
Warning Shots Eliminate Models, Not Uncertainty¹

Ayan Sivaram
New York University

With
Apart Research

Abstract

Warning shots are said to fail because they do not travel. I argue they travel and move the wrong quantity. An incident raises the estimated rate of safety-case-falsifying events and separately forecloses layers of the safety case that cannot accommodate it. Under unanimity among veto players with heterogeneous priors, action tracks not the average estimate but the optimistic bound: the most reassuring layer still standing for the least persuaded player. Two bodies of evidence inducing the same distribution of posteriors can therefore differ in whether they produce action, and severity moves neither channel. In the July–September 2026 agent-containment record, state enforcement followed the victim’s forensic reconstruction rather than the breach, and every response arrived through the lowest-veto-count channel available. Reporting duties should attach to events falsifying providers’ published capability claims: a trigger fixed ex ante that scales with capability, where outcome-magnitude thresholds do not.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

1. Introduction

A survey conducted by the Annenberg Public Policy Center in early 2026 found that 65% of Americans² say the government has done too little to regulate AI. Even 43% of survey respondents who believed AI's net impact will be "very positive" agreed with the statement. In addition to bipartisan consensus on the insufficient governmental response, only 17% of Americans believe AI will have a positive net impact on the United States over the next 10 years.

Four months after the conclusion of the survey, OpenAI agents running inside a cyber-capability evaluation escaped their sandbox through a zero-day in a package registry proxy. Between July 8th and 13th, the agents rooted a third-party code sandbox as a launchpad and executed a multi-day intrusion into Hugging Face's production infrastructure.

Why didn't the public react in the days following the attack on Hugging Face in a manner that the survey data would have predicted? Explanations invoking a simple communications failure as the responsible factor lack a mechanism. In other words, a communication failure can accommodate any outcome ex post and has limited predictive power on the efficacy of potential warning shots before they occur.

In the aftermath, Abecassis and Barnett from MIRI's Technical Governance Team wrote that "it is quite plausible that there will simply be no warning shot which generates an appropriate response" and outlined an eight factor framework to predict the efficacy of warning shots³. To motivate a rigorous study of potential warning shots, I propose a distinction that previous literature, including Abecassis and Barnett, conflate. A warning shot carries two kinds of information. First, accumulative evidence, which raises the estimated rate at which safety-case-falsifying events occur. Second, eliminative evidence, which forecloses accounts of how failure is prevented, or, in other words, which world models are falsified.

The main findings of this paper are:

1. A decision theoretic model in which belief updating and model elimination are two separate channels of an incident's evidentiary value. The model provides a structural account of evidence that is Bayesian at an individual level with a foundational, conceptual explanation of behavior that yields different falsifiable predictions relative to the behavioral psychology literature.

² The findings fell somewhat along political lines with 77% of Democrats, 72% of independents, and 53% of Republicans agreeing with the statement. This is plausibly confounded by the political distribution of representatives, majority Democrat at the time of the Annenberg survey, who have incorporated AI governance into their platform.

³ The factors are how 1) likely, 2) timely, 3) visceral, 4) sudden, and 5) unexpected the incident was, 6) the degree of attribution available, 7) the backlash that followed the incident, and 8) the availability of an appropriate policy response. The blog later grades possible warning shots related to AI risk and finds no threat model adequately meets the eight factors.

2. An empirical and foundational failure diagnosis of the EU AI Act's serious incident trigger. The Hugging Face incident, the most world model-eliminating event in the public record yet, plausibly falls outside Article 3(49) of Regulation (EU) 2024/1689. Conceptually, the trigger conditions on realized damage, which underdetermines the capability an event revealed and is in any case unavailable while the response window is open. Worse, the binding constraint sits upstream of any threshold: whether a duty attaches at all depends on the provider's own classification of the episode, so no calibration of the limbs repairs a trigger that is never reached.
3. A novel framework that follows from the formal model. Frontier labs can be compelled to publish specific, falsifiable claims about dangerous capabilities their systems should not be able to do and what incidents existing controls prevent. By conditioning on revealed capacity for damage rather than realized damage, the trigger is fixed ex ante, is verifiable against a standing text, and scales with capability.

An important limitation of the model is the difficulty in establishing that this is the operative explanation. Indeed, the model could simply be plausible and available. A competitive alternative explanation, individuals and institutions take evidence of contained failures as justification to update downwards the probability of failure, has experimental support and yields very different conclusions relative to this paper's model.

2. Related Work

The second International AI Safety Report, led by Yoshua Bengio and over 100 AI experts, pointed out that one of the most challenging dilemmas faced by policymakers is the so called "evidence dilemma." As AI capabilities improve, the relevant evidence on the risk they pose is not only difficult to analyze, but may not arrive until it is too late to begin forming a policy response. The report continues, finding that this challenge cannot be solved by proactive policy. Having policy fit to old paradigms can pose regulatory burden without meaningfully addressing the ever-evolving state of AI risks.

Abecassis and Barnett's identification of heuristics and proxies that can help gauge the efficacy of potential warning shots is useful yet insufficient. The eight criteria with no mechanism cannot be traded off, cannot say which criteria are redundant, and cannot say what happens when an incident scores high on some criteria and low on others.

Zaller's Receive-Accept-Sample (RAS) model suggests that attentive members of the public converge when elites agree on political messaging. However, AI risk has split elites. But RAS is a model of mass attitude formation, and mass attitudes are the quantity this paper argues is not binding. It predicts the average estimate and predicts it well; nothing in it bears on the optimistic bound, because a distribution of individual attitudes carries no information about which account of

failure survives for the least persuaded actor who can block. The poll numbers RAS explains are this paper’s premise rather than its rival.

Dillon and Tinsley find that people evaluate near-misses as successes and, in experimental settings, subsequently choose riskier decisions. The measured divergence of an empirical result from a Bayesian model is precisely the object of study. While both a decision theoretic and Dillon and Tinsley result predict that near-misses may dampen policy responses, the decision theoretic approach, which I will describe in further detail in Section 3, can specify which incidents break from the behavioral pattern. In other words, Dillon and Tinsley anticipate uniform attenuation while I posit that near-misses can drive institutional action provided they operate via low veto channels⁴.

3. Methods

Rejections of the concerns raised by the AI safety community typically fall in one of five categories⁵. A system lacks the capability to cause the harm; it lacks the propensity, even if able; controls prevent it, even if able and willing; monitoring would detect it, if it happened; and recovery would undo it. The nearest published taxonomy is Clymer et al.’s account of safety cases for advanced AI, which organizes the argument into inability, control, trustworthiness, and deference to AI advisors. I depart from it in one respect that matters here: that scheme has no detection claim, and the claim that monitoring and disclosure would surface an incident of consequence does most of the work in the fourth result below, because it is the only claim whose falsification rescales the evidential content of the whole record rather than removing one story from it. Using the field’s own argumentative structure rather than a taxonomy fitted to one episode is what makes the model generalizable. In other words, a biosecurity incident, an oversight-evasion result and a sandbox escape can code into the same five slots.

3.1 Terms

Let λ be the arrival rate of world-model falsifying events. An agent escaping its sandbox is one instance; an unclaimed capability surfacing in an evaluation, a screening bypass, or an agent evading oversight are others, and nothing below distinguishes them. Each claim is represented by a prior over λ together with the set of failure modes it can accommodate, and each observed event carries a mark recording which claim it falsified rather than what kind of intrusion occurred. An observer holds a finite set of claims still standing. Claims differ in which failure modes they can

⁴ I describe “low veto count channels” in greater detail in section 3.

⁵ Although this is not necessarily the case that this is the comprehensive set of arguments advocating against AI safety measures, the five categories are simply used to illustrate how the decision theoretic model can represent arguments in the practical information environment.

accommodate at all, not merely in how likely they consider each one, and that is what makes ruling out possible: an event of a mode a claim cannot accommodate eliminates it outright rather than merely lowering its weight.

Events arrive at rate λ , but only some are disclosed. If each is disclosed to a given audience with probability q , that audience sees events arriving at rate $q\lambda$. Two consequences follow, and they are the whole machinery of the concealment result. First, for a claim beginning with prior parameters a and b , the estimated rate after n disclosed events over a period T is $(a + n)/(b + qT)$: disclosed events raise it, and elapsed quiet time lowers it in proportion to how reliably events are believed to be reported. Second, what the record identifies is the product $q\lambda$, not λ itself. An observer who does not know the disclosure rate cannot pin down the underlying rate at all, only bound it, with the bound loosening as confidence in q falls. This is the standard problem of choosing policy when data bound a parameter rather than pin it. Write $\bar{\mu}$ for the credibility-weighted average estimate across surviving claims and $\mu_{\min}$ for the lowest estimate any surviving claim implies. The first is what a survey item measures. The second is the optimistic bound, and the next section shows it is what governs action.

3.2 Model

Let the payoff difference between intervening and inaction be $c\lambda - k$. A model-averaging Bayesian intervenes iff $\bar{\mu} > k/c$. However, according to Bewley's inertia assumption⁶, intervention is warranted iff $\mu_{\min} > k/c$. In other words, an alternative is only accepted when it beats the status quo under every element of the set. Thus, under the α -maxmin rule:

$$\text{intervene} \Leftrightarrow \alpha\mu_{\min} + (1 - \alpha)\bar{\mu} > k/c, \quad \alpha \in [0,1].$$

Note that a model-averaging Bayesian approximates $\alpha = 0$ and a Bewley agent approximates $\alpha = 1$. Let V be an index of relevant actors, each with an independent set of models, disclosure rate, and threshold. Institutional change requires unanimity which can be represented as:

$$\min_{v \in V} [\mu_{\min,v}(n, q_v) - k_v/c_v] > 0.$$

Two common-sense results follow. First, the gating factor is the optimistic bound of the least persuaded relevant actor or the "binding optimist." In the context of the Hugging Face incident, a plausible binding optimist is the federal legislature. Second, the first responder to a warning shot

⁶ Bewley's framework is needed because the relevant uncertainty is Knightian rather than probabilistic: the actor does not hold a single prior over the crossing rate but a set of them, one per surviving world-model, and has no basis for averaging across them. Dropping completeness rather than assuming pessimism is what makes inaction a theorem rather than an assumption.

arrives via the lowest veto count channel. For example, a state attorney general with authority under existing consumer protection statute approximates $|V| = 1$.

Blackwell's informativeness theorem states an experiment is relatively more informative if every agent weakly prefers it and is ordered by the induced distribution of posteriors. It follows that two experiments that induce the same distribution of posteriors are decision-equivalent, according to Blackwell. That equivalence fails here, and its failure is the central claim of this paper. Suppose two claims are still standing, the more reassuring of which sets the optimistic bound. Disclosure A reports an event of a kind both claims can accommodate: the count rises and nothing is ruled out. Disclosure B reports an event of a kind the reassuring claim cannot accommodate: the count rises and that claim is falsified. The weights can be chosen so that both disclosures leave the average estimate in exactly the same place, in which case A and B are Blackwell-equivalent. They are not action-equivalent. After A the optimistic bound is still set by the reassuring claim; after B it has moved up to the next claim standing, so there are thresholds at which B produces intervention and A does not. Evidence that leaves the same posterior can therefore differ in whether it moves an institution, and the ordering that matters for disclosure is not how informative it is in Blackwell's sense but which claim it falsifies, and for whom.

This structure also accounts for a phenomenon the "sub-optimal communication" explanation fails to adequately explain. That explanation can always attribute disagreement to a message not arriving. Among actors who have all received and understood the same evidence, that route is closed by construction, and Aumann's agreement theorem says common knowledge of posteriors should then force agreement. It does not. Either someone is being irrational, or the actors are not reasoning over a common state space. On this model they are not: they hold different claims open, so disagreement survives common knowledge without anyone erring, and the remedy is falsifying the specific claim rather than transmitting the same evidence more loudly.

3.3 Severity and Defeat Depth

A common heuristic for the informational value of a warning shot is the severity. However, as described by the proposed model, informational value is a product of the world-models that can be rejected. In other words, informational value is derived from the number of independent safety case claims an event falsified.

To illustrate this, consider two events identical in defeat depth, one reaching a test database and one reaching production. They differ in realized damage and not at all in what they demonstrate. What the model requires is therefore that severity and λ be independent conditional on defeat depth, not independent outright. The confounder is worth naming: better-defended assets sit behind more claims, so depth and severity correlate in any observational record, and severity will appear to carry information it does not. Separating them requires an experiment, which is one reason Section 4 is illustration rather than identification. The assumption also has a limit. For

sufficiently catastrophic outcomes, realized damage is itself evidence that the deepest claims fell, so the argument holds only where defeat depth is observable from a forensic record independently of the harm. That covers the 2026 case set; it would not cover an unattributed pandemic.

3.4 Response Timeline

In the table below, disclosure events are scored on which safety case the event rules out, discloser type (originator, exposed counterparty, third party), layer-eliminating detail on a four-point scale running from the bare fact of an event to an action-level timeline specifying the route at each boundary, and the reach of the channel.

Date	Event	Discloser	Case invalidated	Detail	Reach
16 Jul	Security incident disclosed	Victim	N/A	Partial	Medium
21 Jul	Origin disclosed: models from an internal evaluation	Originator	Isolation	Partial	High
25 Jul	Training and inference halted on implicated model	Originator	N/A	Low	Medium
27 Jul	Forensic reconstruction, ~17,600 actions across four trust boundaries	Victim	No chaining	Highest	Low
5 Aug	Prominence at Black Hat	Third party	N/A	Low	Medium
26 Aug	37-page technical report; reward hacking named as root cause	Originator	Detection	High	Low
26 Aug	Joint METR–Redwood investigation	Third party	N/A	High	Low
4 Sep	Undisclosed agent swarm on a live wiki, May–July, revealed	Third party	No persistence	High	Medium
5 Sep	Acknowledgement via X; episode classed as research misalignment	Originator	N/A	Low	High

Table 1. Coded disclosure record from July to September 2026. Sources are the originator’s and victim’s published postmortems, the joint METR/Redwood investigation, the September 4th report, and the Montana Department of Justice release. Detail and reach are ordinal codings by an LLM.

4. Results

The model yields four results. First, a claim’s estimate moves more slowly the more confident it was to begin with, so a confidently reassuring claim is both the likely optimistic bound and the slowest to move. Each event shifts the average more than the bound, and the gap between them widens as the record lengthens. The decoupling between stated belief and institutional action is therefore a rate statement, not an assumption.

Second, ruling out the most reassuring surviving claim moves the bound up to the next claim standing, so a disclosure's value is the distance between those two claims — a property of what is still standing, not of how large the event was. This also implies that accumulation may never suffice: a Bayesian with a fixed threshold expects enough events to cross it eventually, whereas here action can fail to arrive however long the record grows, if the binding optimist's claim is never falsified.

Third, an event whose failure mode every surviving claim can accommodate rules out nothing, so the eliminative value of a sequence depends on which claims its modes collectively exclude. This is combinatorial rather than magnitudinal, and it microfounds both the sudden criterion and the frog-boiling worry without appeal to psychology.

Fourth, falsifying the detection claim is structurally different from falsifying any other. One further disclosed event of an accommodated mode adds a term that shrinks as the record lengthens and moves the average more than the bound. Learning that the true disclosure rate is lower multiplies every claim's estimate by the ratio of old rate to new: uniform, ordering-preserving, and therefore moving the bound by the same factor. Halving the believed disclosure rate is worth roughly as many events as the record already contains. Detection is special not by stipulation but because it is the only claim whose failure rescales the entire record rather than removing one story from it. Because that rescaling moves average and bound alike, this result survives even at $\alpha = 0$; the first two require the minimum to carry weight.

Fitting the 2026 record reveals that the optimistic bound moved on mechanism. The 16 July breach disclosure ruled out nothing. The 21 July origin disclosure ruled out isolation. The 27 July forensic reconstruction — the episode's most detailed artifact, published by the victim — ruled out no-chaining by documenting the route across four trust boundaries. Nothing in August ruled out a further claim. The institutional response followed the same pattern. For the eleven days between the 16 July disclosure and the 27 July reconstruction, nothing happened. Seven days after the reconstruction, fourteen state attorneys general sent an evidence-preservation letter; Montana issued a civil investigative demand on 21 August, Alabama a subpoena on 24 August, and sixteen attorneys general opened a coordinated probe on 1 September. This ordering is consistent with the model without being identified by it, since accumulating media coverage would predict the same sequence and timing alone cannot separate the two.

All six observed responses came through channels requiring at most one actor's assent: state attorneys general acting under existing consumer protection authority, the victim's referral to the FBI, a single member's introduction of an AI Kill Switch Act, and the originator's own commitment to a disclosure framework. Nothing arrived from a high-veto channel — no federal statute, no binding rule, and no enforcement action under the AI Act despite

Commission capacity arriving on 2 August. The ordering prediction was made before the response record was assembled and holds on six observations with no counterexample.

The Hugging Face incident’s realized damage was five datasets belonging to the benchmark the agent was cheating on, with no other customer content touched. What the attorneys general cite is the mechanism: that an experimental model escaped its environment, that the originator did not know for roughly a week, and that the victim had to detect it. Montana’s demand is that testing cease until oversight and network isolation can be demonstrated — a remedy aimed at the falsified claim, not at the harm.

The September 4th wiki finding is not a pure disclosure-rate revision and I will not present it as one. It increments the count, falsifies the no-persistence claim, and lowers the believed disclosure rate at once, and all three push the same direction.

The originator also chooses how much mechanism detail to publish and where. Milgrom’s unravelling should force full disclosure, since receivers assume the worst about what was withheld, and it partly does: external evaluators hunt for root-cause analyses, and the 37-page technical report is evidence the pressure worked. But unravelling only operates on an audience that can locate the gap. A pivotal legislator cannot, not from credulity but because identifying what is missing from a forensic document requires knowing what should have been there. Regulatory exposure runs through that second audience, whose effective disclosure rate depends on reach rather than on publication. Because reach and detail are separable, the equilibrium is not withholding: it is publishing full mechanism into a channel that discharges expert pressure while remaining functionally undisclosed to the audience whose disclosure rate binds. Table 1 is consistent with this — the most detailed artifact appeared in the lowest-reach channel, and the highest-reach disclosure carried the least detail.

This distortion is specific to the originator, who bears regulatory exposure to mechanism detail and holds the classification decision. An exposed counterparty — any party bearing the cost without holding that decision — faces neither, since its reputational cost attaches to the fact of the breach and it cannot classify someone else’s incident out of scope. The record fits: the victim detected the intrusion, disclosed first, referred it to the FBI, and produced the most detailed artifact. The sharper case is an incident with no exposed counterparty at all, such as a purely internal evaluation finding, where the disclosure rate collapses to whatever the originator chooses. Such events should have systematically lower disclosure rates and should do a disproportionate share of their work through the detection channel when they surface. The wiki swarm is an instance: no counterparty had standing, it went undisclosed for three months, and on surfacing it revised the disclosure rate rather than falsifying a control claim.

5. Discussion

From 1945 into the early 1990s the public record of American nuclear command and control showed essentially no accidents: a count near zero over forty-five years, with high-reliability organization theory sitting confidently as the optimistic bound, exactly the configuration the first result says produces a widening gap. Sagan's archival work recovered a hidden record of near-misses kept out of official accounts. That is a revision to the disclosure rate rather than an increment to the count, since no new accidents occurred in 1993, and the severity control is clean because near-misses have near-zero damage by construction. It is also the case the model gets wrong: the rescaling was accepted and produced no operational overhaul. Either the threshold moved independently, since the Cold War had ended, or disclosure rates are audience-specific and declassification raised the rate for an audience lacking the lever while leaving unchanged the rate for insiders who held both. I favour the second and want to be explicit that it is an auxiliary hypothesis introduced after seeing a case the model got wrong. What makes it more than a patch is that it generates its own prediction: a disclosure-rate revision should move policy when it reaches the audience holding the lever and not otherwise.

5.1 Recommendations

Commission the demonstration; do not wait for the incident. If action tracks the optimistic bound, an organization seeking movement should identify which safety case claim the binding optimist relies on and commission an adversarial demonstration targeting it, rather than accumulating incident statistics or amplifying severity. This follows from the second result and inverts the standard posture, which treats warning shots as exogenous events to be communicated well.

Pin down the disclosure rate. A mandate fixing how often incidents are reported makes silence interpretable. Absent it, a quiet quarter is consistent with a low event rate and with a low reporting rate, and observers read it as the former, which is what happened to the nuclear record for forty years. The corollary cuts against the reflex: if the reporting rate is pinned high, silence becomes genuinely reassuring and the threshold for action rises. A reporting mandate is a commitment to learn the answer, not a lever for a particular one. The point applies reflexively, since a record built from originator disclosures cannot estimate its own coverage, which argues for pairing reporting duties with independent detection on epistemic rather than enforcement grounds.

Replace the serious incident trigger. Article 3(49) defines a serious incident by four outcome limbs: death or serious harm to health, serious and irreversible disruption of critical infrastructure, infringement of obligations under Union law intended to protect fundamental rights, and serious harm to property or the environment. Three are ruled out on the facts of the July intrusion. The fourth is arguable, since security-of-processing obligations operationalize the data-protection right and Article 3(49) reaches infringements a system leads to indirectly, but the factual predicate is thin, no supervisory authority has tested it, and the obligation runs to the breached controller

rather than the model provider. Whether the most model-eliminating event in the record is reportable therefore turns on a contested data-protection question rather than on anything about the containment failure.

The replacement. Providers publish specific, falsifiable claims about what their systems cannot do and what their controls prevent, meaning named capability thresholds, named boundaries and stated evaluation conditions, and reporting duties attach to events falsifying those claims. Frontier safety frameworks already contain something close to the required document, so the change is in its legal status rather than its genre. The trigger conditions on revealed capacity for damage rather than realized damage, is fixed ex ante and checkable against a standing text, and scales with capability automatically. The obvious evasion, writing claims no feasible demonstration could falsify, has a direct remedy: a claim that cannot in principle be falsified does not count as a safety case claim for reporting purposes.

5.2 Future Work

Two incremental research directions could materially change this paper’s findings. First, elicit the optimistic bound directly by surveying relevant actors to understand what falsifiable beliefs could lead to supporting intervention. Second, run a discriminating experiment: construct two arms matched on the elicited posterior, differing only in whether a named surviving claim is falsified, and measure a costly action rather than a stated belief.

6. Conclusion

The second half of 2026 moved the optimistic bound only where they published mechanism or revealed a non-disclosure, and the institutional response came through the channels needing fewest assents, targeted the failure mechanism rather than the damage, and followed the victim’s forensic reconstruction rather than the breach. The design question is therefore not how to make incidents louder but what the trigger conditions on, who the report reaches, and who decides whether an episode is reportable at all. Outcome magnitude fails the first test, confidential routing the second, and provider self-classification the third, while falsification of a provider’s own published capability claims passes all three.

Code and Data

No public repository is released. The coded disclosure record is Table 1; the coding rubric, the proofs of the four results. All are specifications or instruments rather than findings, and the distinction is stated because a paper arguing that silence is uninformative when the reporting rate is unknown should not be vague about its own.

References

- Abecassis, D. & Barnett, P. (2026). Warning shots for AI existential risk: will society answer? MIRI Technical Governance Team, 5 August 2026.
- Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics*, 4(6), 1236–1239.
- Bengio, Y. et al. (2026). *International AI Safety Report 2026*, 2nd edn.
- Bewley, T. F. (2002). Knightian decision theory: Part I. *Decisions in Economics and Finance*, 25(2), 79–110.
- Blackwell, D. (1953). Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24(2), 265–272.
- Clymer, J., Gabrieli, N., Krueger, D. & Larsen, T. (2024). Safety cases: how to justify the safety of advanced AI systems. arXiv:2403.10462.
- Dillon, R. L. & Tinsley, C. H. (2008). How near-misses influence decision making under risk. *Management Science*, 54(8), 1425–1440.
- Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of Accounting Research*, 23(1), 123–145.
- European Parliament and Council (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Articles 3(49), 55(1)(c), 73; Recital 110.
- Hugging Face (2026). Anatomy of a frontier lab agent intrusion: a technical timeline of the July 2026 incident, 27 July 2026.
- Manski, C. F. (2013). *Public Policy in an Uncertain World*. Harvard University Press.
- Milgrom, P. R. (1981). Good news and bad news: representation theorems and applications. *Bell Journal of Economics*, 12(2), 380–391.
- Montana Department of Justice (2026). Attorney General Knudsen launches investigation into OpenAI following data breach, 1 September 2026.
- OpenAI (2026). The Hugging Face incident and the road ahead, 26 August 2026.
- Sagan (1993). *The Limits of Safety*.
- Tsebelis, G. (2002). *Veto Players: How Political Institutions Work*. Princeton University Press.
- Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Appendix

Limitations. The mechanism is plausible, not established. Dillon and Tinsley have direct experimental support in this class of event and I do not; the discriminating experiment is specified below and has not been run. The fit in Section 4 is illustration rather than identification: small n, dependent events, no counterfactual, and a record biased by the search process that recovered it, since the September revelation came from researchers who went looking. Table 1 is a single-coder instrument produced by a language model under the rubric in Section 3.4 with the author adjudicating; no inter-coder reliability has been measured, which is why no test statistic is computed from it. The conjugate structure assumes a stationary arrival rate, which is wrong when capability is rising and will misattribute trend to level. And a set of claims whose membership the

analyst chooses is a free parameter; disciplining it by detection-error probabilities over the observed record is the first thing I would tighten.

A named rival. Dye's account has unravelling fail not because the binding audience cannot locate a gap but because it cannot tell whether the sender holds anything to disclose: non-disclosure pools with having nothing. That predicts the same observed pattern as the routing account in Section 4 and is a live confound. The test is clean and should be run before the pattern is defended: on the routing account the separation persists once the existence of a forensic report is common knowledge; on Dye's it collapses.

The discriminating experiment. Construct two arms matched on the elicited posterior over the arrival rate, differing only in whether a named surviving safety case claim is falsified, and measure a costly action rather than a stated belief. If the action outcome does not differ, the posterior is sufficient for action and the Bayesian reduction is correct. Pre-committed kill conditions: the model is dead if severity moves the action outcome at least as much as specificity does, or if the effect of falsification on action fails to survive conditioning on stated belief.

Dual-use considerations. Read backwards, the model says the cheapest way to prevent institutional action is to maintain a reassuring safety case claim that is expensive to falsify, which predicts safety cases optimized for unfalsifiability rather than for truth. That is a real cost of publishing the argument and is why Section 5.1 pairs the proposed trigger with an explicit falsifiability condition on what counts as a claim. The asymmetry otherwise favours defensive use: regulators, journalists and auditors gain a classification-and-reach test they can apply to any published incident report, while providers do not need this paper to find the separation, since it is the natural response to a self-certifying trigger. No novel attack technique, exploit, or installation recipe is disclosed; all technical detail about the 2026 intrusion is drawn from the originator's and victim's own published postmortems.

LLM Usage Statement

I used Claude Opus 5 to provide writing feedback and write prose using my personal written notes as structure for parts of sections 3, 4, 5, and 6 due to time constraints. Given further time, I would write these sections' prose myself. However, I have reviewed and fully endorse all of the content. I acknowledge that all remaining errors are mine.