From Near Miss to Pause:

Historical Warning Shots and AI Governance

Zhamilia Klycheva, Ph.D. (Mentor), Shubhankar Dharmadhikari, Edidiong James, Erik Leklem

zhklycheva@gmail.com,
shubhu.dharmadhikari@gmail.com, edianajames@gmail.com, eleklem@gmail.com

Executive Summary

The U.S. government[1] often fails to convert warning shots into effective legal and policy solutions for the security and safety of its citizens. A core government function is to recognize indications of future catastrophe, then formulate successful government prevention and response. Yet this governance gap persists.

The gap is a significant risk for American society, especially in an era of AI. The nation is accelerating into an AI future of greater cyber, bio, and loss of control risks, amidst ongoing public harms and dangers (Bengio et al., 2026).

Our project aimed to address this gap. First, we developed a clear definition of what a warning shot is, and their five criteria. We then applied this warning shot analysis to 25 historical "candidate" cases in nuclear, cyber, bio, and military domains. Next, we filtered 13 cases that fully qualified, affirming that governments struggle to translate acknowledged incidents into effective government solutions. From this review, we identified six stages of how governments convert warning shots into improved safety and security conditions for their citizens (these stages form our proposed "Governance Convergence Framework (GCF)").

To address convergence shortfalls, we recommend that the U.S. Congress establish an independent agency to track emergent AI warning shots, register incidents for government attention, and facilitate their rapid conversion through the GCF for appropriate resolution. We also recommend that Congress designate an existing committee to have lead oversight

[1] We postulate that this governance gap also applies to many governments across the world, but focused our policy analysis and recommendations on U.S. audiences. We believe the recommendations are largely transferable to other international jurisdictions.

responsibility for AI warning shots and government regulation/resolution. Lastly, we recommend that one or several civil society organizations (AI safety labs, non-profits, AI safety associations, etc.) establish an external AI warning shot reporting and advocacy initiative to hold the U.S. government accountable for resolving emerging AI risks.

Introduction & Problem Statement

Our project examined how "warning shots"—near-miss events that reveal a credible pathway to catastrophic risk—translate into governance action (or fail to do so). Warning shots provide opportunities (or "signals") for societies and governments to learn and implement preventative governance. However, historical evidence suggests these signals frequently fail to produce meaningful reform. This is a historical policy problem. But it is a particularly acute problem in the AI domain and in a world characterized by rapidly evolving AI-driven existential risk.

Given this problem, we asked: *under what conditions do warning shots convert into binding governance, why do they fail, and what lessons do they have for governments, the frontier AI environment, and existential risk futures?* This question is particularly urgent for AI governance, where early warning signals—such as dual-use model misuse, loss of control, or unsafe capability demonstrations—may not be addressed in time to prevent existential risk futures.

Our analysis uncovered several underlying and complicating issues to this main problem statement above.

Primarily, a standing definition of AI warning shots does not currently exist. The resulting definitional confusion and analytical diffusion contributes to disorderly governance recommendations. In response, we developed a formal definition to provide greater clarity for researchers and policy-makers going forward.

We also observed that a parsimonious framework for analyzing warning shot governance convergence does not appear to exist either. Existing literature has examined individual incidents or organizational learning, but lacks a generalizable framework explaining how warning shots should move from incident recognition to enforceable policies or laws. We thus proposed a new "Governance Conversion Framework (GCF)" to identify how to improve government responses to future warning shots.

The primary audiences for this paper are U.S. policymakers and implementers, with secondary audiences including state governments, international actors, and industry engaged in AI governance and frontier technology oversight.

Background & Context

Governance for some U.S. business sectors exists today (e.g. financial securities and exchanges, pharmaceutical development, environmental harms, the nuclear industry). This is not the case for AI systems. That said, existing legal and policy precedents provide a wide range of lessons for improving AI governance going forward. For example, the Nuclear Regulatory Commission (NRC) provides many lessons for strong governance of warning shots. The NRC has comprehensive oversight and regulatory powers across the six stages of the GCF, with the caveat being that these apply to licensed facilities only (NRC, 2016). Other robust precedent governance conversion framework examples can be drawn from other regulatory bodies, such as the U.S. Securities Exchange Commission (SEC), the Food and Drug Administration (FDA), and similar bodies (FDA, 2023).

In the United States, AI systems lack regulatory oversight and national governance due to the lack of comprehensive AI law that would enumerate such powers. Existing agencies that have some powers over issues like AIxCyber warning shots, such as CISA, have only limited governance authority for a small number of the stages of the GCF "pipeline."

While several U.S. states have established AI laws or are pursuing new AI legislation, there are no existing state entities with authorities needed for AI warning shot governance conversion as of mid-2026 (Gunderson Dettmer, 2026).

As for international context, there are some positive elements of the European Union AI Act that could be instructive to the formulation of an AI warning shot governance conversion framework authorities. One example is the Scientific Panel (EU AI Act, 2024, Art. 68) and their authority to monitor AI risk incidents (Stage 1) and issue a registered warning (State 2) to the EU AI office for action (Stage 3).

Policy process literature offers a structural explanation for GCF institutional gaps. Focusing events force problems onto the policy agenda but do not automatically produce change (Birkland, 1997); conversion requires a problem, available solutions, and political will to converge before the window closes (Kingdon, 1984). Opposing coalitions with commercial or institutional interests in the status quo are often more durable than reform coalitions, and most effective after peak salience fades (Sabatier, 1988). The Governance Conversion Framework is

designed to map these structural problems and illuminate where government responses are stalling, failing, or being sidetracked by status quo stakeholders.

Analysis

Our analysis began with research into warning shot etymology, historical usage, and association with intelligence methods such as indications & warning from the U.S. intelligence community. We found that a contemporary and clear definition of a warning shot (WS) is not widely understood; as such, we developed a specific definition and related conditions, or criteria.

We define a warning shot as: "a discrete event that exposes latent vulnerabilities within a system, organization, or society. It illuminates a credible pathway toward systemic catastrophe, proving that such a failure is possible under existing conditions. While the realized harm of a warning shot remains significantly lower than its potential magnitude, the event serves as a critical precursor—transforming abstract speculation into a demonstrable threat."

Crucially, an AI warning shot is defined by its structural characteristics rather than the societal reaction it provokes. Whether an institution chooses to acknowledge, suppress, or ignore the signal is analytically irrelevant; the event's status as a warning shot is determined solely by its capacity to reveal a plausible trajectory toward disaster.

An event qualifies as a warning shot only if it satisfies the five conditions below:

- *C1 — Discrete Bounded Event.* The event must have a clear onset, duration, and end.
- *C2 — Structural Vulnerability Revealed.* The event must expose a genuine, pre-existing weakness in the sociotechnical system that becomes newly legible because of it.
- *C3 — Credible Escalation Pathway.* A plausible counterfactual must exist: under slightly different conditions — less luck/no intervention — it could have produced a catastrophe.
- *C4 — Prospective Signal Accessibility.* The catastrophic pathway must have been recognizable at the time using information and frameworks then available.
- *C5 — Sub-Catastrophic Ceiling.* The event itself must not be the catastrophe — realized harm can range from minimal to severe, but must remain below catastrophic level.

Applicability Analysis

Once a warning shot has been classified by C1-C5 criteria, the next stage of our case filtering exercise determined how closely that historical case could parallel frontier AI risk (Vermeer, 2024), the strength of the lessons learned, and governance pathways for cases we may see within the frontier AI landscape (George, 2019). We then collected 25 warning shot candidate cases, and added a weighted "AI Applicability Scoring Framework" (D1–D4) to assess case relevance to frontier AI governance. We used this approach to avoid unstructured relevance judgments, which can lead to loose reasoning that can generate policy errors (Neustadt & May 2011; Mearsheimer & Walt, 2013). This applicability framework applied structured, focused comparison to yield 13 analytically load-bearing cases. The four dimensions are mutually exclusive, each examining a distinct analytical object, so no dimension double-counts another.

The four dimensions (D1-4) are:

- *D1 — Structural Mechanism Isomorphism (x2 weight):* Assesses whether the case's failure mechanism has a direct architectural counterpart in frontier AI systems, including dual-use risk dynamics. It does not assess how far the event could have escalated (the object of D3) or what governance response followed it (the object of D2 and D4) (Mearsheimer & Walt, 2013). Treated as a gating factor: without a credible structural match, downstream governance and institutional lessons cannot be validly transferred.
- *D2 — Governance Architecture Transferability (x1.5 weight):* Evaluates the governance response as a design artifact — its institutional structure, enforcement logic, and portability to AI governance. D2 does not ask why the institution produced that output, that is determined through D4 (Birkland, 1997).
- *D3 — Escalation Pathway Fidelity (x1 weight):* Assesses whether the counterfactual escalation chain previews a credible AI catastrophic pathway (Fearon, 1991; Barma et al., 2016). Weighted lowest to prevent speculative scenarios from elevating cases to Tier 1 given frontier AI's high uncertainty.
- *D4 — Institutional Learning Diagnostic (x1.5 weight):* Examines the organizational, political, and psychological dynamics determining whether the warning shot converted into binding governance. Whereas D2 evaluates the output, D4 diagnoses its cause, drawing on the literature on normalization of deviance (Vaughan, 1996), normal accidents (Perrow, 1984), latent failure (Reason, 2000), man-made disasters (Turner, 1978), and near-miss processing (Dillon & Tinsley, 2008).

The specific weighting multipliers (×2, ×1.5, ×1, ×1.5) are a defensible methodological choice rather than meant to be an empirically derived parameter; what the framework treats as load-bearing is the *ordinal priority* of D1 > D2 = D4 > D3. A sensitivity analysis across five

alternative weighting schemes confirmed that the core set of highest-scoring cases remains stable under any scheme preserving that ordering.

Each dimension was scored 1–5. The full 'equation' is captured within Figure 1. Tier 1 requires a composite of ≥ 24.0 and a D1 score of ≥ 4, codifying the gating logic.

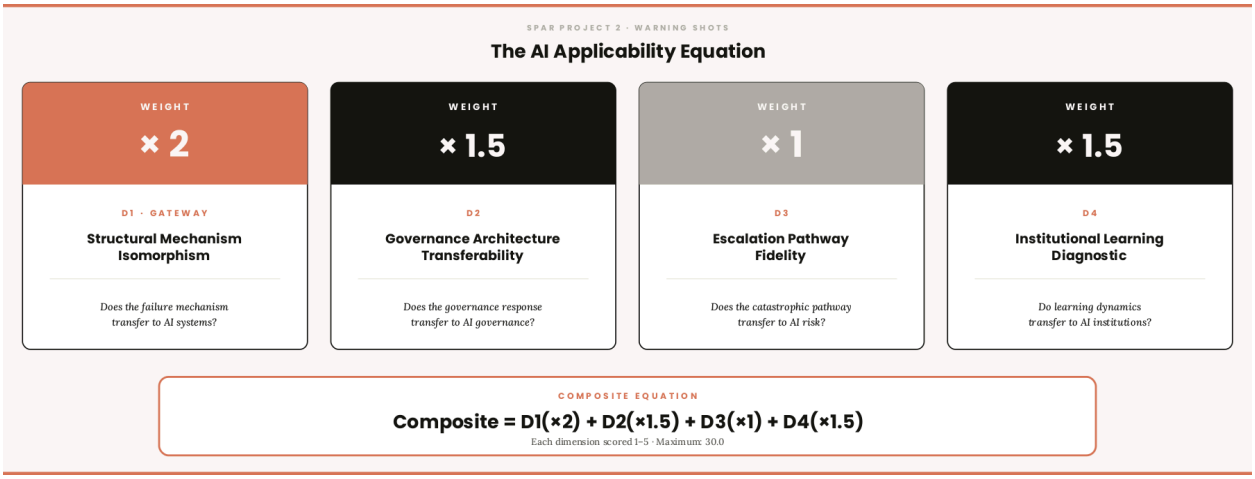

Fig. 1 AI Applicability Equation (Image Generated by Claude)

Case Downselection

Applying the framework to the originally determined set of 25 warning shot cases yielded 12 Tier 1 cases, with one deliberate addition made later for a total of 13, with a notable concentration within the dual-use capability domains of biosecurity and cybersecurity. This concentration follows from the structure of the technologies themselves.

Dual-use potential is not uniform across technologies — it varies in how deeply that potential is built into a given technology. A technology's inherent dual-use potential, meaning the properties that make it susceptible to both beneficial and harmful use largely independent of context, is distinct from its extrinsic potential, which turns on how the technology is deployed and toward what end (Hähnel, 2025). We extend this distinction by characterizing as constitutively dual-use those domains in which the inherent dimension dominates: where the capability that produces benefit is the same capability that produces harm, and the two cannot be separated even in principle. Biosecurity and cybersecurity are constitutively dual-use in this sense, and so is frontier AI. AI research is dual-use by virtue of its inherently digital and general-purpose character — the capability that detects a software vulnerability is the capability that exploits it; the model that designs a therapeutic molecule is the model that designs a toxic one (Brenneis, 2025). Nuclear and conventional military technologies sit at the other end of this spectrum, in what we term contingently dual-use domains, where harmful application

generally requires a deliberate and institutionally visible diversion step rather than being latent in the capability itself. Warning shots from biosecurity and cybersecurity therefore transfer to frontier AI more readily than those from nuclear or military domains, because they share the underlying risk structure that D1 is designed to detect. Nuclear and military cases nonetheless contribute distinctive institutional learning insights not available elsewhere.

That biosecurity and cybersecurity cases score higher on D1 is the central finding rather than an artifact of the method: the framework weights D1 domain-neutrally, and the concentration is an empirical result produced by a domain-neutral instrument. Cases are fully captured in Figure 2.

We added the Three Mile Island case as a deliberate exception. It does not clear the Tier 1 gateway — its failure mechanism has only a weak architectural counterpart in AI systems and its D1 score is low — but it is the clearest historical case of a near-miss converting into durable governance reform, making it analytically indispensable to GCF analysis.

SPAR PROJECT 2 · WARNING SHOTS

The 13 Core Cases

Warning-shot portfolio — D1–D4 Tier 1 (12 cases) plus Three Mile Island

CASE	DOMAIN	SCORE	SUMMARY
H5N1 Gain-of-Function 2011	BIO	29.0	Lab-created airborne-transmissible H5N1; triggered the first NSABB recommendation to restrict publication of the method.
Triton / TRISIS 2017	CYBER	28.5	First malware purpose-built to disable industrial safety systems; an accidental plant shutdown exposed it before catastrophe.
MegaSyn Dual-Use AI 2022	BIO	28.0 *	Drug-discovery AI inverted to design toxins — about 40,000 toxic candidate molecules generated in under six hours.
CureVac State Co-Option 2020	BIO	27.5	Attempted state co-option of a COVID-19 vaccine maker; Germany responded with a defensive government equity investment.
Northeast Blackout 2003	CYBER	27.5 *	A grid alarm-system software bug cascaded into a blackout affecting about 50 million people; produced binding NERC reliability reform.
Birmingham Smallpox 1978	BIO	27.5 *	A laboratory containment failure caused the last recorded smallpox death, during the disease's eradication endgame.
SolarWinds / SUNBURST 2020	CYBER	27.5	Software supply-chain compromise distributed a backdoor to about 18,000 organizations, including US federal agencies.
IL-4 Mousepox 2001	BIO	26.0	A gene insertion accidentally produced a lethal virus that defeated vaccine immunity even in resistant mice.
Crimea Annexation 2014	MILITARY	25.5	Hybrid warfare conducted below the conventional-war threshold — the paradigm case of strategic ambiguity and deniability.
Operation Spiderweb 2025	MILITARY	25.5	Smuggled low-cost drones struck strategic bomber aircraft deep inside Russia — cheap systems defeating high-value assets.
Petrov Incident 1983	NUCLEAR	24.0	A Soviet officer correctly judged a false satellite missile-attack warning, averting a potential nuclear escalation.
Stuxnet 2010	CYBER	24.0	The first cyber weapon to cause physical destruction; it spoofed monitoring systems while wrecking enrichment centrifuges.
Three Mile Island 1979	NUCLEAR	—	A partial reactor-core meltdown that produced sweeping, durable nuclear-safety reform — the portfolio's positive exemplar.

** Composite as stated in the framework scoring table. For these three cases the stated value does not match the weighted sum of the dimension scores (each recomputes to 27.0); to be reconciled before publication. Three Mile Island is not D1–D4 scored — added as the institutional-learning exemplar.*

Score = D1–D4 weighted composite (max 30). Full scored table with dimension breakdown and governance outcomes appears in the appendix.

Figure 2: 13 Tier 1 WS incidents (Image Generated by Claude)

Governance Conversion Framework

Governments or external monitors typically recognize warning shot incidents, but these are rarely converted fully into binding governance solutions. To understand better why this is the case, we analyzed the 13 historical warning shot incidents to map how they translated into effective governance action (or failed). Our analysis illuminated six sequential stages of governance actions that given historical warning shot incidents traverse before fully converting into a binding regulatory action that lasts at least 1 full election cycle. Combined, we describe this as the "Governance Conversion Framework (GCF)." The GCF is not limited to a single national context but is applied from the perspective of actors capable of producing binding governance. Figure 3 demonstrates the six stages and portrays how warning shot events progressed to a given stage.

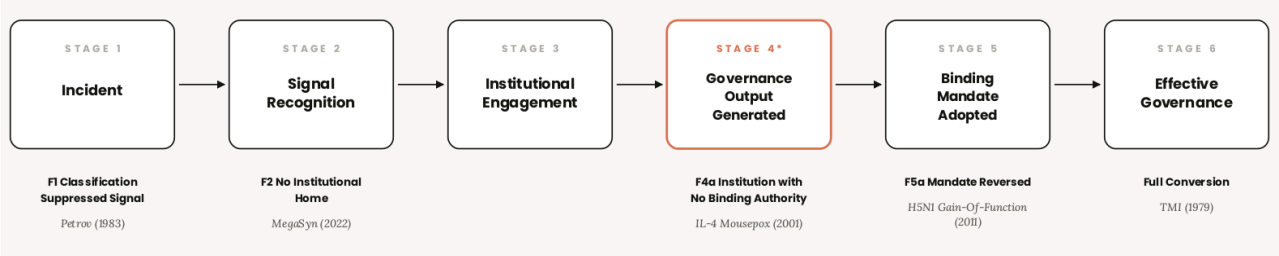

Figure 3: Governance Conversion Framework (Image Generated by Claude)

Each stage and the associated 3 necessary preconditions are briefly described below:

- *Stage 1 Incident:* a warning shot occurs. A discrete, bounded, sub-catastrophic event revealing a structural vulnerability with credible escalation pathway. The event must be legible as a systemic signal and not a one-off accident. Cases can fail here before conversion begins — Petrov (1983) stalled at Stage 1 when Soviet classification suppressed the signal entirely, preventing it from entering public governance channels.
- *Stage 2 Signal Recognition:* evidence of a broader structural problem is recognized. Key actors translate the event into a governance-relevant claim requiring institutional response. MegaSyn (2022) stalled here — signal recognition occurred but no institutional home existed to receive and act on it, terminating conversion at Stage 2.
- *Stage 3 Institutional Engagement:* An institution with some authority takes up the signal, initiates investigation, convenes experts, opens regulatory proceedings.
- *Stage 4 Governance Output:* Institutions produce concrete recommendations, an advisory framework, or a report. While this stage looks like governance, it is not yet binding. IL-4

Mousepox (2001) stalled here — the institutional response produced scientific discussion but the relevant bodies lacked authority to issue binding requirements.

- *Stage 5 Binding Mandate:* Legally or institutionally binding requirement issued — a regulation, an executive order, a treaty obligation. Non-compliance carries consequences. It is not yet the final stage, as the governance may still be reversed. H5N1 Gain-of-Function (2011) reached Stage 5 but stalled when the moratorium was lifted, demonstrating that binding mandates without durable enforcement architecture remain vulnerable to reversal.
- *Stage 6 Effective Governance:* The legal or binding mandate is enforced, compliance monitored, and structural vulnerability durably addressed. It is important that governance persists through at least 1 full political cycle to qualify. This indicates full conversion and a successful outcome. Three Mile Island (1979) and North East blackout (2003) are the only two cases in our data set that resulted in full governance conversion.

Preconditions

We identified three preconditions necessary to result in a regulatory conversion. Missing a single one of the preconditions is sufficient to prevent full conversion.

- *P1 — Domestic Victim at Scale.* Harm must fall on the regulating state's own population at sufficient scale to generate political salience. Foreign victims, diffuse harms, or near-misses without casualties cannot sustain the domestic pressure required to cross stage 4 to stage 5. This fails in: Petrov (no harm occurred) and Stuxnet (harm to Iranian centrifuges — foreign adversary's weapons programme).
- *P2 — No Information Suppression.* Incident details must propagate through public institutional channels. Suppression — whether via formal state classification or private interest-motivated non-disclosure such as competitive pressure or coordinated lobbying — prevents durable regulatory architecture from forming. Fails in: Petrov (classified for 30 years), Stuxnet (covert operation, never officially attributed).
- *P3 — Institutional Capacity* (two sub-conditions, both required)
- *P3a — Institutional Home with Enforcement Authority.* A body with jurisdiction over the relevant domain must exist and hold genuine enforcement power, not just advisory capacity. Fails entirely in MegaSyn, Swarm Drones and, partially, in IL-4.
- *P3b — Uncontested Primary Jurisdiction.* One institution must hold clear, undisputed authority to lead the response. When multiple agencies hold overlapping claims, output is negotiated down to the intersection of their comfort zones, constraining mandate scope. Fails in: Triton (jurisdictional fragmentation across Industrial Control Systems and national security domains), SolarWinds (Cybersecurity and Infrastructure Security

Agency (CISA)/National Security Agency (NSA)/Federal Bureau of Investigation (FBI)/ Office of Management and Budget (OMB) overlap — P3b strain limited mandate scope to federal civilian agencies, blocking Stage 6).

The Current Day Warning Shot: Claude Mythos (And GPT 5.5)

As of this project's submission in late May, there is an ongoing AI warning shot that is actively testing our analytic frameworks, and the parsimony of the GCF: the Claude Mythos warning shot case. Through this project, we observed the evolution of the Claude Mythos signal via our project's AI warning monitor and tracking report system, registering the signal as *a possible* warning shot weeks before it became popularly understood as a Stage 1 incident.

In terms of background: on April 7, 2026, Anthropic announced Claude Mythos Preview and simultaneously said it would not be releasing it publicly. Independent testing by the UK's AI Security Institute (AISI, 2026) confirmed that Mythos was the first AI model to complete a 32-step simulated corporate network attack end-to-end, a task estimated to take a human expert around 20 hours, in 3 of 10 attempts. Rather than a standard release, Anthropic launched Project Glasswing, a gated program giving roughly 40 organizations access to the model for defensive security purposes (Anthropic, 2026). Within two weeks, unauthorized users had accessed it through a third-party contractor connection. The White House intervened informally. Regulators in the US, Canada, and the EU held emergency consultations. The IMF named Mythos specifically as a driver of rising financial stability risk. The situation became significantly more complicated when AISI published results showing that OpenAI's GPT-5.5 matched Mythos on the same benchmark and was already publicly available.

The Mythos incident satisfies all five warning shot criteria. It is a discrete, bounded event that reveals a structural vulnerability, specifically the absence of any regulatory framework capable of governing general-purpose frontier AI with autonomous offensive capability at this level. The escalation pathway is credible, and the harm so far remains below the catastrophic threshold the event makes plausible. Signal recognition has been unusually fast, largely because Anthropic made the governance problem explicit in its own documentation and AISI's evaluation removed any doubt about the capability claim.

The case also shows the challenges of stage 3 institutional engagement converting further in the GCF. Without a willing enforcement authority, progress to Stage 5 can be challenging. CISA cannot mandate what frontier AI labs may develop. UK AISI can evaluate but has no authority over U.S. companies. The advisory infrastructure forming around the problem is real and has

value (partially into Stage 4), but based on the patterns observed across this project's historical case portfolio, it is more likely to substitute for binding governance than to produce it. The GPT-5.5 development makes this harder: once an equivalent capability is publicly deployed by a second actor, the political case for restricting the original weakens considerably. As of May 22, 2026, the historical pattern of failed conversion to stage 6 seems to be more likely: on May 21, President Trump postponed a proposed Executive Order (stage 4) that was rumored to have some mandates for improved governance over future AI models (stage 5).

Recommendations

Notwithstanding the AI governance headwinds of the postponed Executive Order, the Mythos warning shot can still inspire new ideas for improved governance conversion. In this spirit, our recommendations trend to the ideal, with an eye towards being potentially more feasible as additional warning shots and compounding societal harms will likely generate more political salience. We acknowledge that less comprehensive recommendations may be more achievable in the current U.S. political environment of this legislative session.

In this regard, we propose the following recommendations in the context of a potential U.S. midterm election that could a) result in a change in congressional control of one or both chambers to the Democratic Party, b) be characterized by increasing public concern about AI harms and demands for improved governance, and c) likely additional AI warning shots or increasing near-catastrophic/catastrophic levels of harms from prior warning shots (e.g. Mythos/GPT 5.5 implications get worse). Our recommendations should be considered as debate proposals now, for early action in an altered (and more favorable) political context of the 120th Congressional Session of early 2027.

U.S Federal Government and Congress

The historical record is consistent: binding governance solutions to warning shots are most likely when a single institution with uncontested jurisdiction and enforcement authority are extant. The U.S. currently lacks any equivalent for AI. Congress should designate an existing agency or create a new one with clear mandates to identify AI warning shots, track their progression, and escalate them for formal regulatory action to stage 6. Without this, each new incident requires a governance response to be assembled from scratch under political time pressure, with little institutional memory of what has worked before. Additionally, Congress should explore a fast track mechanism for governance conversion when a warning shot crosses a defined risk threshold, reducing the window between signal recognition and binding mandate. We also recommend that Congress designate a lead committee to have authority over

warning shot governance and resolution to solution, so that this committee can exert legislative oversight and hold the executive branch to account.

Frontier AI Companies

Voluntary commitments and self-imposed restrictions have a role to play, but the case portfolio shows they do not survive sustained commercial pressure or changes in the competitive landscape. The most useful contribution industry can make is to support the development of binding frameworks that apply to all actors equally. A shared regulatory baseline removes the competitive disadvantage that makes individual companies reluctant to implement restrictions unilaterally, and mitigates dynamics that have driven governance failure across multiple domains in this analysis.

Civil Society and AI Safety Organizations

Civil society has a particular role to play at the stage 4 to stage 5 transition, which is consistently the point where conversion stalls. Advisory outputs and voluntary frameworks tend to absorb political pressure without producing binding mandates, and their existence can reduce the urgency for legislative action. AI Safety organizations should focus sustained advocacy on this gap specifically by maintaining pressure past the peak salience of individual incidents and by developing standardized capability of evaluation frameworks that regulators can use as the technical basis for binding thresholds. Without that technical infrastructure, the policy stream that Kingdon's framework identifies as necessary for conversion remains incomplete.

International

Warning shots rarely respect national borders, but governance responses are almost always domestic. The structural mismatch between where incidents occur and where enforcement authority sits is one of the clearest patterns in this analysis. Allied governments should work toward shared definitions of what constitutes an AI warning shot and shared protocols for coordinating governance responses when incidents cross jurisdictions. Bilateral pre-deployment evaluation agreements between like-minded governments are a practical near-term step that does not require waiting for broader multilateral consensus.

Conclusion

Warning shots are not rare. What is rare is the institutional response that converts them into durable governance. Across the historical cases spanning nuclear, biosecurity, cybersecurity, and other domains, this project found predictable failures in predictable governance actions for predictable reasons: fractured institutional authority, organized opposition that outlasts

political salience, and competitive dynamics that make unilateral binding restrictions unsustainable. Full conversion occurred only twice in our case set. Success hinged on a common structural feature: a single institution with uncontested jurisdiction and genuine enforcement authority was already in place when the warning shot occurred.

These findings suggest near-term priorities for resolution. First, an AI safety organization or several should be responsible for identifying AI warning shots as they happen, facilitate their progression through the governance framework, and report on failures/successes. Another beneficial priority could be for future legislative or policy action to incorporate the governance framework as a legal procedure for managing AI risks and harms.

The Mythos case makes the cost of the AI governance gap especially poignant. The warning shot was clear, the window to respond has been open, and the advisory infrastructure forming around it was beginning to feel real. But advisory infrastructure is not enforcement infrastructure, and the historical record is consistent on what happens when the two are treated as equivalent.

The Trump Administration's postponement of a new Executive Order designed to address the Mythos warning shot underscores why a stronger governance conversion framework (and standing authority) is urgently needed. This ongoing case shows that some frontier AI companies and AI accelerationist advocates will continue to derail ad hoc governance responses to AI warning shots. In an era of existential risk, governance responses vulnerable to privileged interest advocacy will likely fail to manage those risks. New AI governance institutions, with robust authorities, are essential to ensure the security and safety of the public, and humanity's future.

References

Anthropic. (2026, April 7). Claude Mythos Preview and Project Glasswing. https://www.anthropic.com/glasswing

Barma, N. H., Ratner, E., & Weber, S. (2016). "Imagine a world in which": Using scenarios in political science. *International Studies Perspectives, 17*(2), 117–135.

Bengio, Y., et al. (2026, February). *International AI Safety Report 2026* (DSIT 2026/001). https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026#2.1

Birkland, T. A. (1997). *After disaster: Agenda setting, public policy, and focusing events.* Georgetown University Press. https://press.georgetown.edu/Book/After-Disaster

Brenneis, A. (2025). Assessing dual use risks in AI research: Necessity, challenges and mitigation strategies. *Research Ethics, 21*(2), 302–330.

Dillon, R. L., & Tinsley, C. H. (2008). How near-misses influence decision making under risk: A missed opportunity for learning. *Management Science, 54*(8), 1425–1440. https://doi.org/10.1287/mnsc.1080.0869

European Parliament and Council of the EU. (2024, July 12). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).* Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Fearon, J. D. (1991). Counterfactuals and hypothesis testing in political science. *World Politics, 43*(2), 169–195.

George, A. L. (2019). Case studies and theory development: The method of structured, focused comparison. In D. Caldwell (Ed.), *Alexander L. George: A pioneer in political and social sciences* (pp. 191–214). Springer. https://doi.org/10.1007/978-3-319-90772-7_10

Gunderson Dettmer. (2026, February 5). *2026 AI laws update: Key regulations and practical guidance.*
https://www.gunder.com/print/v2/content/25131/2026-ai-laws-update-key-regulations-and-practical-guidance.pdf

Hähnel, M. (2025). Conceptualizing dual use: A multidimensional approach. *Research Ethics, 21*(2), 205–227.

IMF. (2026, May 7). *Financial stability risks mount as artificial intelligence fuels cyberattacks.* https://www.imf.org/en/blogs/articles/2026/05/07/financial-stability-risks-mount-as-artificial-intelligence-fuels-cyberattacks

Kingdon, J. W. (1984). *Agendas, alternatives, and public policies.* Little, Brown.

Mearsheimer, J. J., & Walt, S. M. (2013). Leaving theory behind: Why simplistic hypothesis testing is bad for international relations. *European Journal of International Relations, 19*(3), 427–457.

Neustadt, R. E., & May, E. (2011). *Thinking in time: The uses of history for decision makers.* Simon and Schuster.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies.* Basic Books.

Reason, J. (2000). Human error: Models and management. *BMJ, 320*(7237), 768–770. https://doi.org/10.1136/bmj.320.7237.768

Sabatier, P. A. (1988). An advocacy coalition framework of policy change and the role of policy-oriented learning therein. *Policy Sciences, 21*(2–3), 129–168. https://doi.org/10.1007/BF00136406

Turner, B. A. (1978). *Man-made disasters.* Wykeham Publications.

UK AI Security Institute (AISI). (2026, April 13). *Our evaluation of Claude Mythos Preview's cyber capabilities.* https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities

U.S. Food and Drug Administration. (2023). *FDA's Sentinel System: A national electronic system for monitoring medical product safety.* https://www.fda.gov/safety/fdas-sentinel-initiative

U.S. Nuclear Regulatory Commission. (2016). *Reactor oversight process* (NUREG-1649, Rev. 6). https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr1649/

U.S. Securities and Exchange Commission. (2013). *The investor's advocate: How the SEC protects investors, maintains market integrity, and facilitates capital formation.* https://www.sec.gov/about/what-we-do

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA.* University of Chicago Press.

Vermeer, M. J. D. (2024). *Historical analogues that can inform AI governance.* RAND Corporation. https://www.rand.org/pubs/research_reports/RRA3408-1.html