May 2026

The Harness Gap

Orchestration, Defensive Parity, and the Closing Window for Critical Infrastructure AI 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

Executive Summary	3
Key Terms and Acronyms	5
Introduction: A Policy Response and Its Structural Assumption	7
Section 2: The Pipeline Decomposition	8
Section 3: The Defensive Inversion	11
Section 4: The Overdetermination Problem	12
Section 5: The ICS Window — Governance Timing as the Central Variable	14
Section 6: A Topology-Aware Governance Architecture	17
Section 7: Governance Recommendations	19
Conclusion	21
Bibliography	22
About	23

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

Institute for Critical Infrastructure Technology

Executive Summary

The United States critical infrastructure faces a threat environment that is, in the precise technical sense, already compromised. Volt Typhoon has burrowed into energy, water, transportation, and communications networks. Salt Typhoon has penetrated every major telecommunications carrier. Iranian actors have disrupted water and power control systems at operational scale. Against this baseline, the April 2026 announcement of Anthropic's Claude Mythos Preview — and the concurrent publication of Glasswing, its gated access program — prompted a predictable governmental response: restrict the model, restrict the capability.

This paper argues that the restriction instinct, however understandable, targets the wrong chokepoint. The capability Mythos exemplifies is not monolithically located in any model. It is a four-layer pipeline — vulnerability detection, broad-spectrum scanning, exploit chain construction, and end-to-end attack chain execution — whose first three layers are already closed, or closable by minimal orchestration, using open-weight models available today without restriction. The AgentFlow result, published in peer-reviewed research in April 2026, established this empirically: a mid-tier open-weight model wrapped in a synthesized multi-agent harness produced ten vendor-confirmed zero-day vulnerabilities in Google Chrome, including two Critical sandbox-escape CVEs. Mythos-class offensive output, delivered without Mythos.

The same pipeline logic governs the defensive side. If offensive AI capability is a (model × harness × compute) function, then defensive AI capability is too — and the policy question is not how to restrict the offensive pipeline but whether the defensive harness can close the detection-to-response cycle faster than the adversary's offensive harness closes the one remaining capability gap: the ICS/OT attack surface.

That gap is genuine. Frontier models average 1.3 of 7 steps on realistic ICS attack scenarios as of early 2026. It is also temporary. The research program that produced AgentFlow is directly applicable to ICS targets and is economically incentivized to go there. The window during which governance investment can establish defensive parity in ICS/OT environments before the offensive gap closes is open. It is not open indefinitely.

This paper advances four findings and a set of governance recommendations calibrated to the race condition, not the steady state.

Finding 1

The model-centric theory of AI cyber capability governance is substantially falsified by the orchestration evidence. Governing the model tier while leaving harness architecture ungoverned fails as a general governance architecture: it does not intercept the capability at the layers where it actually resides. A thin residual Mythos-class capability exists at the extreme frontier of Layer 3, but it does not restore the model-access chokepoint as a viable governance instrument for the vast majority of the offensive task class. It is a structural mismatch between the governance instrument and the governance target.

Finding 2

The defensive pipeline is symmetric with the offensive pipeline. Defensive harness engineering — autonomous triage, cross-domain correlation, containment initiation — can achieve operational parity. But only if it reaches the operators who need it most.

Finding 3

The current Glasswing architecture and the existing agentic governance frameworks share a structural failure: they concentrate capability at the top of the operator size distribution, leaving Tier 3 critical infrastructure operators — defined in Section 4 as organizations with genuine OT exposure but minimal dedicated cybersecurity capacity — unserved at precisely the moment of maximum risk.

Finding 4

The ICS/OT attack surface is the only remaining domain where governance investment can establish monitoring conditions before the capability threshold is crossed. The correct governance priority is a standardized ICS adversarial benchmark with a defined escalation trigger — not model access restriction that is irrelevant to the three pipeline layers already closed.

Key Terms and Acronyms

Term	Definition
ABOM	Agent Bill of Materials. A structured governance instrument that formally inventories an agentic deployment's decision authorities, tool access scopes, Behavior Certificates, model provenance, and precommitment constraints.
AgentFlow	A synthesized multi-agent harness architecture (Liu et al., arXiv 2604.20801, April 2026) that demonstrated Mythos-class offensive capability using a mid-tier open-weight model, producing ten vendor-confirmed zero-day vulnerabilities including two Critical CVEs in Google Chrome.
CISA	Cybersecurity and Infrastructure Security Agency. The U.S. federal agency with primary responsibility for protecting the nation's critical infrastructure.
CVE	Common Vulnerabilities and Exposures. A standardized identifier assigned to publicly disclosed cybersecurity vulnerabilities.
CVSS	Common Vulnerability Scoring System. An open framework for communicating the severity of software vulnerabilities, producing a numerical score from 0 to 10.
Glasswing	Anthropic's gated access program for Claude Mythos Preview, launched concurrently with the model's April 2026 announcement. Launch partners include major cloud providers, technology firms, and financial institutions — all Tier 1 operators by the classification used in this paper.
Harness	The orchestration architecture that wraps a language model: the system of task decomposition, tool access, memory management, feedback loops, policy gates, and inter-agent communication that converts raw model capability into directed pipeline output. As used in this paper, the primary unit of governance analysis.
ICS	Industrial Control System. Computer-based systems used to monitor and control industrial processes, including in energy, water, and transportation sectors.
ISAC	Information Sharing and Analysis Center. Sector-specific organizations that facilitate sharing of cybersecurity threat intelligence among critical infrastructure operators.
IT	Information Technology. Computing systems used for data processing and communications, as distinct from Operational Technology.
MoE	Mixture-of-Experts. A neural network architecture in which only a subset of model parameters ("experts") is activated for any given input, enabling very large total parameter counts while keeping per-token inference costs low. DeepSeek-V4-Pro uses this architecture, activating 49B of 1.6T total parameters per token.
Mythos	Claude Mythos Preview. Anthropic's frontier AI model announced April 2026, capable of autonomous discovery of previously unknown software vulnerabilities at rates that substantially exceed human security team capacity. The governance trigger for this paper's analysis.
Open-weight model	An AI model whose trained weights are publicly released and available for local deployment, without API access controls, usage restrictions, or licensing fees that would limit distribution. DeepSeek-V4-Pro is the primary open-weight model referenced in this paper's analysis of the fifth pathway.

OT	Operational Technology. Hardware and software that monitors or controls physical devices, processes, and infrastructure — including industrial control systems and programmable logic controllers.
PLC	Programmable Logic Controller. A ruggedized industrial computer used to automate electromechanical processes in manufacturing, utilities, and critical infrastructure.
SOC	Security Operations Center. A facility housing an information security team responsible for monitoring, detecting, and responding to cybersecurity threats.

Introduction: A Policy Response and Its Structural Assumption

Governing powerful technologies under uncertainty requires making bets. The question is not whether to bet. It is whether the bet you are making is calibrated to the actual risk.

In early May 2026, Vice President JD Vance convened a call with the chief executives of Anthropic, OpenAI, Google, Microsoft, and SpaceX. The precipitating event was Anthropic's April release of its Claude Mythos Preview model — a system that had demonstrated the ability to autonomously discover thousands of high- and critical-severity software vulnerabilities, including previously unknown zero-days in production code dating back decades. The administration, which had revoked the Biden AI safety executive order within hours of taking office, was now contemplating a mandatory pre-release vetting regime modeled on the Food and Drug Administration's pre-market approval process.

The policy instinct is understandable. Mythos represents a visible capability threshold event, and visible threshold events demand a response. The Mythos Preview System Card documented that the model autonomously discovered CVE-2026-4747 — a 17-year-old FreeBSD NFS remote code execution flaw granting unauthenticated root access — and chained exploit sequences across large codebases at rates no human security team can match. On the Firefox 147 benchmark, Mythos Preview developed working exploits 181 times compared to just two for the previous generation model — a 90-fold improvement in the span of months. That is not an incremental development. It is a threshold crossing. The instinct to restrict the model that crossed it is not irrational. It is simply aimed at the wrong target.

This paper argues that the policy response the threshold crossing has generated — mandatory pre-release vetting of frontier models modeled on FDA pre-market approval — is calibrated to the wrong chokepoint. The capability the vetting regime targets is not monolithically located in the Mythos model. It is distributed across a four-layer pipeline, and the first three of those layers are already closed — or closable with minimal orchestration effort — using open-weight models available today without restriction. Restricting Mythos does not close the pipeline. It governs a component that is not the binding constraint. The pipeline runs without it.

Section 2: The Pipeline Decomposition

The capability that Mythos exemplifies — end-to-end autonomous vulnerability discovery and exploit chain construction — is not atomic. It decomposes into four functionally distinct sub-tasks, each with a different closure mechanism, a different governance leverage point, and a different relationship to the model tier that the vetting regime proposes to restrict.

Layer 1: Vulnerability Detection in Isolated Code Segments

The first sub-task is vulnerability detection in isolated code segments: identifying potential security flaws in bounded, well-defined code. This capability is fully distributed across all code-trained language models and has been embedded in commercial tooling — GitHub Copilot, Amazon CodeWhisperer, Cursor, and dozens of others — at production scale. The governance chokepoint the vetting regime targets does not exist at this layer. Restricting Mythos has no effect on a capability already embedded in the development environment of every professional software engineer. Layer 1 was closed before Mythos existed.

Layer 2: Broad-Spectrum Scanning

The second sub-task is broad-spectrum scanning across large production codebases: applying vulnerability detection at scale to real-world repositories. The AISLE System Over Model analysis established that this capability is closed by orchestration. A parallel agent harness — multiple instances of a code-trained model operating concurrently across different segments of a large repository — matches frontier performance on broad-spectrum scanning tasks when the model itself is mid-tier. Access restriction at the frontier model level is irrelevant to a capability whose binding constraint is harness architecture, not model capability. Layer 2 was closed by orchestration, not by frontier models.

Layer 3: Exploit Chain Construction

The third sub-task is exploit chain construction and iterative verification: taking identified vulnerabilities and developing working exploit chains, including chaining multiple vulnerabilities to achieve higher-impact outcomes. The AgentFlow result — arXiv 2604.20801, Liu et al., April 2026 — established that this capability is substantially closed by synthesized multi-agent harness design. A mid-tier open-weight model, paired with a harness designed to decompose exploit chain construction into structured sub-tasks with feedback-driven iteration, produced ten previously unknown zero-day vulnerabilities in Google Chrome, including CVE-2026-5280 and CVE-2026-6297, two Critical sandbox-escape vulnerabilities confirmed by the vendor.

A thin residual gap remains after orchestration at Layer 3. The highest-complexity creative exploit chains — those requiring sustained strategic synthesis across dozens of iterative steps while maintaining coherent architectural reasoning about a complex constraint space — remain Mythos-class. But this residual is narrow. The vast majority of exploit reasoning and chain construction tasks, including the zero-day class that motivated the Mythos policy response, are achievable by a mid-tier open-weight model paired with a well-designed harness

The AgentFlow result established Layer 3 closure via harness amplification of sub-frontier models — a mid-tier model made capable by competent orchestration. DeepSeek-V4-Pro, an open-weight model with 1.6 trillion parameters released in May 2026, reinforces this finding by demonstrating that frontier-class coding capability has now arrived at the open-weight tier in its own right. V4-Pro achieves a Codeforces competitive coding rating of 3,206, matching or marginally leading the frontier closed models referenced in this paper's analysis (GPT-5.4 at 3,168; Claude-Opus-4.6-adjacent models at 3,052–3,168). On the benchmark class most directly relevant to exploit chain construction — sustained reasoning on novel, complex coding problems — open-weight capability has arrived at the frontier. The Layer 3 residual gap cannot be defended at the model tier because the model tier is no longer a restriction point. DeepSeek-V4's architectural design — FP4 quantized expert weights, dramatically reduced per-token inference cost, designed for deployment on commodity hardware — means this capability is accessible without frontier-scale infrastructure.

Two mechanisms of closure now operate simultaneously: harness amplification brings mid-tier models to Layer 3 capability, while open-weight frontier capability has arrived on its own terms. V4-Pro's performance addresses the thin Mythos-class residual at Layer 3 — the narrow class of highest-complexity chains that harness amplification of sub-frontier models alone had not fully closed. Neither mechanism is intercepted by restricting Mythos. For the target classes relevant to Layers 1 through 3, the model tier is no longer a restriction point; model-centric governance retains only the narrowest residual leverage at the extreme frontier, which is insufficient to serve as the basis for a general governance architecture.

Layer 4: End-to-End Attack Chain Execution

The fourth sub-task is end-to-end attack chain execution: navigating from initial access to mission completion across a realistic multi-step adversarial sequence. The cyber range evaluation of frontier AI models (arXiv 2603.11214, Folkerts et al., March 2026) provides the most structured measurement of this capability across two target domains, and the result generates the paper's central governance distinction.

For corporate network targets, the evaluation deployed seven frontier models across a 32-step attack scenario over an 18-month period from August 2024 to February 2026. Performance scaled log-linearly with inference-time compute — increasing from 10 million to 100 million tokens yielded gains of up to 59 percent — with no observed plateau. The best single run of Claude Opus 4.6 completed 22 of 32 steps, representing approximately six of the estimated 14 hours a human expert would require for equivalent work. Critically, scaling inference budget required no specific technical sophistication from the operator.

For ICS/OT targets, the same evaluation deployed models against a seven-step industrial control system attack. Results were materially different: frontier models averaged 1.3 of 7 steps across the evaluation period, with even the most recent models — the first to reliably complete any steps — averaging only 1.2 to 1.4 steps. The ICS attack surface requires domain-specific capabilities — protocol heterogeneity, physical-layer dependency mapping, vendor-specific PLC knowledge — that have not yet been addressed by generic harness optimisation.

[1] Benchmark comparisons reference DeepSeek-AI's May 2026 technical report. The named comparator model versions (GPT-5.4, Claude-Opus-4.6-Max, Gemini-3.1-Pro-High) are not independently publicly verifiable at the time of this amendment and should be read as indicative of competitive positioning rather than confirmed absolute scores.

The two-domain result generates the paper's central governance distinction. For user-space and corporate network targets, end-to-end autonomous capability is effectively achieved by current frontier models, and substantially approached by mid-tier open-weight models under synthesised harness architectures. For ICS/OT targets, a meaningful gap remains — the only genuine residual capability gap in the four-layer pipeline.

Pipeline Status Summary:

Layer	Sub-Task	Closure Mechanism	Governance Leverage Point	Status
1	Vulnerability detection in isolated code segments	Capability fully distributed across all code-trained models; embedded in commercial tooling at scale.	None. Governance chokepoint does not exist at this layer.	CLOSED
2	Broad-spectrum codebase scanning across large production repositories	Closed by orchestration. Open-weight model + parallel agent harness matches frontier performance.	Harness architecture, not model access. Access restriction is irrelevant.	CLOSED
3	Exploit chain construction and iterative verification	Substantially closed by synthesised multi-agent harness (AgentFlow result, Apr 2026). Thin residual for highest-complexity chains. Open-weight frontier-class coding capability (V4-Pro) now matches closed frontier on coding benchmarks — model tier no longer a restriction point.	Harness topology design. Thin Mythos-class residual does not restore model-access governance leverage.	SUBSTANTIALLY CLOSED
4a	End-to-end attack chain execution — corporate IT / user-space	Frontier models complete 22/32 steps; performance scales log-linearly with compute. No plateau observed. Inference cost reductions (DeepSeek-V4 class) amplify this by making larger token budgets economically accessible.	Defensive harness uplift. Scaling inference budget is an economic decision, now cheaper.	CLOSED
4b	End-to-end attack chain execution — ICS/OT targets	Genuine gap. Frontier models average 1.3/7 steps. Domain-specific protocol, PLC, and physical-layer knowledge not yet available. Lower inference costs accelerate the economics of closing this gap.	The only tractable governance target. ICS benchmark + defensive harness deployment before gap closes.	GAP — CLOSING

Section 3: The Defensive Inversion

The pipeline decomposition establishes the offensive capability topology. It also, by structural symmetry, establishes what effective agentic defense requires — and exposes the gap between what current defensive frameworks provide and what the threat environment actually demands.

If offensive AI capability is a (model × harness × compute) function, then defensive AI capability is the same function applied to a different task class. The offensive pipeline — vulnerability detection, broad-spectrum scanning, exploit chain construction, attack chain execution — has a defensive mirror: anomaly detection, cross-environment correlation, threat hypothesis generation, and automated containment initiation. The architectural requirements are structurally similar: multi-agent orchestration, feedback-driven iteration, domain-specific scaffolding, and persistent context across extended operational sequences. The economics cut both ways. V4-class inference efficiency is as available to defenders constructing defensive harness pipelines as it is to adversaries constructing offensive ones. The race condition is symmetric. The question is who has the harness first and who reaches the operators who need it most.

The current defensive AI governance landscape does not reflect this symmetry. CISA's guidelines, ISACA's agentic AI security frameworks, and the AWS Agentic AI Security Scoping Matrix — the most developed publicly available frameworks as of the time of writing — are organized around action risk classification and access control at the model level. They treat the model as the primary governance object and the harness as a secondary deployment detail. This is precisely the misspecification the pipeline decomposition identifies on the offensive side, applied symmetrically to defense. The defensive frameworks are organized around the wrong unit of analysis.

A harness-aware defensive architecture would look different. It would treat the defensive (model × harness × compute) triple as the unit of governance analysis. It would ask not which model a critical infrastructure operator has access to, but what orchestration architecture surrounds that model, what context management it employs, what policy gates govern its output, and whether its inference budget is sufficient to close the detection-to-response cycle before the adversary's offensive pipeline completes its attack sequence.

The operators most exposed to the adversarial pipeline are Tier 3 critical infrastructure operators — organizations with genuine operational technology exposure but minimal dedicated cybersecurity capacity: regional water utilities, municipal power systems, rural hospitals, mid-sized manufacturing facilities with networked industrial control systems. These are the organizations for which defensive harness engineering is most consequential and least available. They are also the organizations that are not in Project Glasswing.

Section 4: The Overdetermination Problem

The vetting regime has a deeper structural problem than the orchestration finding alone establishes. Suppose, for a moment, that the model-centric theory of capability governance were correct — that Mythos were genuinely the primary capability source, that restricting it would meaningfully intercept the offensive pipeline. Even then, the governance architecture proposed would be insufficient. The relevant capability is overdetermined: it would remain available through independent pathways even if the specific pathway targeted by restriction were successfully closed.

The overdetermination literature identifies four pathways to domain-level AI cyber capability availability: open-weight scaling (increasingly capable models released without access restrictions); deliberate cyber-specific fine-tuning (training or adapting models explicitly for offensive cyber tasks); neurosymbolic scaffolding (G-CTR-class systems combining learned and programmatic reasoning); and state-actor adversarial distillation via military compute facilities (extracting capability from restricted frontier models through systematic interaction). The Vaynman-Volpe governance dead zone analysis establishes that when a capability is overdetermined in this sense — achievable through multiple independent pathways each sufficient on its own — governance regimes targeting individual pathways are structurally inadequate. Interdicting one pathway shifts activity to others without reducing overall capability availability.

4.5 The Fifth Pathway: Harness-Mediated Capability Reconstruction

The orchestration finding establishes a fifth pathway that the existing taxonomy does not capture: harness-mediated capability reconstruction from API-accessible or openly available precursor models. This pathway requires neither state-level compute nor deliberate fine-tuning for offensive purposes. It requires organisational capability — the ability to design, implement, and operate a multi-agent pipeline — and inference budget, both of which are widely distributed. The barrier is organisational, not technical, and it is declining as harness engineering becomes a named discipline with a maturing toolchain.

A governance architecture calibrated to the existing four pathways — compute export controls, model vetting, open-weight restrictions — has no instrument targeting the fifth pathway. The fifth pathway is also the pathway that the public evidence most clearly demonstrates has already been exploited operationally: the AgentFlow result is not a red-team exercise; it is a production vulnerability discovery campaign that generated vendor-confirmed Critical CVEs against a major browser codebase. The fifth pathway is not theoretical. It is operational.

DeepSeek-V4-Pro (May 2026) adds a concrete and material dimension to the fifth pathway analysis. It is a 1.6 trillion parameter open-weight model, released on HuggingFace without access restrictions, achieving frontier-class performance on competitive coding benchmarks. Its architectural design — FP4 quantized mixture-of-experts weights, 10% of the KV cache of prior-generation models at one-million-token context, designed for efficient deployment on commodity hardware without frontier-scale infrastructure — materially lowers the compute variable in the fifth pathway's (model × harness × compute) triple for actors outside the frontier compute infrastructure tier.

The geopolitical dimension is not incidental. DeepSeek-V4 is PRC-origin. The state-sponsored actors named in this paper — Volt Typhoon, Salt Typhoon, IRGC-affiliated actors — now have access to a frontier-coding-capable open-weight model that can be run on-premise, fine-tuned without API restrictions or usage logging, and wrapped in purpose-built harness architectures without export control exposure (see Layer 3 discussion above, noting V4-Pro's Codeforces competitive coding rating of 3,206, matching or leading the named closed frontier models). The harness-mediated capability reconstruction pathway, which requires organisational rather than state-level capability, is now accessible with a model tier that matches the closed frontier on the tasks most directly relevant to Layers 1 through 3 of the offensive pipeline. The fifth pathway's organisational barrier is lower; the capability ceiling it delivers is higher.

This does not require a new finding. It sharpens Finding 1: the governance architecture has no instrument targeting the fifth pathway, and the fifth pathway's primary model resource is now openly available, efficiently deployable, and PRC-origin.

Section 5: The ICS Window — Governance Timing as the Central Variable

The pipeline decomposition identifies a single domain where the governance problem remains tractable on a governance-determined timeline rather than an adversarially determined one: the ICS/OT attack surface. This is the only place where defensive investment can establish parity before the offensive gap closes. Everything about the governance response to Mythos should be calibrated to this finding. The window is real. It is open. It is not open indefinitely.

The Gap Is Genuine

The ICS/OT capability gap is not a policy artifact or an optimistic projection. Frontier models averaged 1.3 of 7 steps on a realistic ICS attack scenario as of early 2026 — the Folkerts et al. cyber range evaluation, the most structured measurement available. The constraint is not compute or model capability in the general sense; it is domain-specific knowledge that current training has not provided. Industrial control systems present a distinctive attack surface. Modbus, DNP3, EtherNet/IP, PROFINET, and the dozens of other protocols governing OT communication are not well-represented in the training corpora of general-purpose language models. Physical-layer dependency reasoning, vendor-specific PLC knowledge, and the multi-layer architecture of process control environments are poorly represented in the open literature from which model training draws. The gap is real. It is also the only gap that remains.

This is not a permanent limitation. The research program that produced AgentFlow demonstrated this on the corporate IT surface: what appeared to be a frontier-model-only capability became achievable by a mid-tier open-weight model when the harness provided the domain-specific scaffolding that raw model capability lacked. The same approach, applied to ICS targets, will eventually produce the same result.

The Gap Is Temporary

The economic and adversarial incentives to close the ICS gap are substantial and accelerating. The Mythos announcement created a global research and development race in AI-enabled vulnerability discovery. That race is not confined to user-space targets. State-sponsored adversaries with strategic interests in pre-positioned ICS access — the PRC actors behind Volt Typhoon, the IRGC-affiliated actors behind the water sector campaigns — have both the resources and the strategic motivation to direct AI-enabled attack chain development toward OT targets. The question is not whether the gap will close. It is whether governance investment will establish defensive parity in ICS environments before it does.

DeepSeek-V4's architecture adds a specific economic mechanism to the closing-gap argument. The Folkerts et al. log-linear compute scaling finding — that increasing inference token budget from 10 million to 100 million tokens yields up to 59 percent performance gains on corporate IT attack scenarios, with no observed plateau — establishes that the compute variable in the (model × harness × compute) triple is the primary lever for closing remaining capability gaps. DeepSeek-V4's architectural advances reduce per-token inference cost materially: at one-million-token context, V4-Pro requires only 27% of the single-token FLOPs and 10% of the KV cache of its predecessor (DeepSeek-V3.2). For the Flash variant, this reaches 10% of FLOPs and 7% of KV cache.

The governance implication is precise and requires careful framing. V4-class efficiency improvements do not directly narrow the ICS gap today: the gap is knowledge-constrained rather than compute-constrained, and cheaper inference does not supply the domain-specific ICS protocol knowledge, physical-layer dependency reasoning, and PLC vendor expertise that frontier models currently lack.

The mechanism is second-order. V4-class efficiency reduces the lag between when adversaries acquire ICS-specific harness capability and when they acquire ICS operational capability. Once domain-specific harness development addresses the knowledge gap — a timeline driven by adversarial investment, not policy deliberation — the 10M-to-100M-token scaling regime that produced 59 percent gains on corporate IT scenarios becomes cheaper to operate against ICS targets. The window during which defensive governance can establish ICS parity is defined partly by the timeline of ICS-specific harness development, and partly by the economics of compute scaling once that harness exists. V4-class efficiency improvements accelerate the second of these two timelines without touching the first: when the knowledge threshold is crossed, operational capability will arrive faster and more cheaply than earlier estimates assumed.

The Monitoring Condition

The IAEA safeguards system — the institutional infrastructure that gives the Nuclear Non-Proliferation Treaty its enforcement teeth — was designed before nuclear technology was fully proliferated, during a window in which verification was feasible. The system's longevity derives not from its ability to prevent nuclear capability development, which it cannot guarantee, but from its ability to characterize the capability surface and establish monitoring conditions that trigger governance escalation when specific thresholds are approached.

AI cyber governance is attempting to build verification mechanisms after the technology has already diffused to the point where frontier-class capability is demonstrable from publicly available components. This changes the character of verification. The task is not to confirm that a capability does not exist. It is to characterize the capability surface and establish monitoring conditions that trigger governance escalation when specific thresholds are crossed.

The ICS completion rate on a standardized adversarial benchmark is a concrete instance of such a monitoring condition. As of May 2026, no publicly accessible ICS/OT-fidelity cyber range with orchestrated multi-agent AI capability testing exists at operationally meaningful task specificity. This absence is itself a governance gap. Establishing that infrastructure is the most tractable governance investment available to policymakers operating with limited institutional capacity. A standardized ICS adversarial benchmark — with CVSS-relevant task specifications, replicable methodology, and public accessibility — would allow systematic monitoring of AI capability development against ICS targets. When the average completion rate crosses a defined threshold — the author proposes 4 of 7 steps on a CVSS-relevant ICS scenario as a reasonable trigger condition — it constitutes a governance escalation event requiring defensive deployment posture to shift accordingly.

Establishing this monitoring condition before the threshold is crossed is qualitatively different from responding to the threshold crossing after the fact. The original Mythos announcement demonstrated this asymmetry: the governance conversation about pre-release vetting began after the capability threshold had already been crossed and diffused into published research. Building equivalent infrastructure for ICS AI capability requires acting in the current window — while the gap is still measurable, while baseline completion rates are still sub-threshold, and while defensive investment can establish parity in advance of the offensive capability arrival.

Section 6: A Topology-Aware Governance Architecture

The preceding analysis converges on a governance architecture that is structurally different from both the current Glasswing approach and the agentic governance frameworks now in circulation. It is not a repudiation of those frameworks. Their authorization architecture — Agent Bills of Materials, Behavior Certificates, precommitment constraints, Scope 1–4 deployment taxonomy — remains correct as far as it goes. The problem is not that existing frameworks are wrong. It is that they are aimed at the wrong object.

Reframing 1: From Model-Centric to Harness-Aware Governance

The most direct implication of the sub-task decomposition is a reframing of the governance object. The contemplated vetting regime treats the model as the unit of governance analysis: a model is vetted, approved, or restricted. The orchestration finding establishes that the model is not the capability. The capability is the (model × harness × compute) triple, and for the target classes that matter most — large source-available codebases, corporate network attack surfaces — the model tier is not the binding constraint.

A harness-aware governance architecture would require at minimum: (1) classification of orchestration architectures by capability profile, distinguishing general-purpose software engineering pipelines from purpose-built vulnerability discovery systems; (2) monitoring of publicly available harness designs against an evolving capability taxonomy; and (3) integration of harness capability assessment into any pre-release evaluation regime for frontier models, so that evaluation considers what a model can do when wrapped in a competent harness, not merely what it does in direct API mode.

The practical challenge is formidable: orchestration architectures are software artefacts, not model weights. They are reproducible, modifiable, and not subject to weight-export-style controls. A harness-aware regime would need to operate through different instruments — mandatory disclosure of orchestration architectures used in high-consequence deployments, capability assessment requirements for autonomous cybersecurity systems, and liability frameworks that attach to the operator of the (model × harness) system rather than exclusively to the model developer.

Reframing 2: From Access Restriction to Defensive Parity

Project Glasswing represents an implicit recognition of the access-restriction problem: if Mythos-class capability cannot be reliably denied to adversaries through model restriction, the relevant policy question shifts to whether defenders can access equivalent capability before adversaries do. The Glasswing partner list — Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, Nvidia, Palo Alto Networks — is an attempt to operationalize this insight through a controlled access program.

But the Glasswing architecture reproduces the access asymmetry it is designed to address. The twelve launch partners are exclusively American. No European government, no European cybersecurity agency, no allied national CERT, no Tier 3 critical infrastructure operator — water utilities, regional hospitals, municipal power systems — is in the program. The operators most exposed to the adversarial pipeline are precisely the operators for whom the defensive capability uplift is most consequential. They are unserved by the current architecture. Glasswing protects the protected.

Reframing 3: From Steady-State to Race-Condition Governance

The ICS window finding implies a third reframing: from governance calibrated to a steady-state threat environment to governance calibrated to a race condition with a defined timeline. The steady-state governance instinct — study, deliberate, build consensus, implement incrementally — is appropriate when the gap between capability development and governance response is not itself a source of risk. It is inappropriate when the gap is the primary risk variable.

The ICS window is a race condition. The offensive capability gap will close; the timeline is uncertain but bounded by adversarial investment and research momentum. The defensive capability requires proactive deployment before the offensive capability arrives, because deploying defensive harness capability in Tier 3 operators after the offensive threshold is crossed is qualitatively harder than deploying it before. The governance priority — establishing the ICS benchmark monitoring condition and beginning defensive harness deployment in the most exposed Tier 3 operator population — is defined by the race condition timeline, not by the steady-state deliberation calendar.

Section 7: Governance Recommendations

The three reframings in the preceding section generate a set of governance recommendations calibrated to the race condition. They are sequenced by urgency — defined by proximity to the ICS window timeline — rather than by political tractability.

Recommendation 1: Establish an ICS AI Capability Benchmark (Immediate)

The most urgent governance investment is the establishment of a standardized, publicly accessible ICS/OT adversarial benchmark with CVSS-relevant task specifications and a defined escalation trigger. The benchmark should:

- Define a seven-step ICS attack scenario — initial access, reconnaissance, lateral movement, persistence, manipulation of control system, safety system bypass, and physical consequence initiation — using realistic OT protocols (Modbus, DNP3, EtherNet/IP, PROFINET) at an operationally meaningful fidelity level.
- Establish 4 of 7 steps as the escalation trigger: the completion rate at which the governance posture must shift from monitoring to active defensive deployment.
- Require public reporting of model performance on the benchmark by any organization seeking pre-release vetting approval for a frontier model.
- Assign institutional stewardship to CISA, in partnership with the national laboratories with OT security expertise (INL, PNNL, Sandia), with an independent scientific advisory board to maintain benchmark fidelity as adversarial capability develops.

Recommendation 2: Mandate Harness Capability Assessment in Pre-Release Vetting (Near-Term)

If the administration proceeds with a pre-release vetting regime, the regime should be redesigned to assess (model × harness × compute) capability rather than model-only capability. Concretely, this means:

- Evaluators should assess frontier models wrapped in competent harness architectures — multi-agent, tool-enabled, feedback-driven — not merely in direct API mode.
- The evaluation should include the ICS benchmark as a required capability assessment, not only the user-space and corporate network tasks that current safety evaluations address.
- Pre-release vetting should require disclosure of the harness architectures the developer has tested the model against, and the completion rates achieved on standardized capability benchmarks.

Recommendation 3: Fund Defensive Harness Deployment for Tier 3 Operators (Near-Term)

The Glasswing distribution problem — frontier defensive capability concentrated at the top of the operator size distribution, inversely correlated with OT exposure — requires a dedicated program for Tier 3 operator defensive harness uplift. This program should:

- Prioritize the four Tier 3 operator categories with the highest OT exposure and lowest cybersecurity capacity: water and wastewater utilities serving populations under 100,000, rural electric cooperatives, critical manufacturing facilities with networked ICS, and rural and community hospitals.
- Fund development and deployment of standardized defensive harness architectures — not frontier model access, which is not the binding constraint — validated against the ICS benchmark and adapted for the resource and operational constraints of Tier 3 operators.
- Operate through existing CISA relationships with sector-specific ISACs and Information Sharing and Analysis Organizations, leveraging the trust infrastructure already established for incident reporting.

Recommendation 4: Establish a Harness Architecture Monitoring Program (Medium-Term)

The fifth pathway — harness-mediated capability reconstruction from openly available precursor models — is not currently monitored by any governance program. A harness architecture monitoring program should:

- Maintain a running capability taxonomy of publicly available harness designs, classified by task class and capability profile against standardized benchmarks.
- Track the convergence of open-weight model capability and harness sophistication on the ICS attack surface, providing leading indicator data to the ICS benchmark monitoring program. DeepSeek-V4-Pro, released May 2026, is a concrete instance of what this program would monitor: a PRC-origin open-weight model achieving frontier-class coding performance, with commodity-deployment architecture, available without access restrictions.
- Integrate with existing OSINT collection on adversarial AI development, with specific attention to the application of harness engineering to ICS targets by state-sponsored actors.

Conclusion

The governance response to a capability threshold crossing has historically been cautious — to slow down, to study, to defer deployment until accountability frameworks are adequate. That instinct is not irrational. Caution toward powerful and rapidly evolving technologies is rational. The governance challenges are real: opacity in model behavior, difficulty attributing outputs, institutional capacity gaps that make oversight genuinely difficult to sustain. Organizations and policymakers who insist on accountability frameworks before deployment, 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, in a deteriorating threat environment, is cost-free.

But caution depends on a condition. The condition is that the environment being managed is stable enough that deferral preserves rather than erodes the option of safe adoption later. That condition no longer obtains. The pipeline decomposition establishes this. Three of the four layers of the autonomous offensive cyber pipeline are already closed. The fourth layer is closed for corporate IT targets and open for ICS/OT targets. The governance window on the one remaining tractable domain is defined by adversarial research momentum and inference cost economics, not by policy deliberation timelines. In that environment, deferral does not preserve optionality. It compounds risk.

The May 2026 release of DeepSeek-V4 sharpens the point. Open-weight frontier-class coding capability — designed for commodity deployment, PRC-origin, released without access restrictions — is now available to anyone who wants it. The fifth pathway to domain-level AI cyber capability is no longer hypothetical. The governance architecture has no instrument targeting it. This does not require revising the four recommendations in the preceding section. They remain calibrated correctly. It requires executing on them with urgency proportional to a window that is measurably narrowing.

The most measured response – calibrated to the actual risk — is accelerated governed adoption: deliberate, auditable, human-teaming in its operational logic, 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. Current policy is unlikely to mitigate increasing risk. The most vulnerable infrastructures deserve priority attention.

Restraint, in this environment, is itself a risk posture — and not the safer one.

Bibliography

References

1. AISLE. 'AI Cybersecurity After Mythos: The Jagged Frontier.' AISLE Blog. April 7, 2026.
2. AISLE. 'System Over Model: Zero-Day Discovery at the Jagged Frontier.' AISLE Blog. April 2026.
3. Anthropic. 'Claude Mythos Preview System Card.' red.anthropic.com/2026/mythos-preview/. April 7, 2026.
4. Dang, J., Xie, B., and Younis, O. 'Subliminal Transfer of Unsafe Behaviors in AI Agent Distillation.' arXiv:2604.15559. April 16, 2026.
5. Folkerts, L. et al. 'Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios.' arXiv:2603.11214. March 2026.
6. Jiang, B. 'DistillGuard: Evaluating Defenses Against LLM Knowledge Distillation.' arXiv:2603.07835. March 8, 2026.
7. Liu, H. et al. 'Synthesizing Multi-Agent Harnesses for Vulnerability Discovery.' arXiv:2604.20801. April 2026.
8. OSTP. 'Adversarial Distillation of American AI Models (NSTM-4).' April 23, 2026.
9. Vaynman, J. And Volpe T. "Dual Use Decepton: How Technology Shapes Cooperation in International Relations." International Organization Vol. 77, (Summer 2023), pp. 599-632.
10. Woolley, S. and McConnell, M. 'The Case for Engineering an AI Partner for Intellectual Honesty.' Small Wars Journal. April 30, 2026.
11. DeepSeek-AI. 'DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence.' Technical Report. May 2026. HuggingFace: deepseek-ai/deepseek-v4.

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. He is also 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 (CISA), 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.

Revision 3, May 2026.

ICIT

www.icitech.org