISA's Cyber Safety Review Board. David has extensive public and private sector experience in cyber and infrastructure security, selected for the Senior Executive Service and assigned to the Office of the Secretary of Defense in the role of Senior Advisor for Cyber Policy, later joining the NSC staff as Director for Surface Transportation Security Policy. As a researcher at RAND Corporation and later at the Institute for Defense Analyses, David directed cybersecurity studies for the Department of Homeland Security (DHS), the Office of the Director of National Intelligence (ODNI), the Federal Communications Commission, the Bank of Canada, and NATO. David has a Ph.D. in Political Science from Canada's Carleton University, and M.A. and B.A. degrees from the University of Toronto. He undertook postdoctoral study at Harvard's Belfer Center and at the UK's International Institute for Strategic Studies. In 2021 David was elected a life member of the Council on Foreign Relations. In 2023 David was awarded Homeland Security Today's Mission Award, for contributions to the U.S. Critical Infrastructure Security and Resilience mission. In 2024 he received the 2024 Impact Award from the Institute for Critical Infrastructure Technology (ICIT) for leadership in critical infrastructure policy and strategy.

ICIT

www.icitech.org