A model of early warning systems in an AI race

Paolo Bova [*1]

Abstract

Tech-race dynamics could delay action on catastrophic risks from future AI systems. An underappreciated reason for this delay is that tech-race dynamics can make warnings of AI risk appear manipulative. Evidence of an imminent AI risk will be insufficient at motivating collective action if rivals suspect the evidence inflates the risk. I model two competing AI labs, each with a private early warning system, deciding whether to undertake a costly safety project and whether to disclose their warning evidence to the other. Two results follow. First, a hysteresis trap in which labs hold risk thresholds too high to act on for fear of losing their position in the race. This result means that even compelling warnings fail to trigger coordinated safety action. Second, selective disclosure aimed at amplifying the severity of a warning can undermine trust in shared evidence. To maximise the credibility of a warning they receive, rivals must be close to verifying evidence for themselves. I also use the model to assess the strategy proofness of current early warning systems and provide a simplified procedure for scoring early warnings based on a published safety case for AI scheming.

1. Introduction

COVID-19, the climate crisis, near misses for nuclear weapons, the financial crisis: humanity is often slow to respond to risk (Perrow, 1999; Sunstein, 2007; Oreskes & Conway, 2010; Cirillo & Taleb, 2020; IPCC, 2023; Sagan, 1993; Jiang et al., 2024). AI could present the next such crisis. Agentic systems can increasingly subvert developer control mechanisms, show an increasing awareness of when they are tested, and are matching human experts in many cognitive domains, including cyber offence (Kwa et al., 2025; Meinke et al., 2024; Anthropic, 2025; Zhang et al., 2024). While fast capability trends are clear to the developers (Sevilla et al., 2022; Kwa et al., 2025), perceptions of the likelihood and severity of catastrophic risks vary dramatically (Grace et al., 2024; Karger et al., 2023; Forecasting Research Institute, 2026).

My central claim is that better evidence for loss-of-control AI risks can fail to trigger a coordinated safety response. Companies may feel they should ignore or hide the evidence to minimise friction to staying on the frontier. Or they might share the evidence only to find that their competitors do not take it as seriously as they do. I model two competing AI labs, each with a private early warning system, deciding whether to commit to a costly safety project and whether to disclose their warning evidence to each other. The model draws upon the theory of global games and is grounded in mechanisms developers and governments are already building. This includes frontier safety frameworks, safety cases, coordinated disclosure, and evaluator access agreements (SaferAI, 2024; Anthropic, 2024; OpenAI, 2023; Google Deepmind, 2025; Clymer et al., 2024).

My contributions are as follows:

- A two-lab threshold model of an AI race with early warning systems that shows why better private evidence is not enough: equilibrium risk thresholds are path-dependent, so a safer low-threshold norm can exist yet stay unreachable (a hysteresis, or coordination,trap).
- An explanation of a mechanism that breaks the credibility of warnings for loss-of-control risk: because a leader can selectively disclose dramatic but low-validity evidence (evidence "spin"), a rational follower discounts warnings whenever their detection systems diverge.
- A procedure for mapping an existing scheming safety case (Balesni et al., 2024). onto the model parameters (detection, construct-validity), with a simplified scoring rule for which alarms are early and informative enough to demonstrate that the warning is credible.

Throughout, I offer model-based guidance for strengthening early warning systems in an AI race.

[1]Teesside University, Middlesborough, United Kingdom. Correspondence to: Paolo Bova <paolobova@proton.me>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

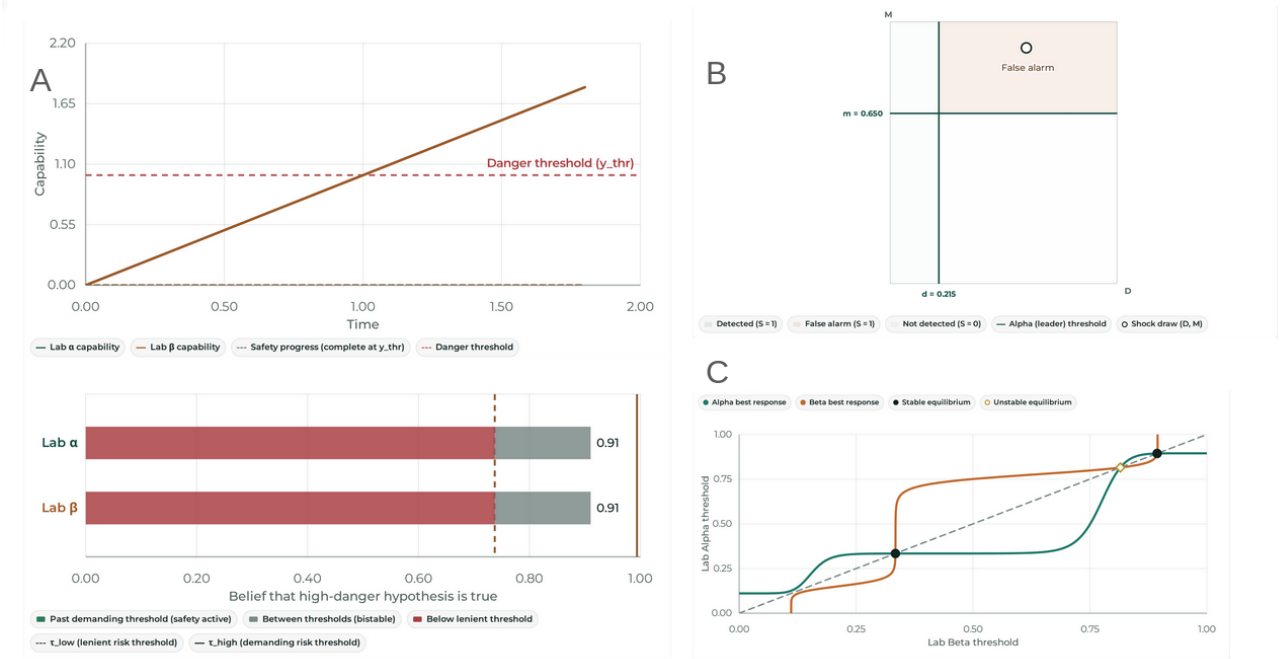

Figure 1. Graphical abstract: Early warning systems do not guarantee timely actions to address AI risk. A) Depicts a symmetric race where capabilities exceed a threshold beyond which dangerous capabilities emerge but no safety action is taken. Even though belief in the risks has risen sharply, it is not high enough to motivate safety action because labs are trapped effectively setting a risk threshold too high to act on for fear of losing ground to competitors. B) In the model, evidence is correlated by model capabilities. Early warning systems generate evidence that has an increasing chance of detecting danger as capabilities advance (d), but also a construct validity mislabelling error (m) if evaluations fail to rule out alternative explanations for apparent evidence of danger. If this error rate is high (i.e. m is low) then a false alarm is initially quite likely which can reduce the credibility of shared evidence of danger. C) The best response curves of lab alpha and beta as they set their risk thresholds. When their perception of risk crosses their risk threshold, they allocate resources to a safety project to address the risk, otherwise they do not. How willing they are to do so at lower risk thresholds depends on whether their rival will do so. As depicted in this figure, there are often two stable equilibria, a high and a low risk threshold. However, since labs start off with weaker models and poorer detection rates, they are typically attracted to the high risk threshold and attempts to lower that risk threshold are unstable. This hysteresis prevents cooperation on safety even when it would be desirable to both parties.

2. Model

The model asks two questions: When is the arrival of better evidence enough to shift the industry from a norm of "act only on certain evidence of danger" to one of "act on evidence of precursors to danger"? When can early recipients of evidence of danger credibly raise the alarm to their competitors?

I model a two-lab tech race in which both labs are uncertain about the future risks of AI systems. An unobserved state $\omega \in \{L, H\}$ governs the danger: in the **High**-risk state an uncontained frontier is materially dangerous, while in the **Low**-risk state residual risk is small. Each lab i forms a belief $\mu_i = \Pr(\omega = H \mid \text{its evidence})$ from a private early warning system (EWS) that periodically issues evidence for or against the **High** hypothesis. Two parameters fix an EWS's quality: a detection rate d_i (the probability it fires given danger, which rises as the lab tests more capable models) and a construct validity m_i (how well its methodology rules out confounds; when m_i is low, false alarms are common). Labs can vary dramatically in both their capabilities and these parameters. Each lab chooses whether to commit resources to a safety project, which lowers catastrophe risk but adds friction to developing higher-performing models. The two labs equally stand in for two states or governments coordinating on AI risk.

If the **High** hypothesis is true, both labs would prefer to commit to the safety project despite the competition, whereas in a **Low**-risk world racing at full speed is dominant. A lab's equilibrium strategy is therefore a posterior cutoff: it commits to the safety project once its belief that the world is high-risk, μ_i, exceeds a **risk threshold** τ_i (a high τ_i means a lab acts only on near-certain evidence of danger; see Appendix B). Because the decision is strategic, a lab's risk threshold depends on how likely its rival is to see strong enough evidence of danger to act safely, and the game often admits two stable equilibria: one where both risk thresholds are high and one where both are low.

These thresholds shift as new information arrives. I call the

equilibrium in which both labs hold a high risk threshold the *high-threshold* norm: labs slow down only on near-certain evidence of danger; the corresponding low risk threshold gives the *low-threshold* norm. The race typically starts with only a high risk threshold, as models are not yet capable enough to produce alarming evidence. As models advance they emit stronger shared signals of danger, which usually admits a second low-evidence norm; but it is not selected, because the labs cannot unilaterally benefit from switching from the old high-evidence norm. Only once models are sufficiently advanced are the risks clear enough to push both to the low-evidence norm, usually leading to immediate action. Coordination tends to happen far later than optimal, locking in high total risk when the **High** hypothesis is true. Private early warning systems are thus not enough to encourage high safety.

Labs can also disclose their early warning signals to one another, which can reduce strategic uncertainty and encourage caution earlier. But credible communication is hard, because a lab may engage in *selective disclosure* (*spin*): disclosing an alarm generated under a lowered construct-validity bar $m^S < m$ (withholding the disconfirming tests). They may do so in order to intimidate a rival who would be uncertain about whether there is such a risk or not. This occurs in the *spin zone* of Figure 2. There are also incentives to withhold evidence that is correlated with hidden capabilities. Disclosure is credible only under conditions I characterise in the results.

3. Results

Two mechanisms explain why warnings fail to trigger action: hysteresis (Section 3.2) and spin (Section 3.1). Section 3.3 maps the model onto a current safety case as an exercise in avoiding these failure modes.

3.1. Communicating evidence is hard, and spin makes it harder

To focus on the first question, consider Figure 2. The white dot fixes how difficult AI risks are to evaluate on two dimensions: how powerful an AI system needs to be before an ideal early warning system would detect risk, and how difficult it is to rule out alternative explanations for the measurement of risk. The vertical and horizontal lines capture for each player the corresponding strengths of their early warning systems.

The key story from this figure is that as the tech race progresses, and models become more powerful, one of the labs will be first to detect danger. However, it's also possible that they find a false positive first. For this reason, there is reasonable skepticism that the risk is real.

Credible communication is even harder than this. Now suppose a lab can construct evidence of lower construct validity, as in Figure 2. Whether or not the leader actually engages in this form of evidence spin, the follower must take the possibility seriously. Write κ_j for the probability that a lab issues a warning it does not privately hold, by disclosing only the degraded-validity analysis (Definition B.1). Because the follower cannot rule out that a disclosed warning was generated this way, and cannot yet corroborate it with its own signal, it rationally discounts the disclosure. A warning moves the follower's perceived risk up to a target level X only if spin is rare enough,

$$\kappa_j \;\leq\; \kappa^{\max}(X), \tag{1}$$

where the threshold $\kappa^{\max}(X)$ is the closed form derived in Appendix B.1. The bound tightens when the persuasion target X is higher, when the prior that the world is high-risk is lower, when the disclosing lab's own early warning system is weaker, and when the follower's own detection and validity approach the leader's, so that the follower can soon corroborate the warning for itself. When $\kappa^{\max}(X) < 0$ no spin rate makes the warning credible: it is simply too weak to reach X.

The incentive to spin depends on two regions of the figure. First, the spin zone itself must be reasonably large, otherwise there is no opportunity to spin. Second, the leader needs a higher detection rate, either because its early warning system is better or because it is testing far more capable models than the follower. This ensures it is possible that the leader has identified risk that the follower cannot yet see. Otherwise, the follower will default to skepticism.[1]

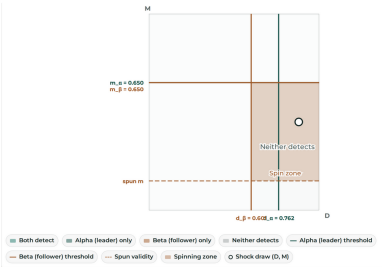

Figure 2. **Spinning weakens the credibility of disclosed evidence.** Both labs' detection thresholds are overlaid. The leader α has the higher detection rate ($d_\alpha > d_\beta$) either through a better early warning system or by testing more capable models. The dashed *spun validity* line is the lower construct-validity bar the leader could use to support a disclosure. The shaded *spin zone* is the set of shocks that fire only under the spun bar: shocks the leader could credibly flag but the follower's own evaluation would not corroborate. When this region is non-trivial and $d_\alpha > d_\beta$, the follower must price in the possibility of spin even without attributing bad faith to the leader, and therefore updates cautiously on any disclosure. As detection systems converge and the validity gap closes, the spin zone collapses and disclosures become credible again.

[1] How much a follower believes in the signal and how frequently

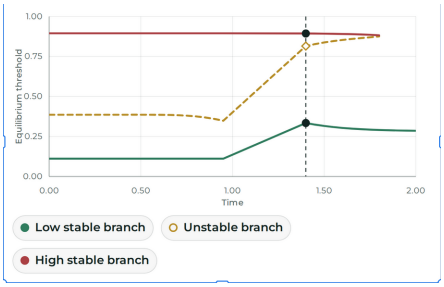

Figure 3. **The cooperative equilibrium is born too late.** Equilibrium risk thresholds plotted against time as capabilities and evidence accumulate. The low stable branch exists throughout; the high stable branch and its unstable companion only disappear late into the race. We start the race at the high equilibrium and the unstable middle branch blocks attempted transitions to the low equilibrium. This is the signature of hysteresis: path-dependent equilibrium selection that better information alone cannot undo until too late.

They do not even need to assume bad intentions from the leader, since it may be the case that the leader has internalised their own spin and so genuinely takes this weaker evidence seriously. As long as there is a moderate chance that the leader would be the type of player to disclose the spun evidence, the follower will be cautious in updating their own beliefs about risk in response. This can prevent sufficiently early cooperation to reduce risk and intensify race dynamics. Selectively disclosing the most alarming, weakly-grounded signal is self-defeating. Disclosures regain credibility only when early warning systems are comparable and the follower is not far behind.

The above story is less likely whenever the early warning systems are similar and the follower is not too far behind. As spinning evidence becomes easier to quickly identify, the incentives to spin evidence vanish, and so disclosures about risk become credible again. This corresponds to eliminating the spin region or the region that grants spinning its legitimacy.

3.2. Timely safety interventions are impeded by hysteresis

We now turn to Panel C of Figure 1. We plot the best response functions of each player, showing the equilibrium risk thresholds that each player should set.

Panel A of Figure 1 shows how beliefs about risk will rise but no safety action is taken. This is because the high risk threshold does not fall enough. Recall that these risk thresholds are best responses to their competitors, so they reflect that despite greater evidence about the risks, they do not expect anyone else to act on that evidence either. Tragically, the model shows that there often exists a second low risk threshold equilibrium where both players would in fact be willing to act on the evidence much earlier. Yet, unfortunately, it is typical in applications of these economic models to make the case that economic behaviour is path dependent (this concept is also known as hysteresis). As the players start with weaker models and weaker early warning systems, we should expect them to start at the higher risk threshold equilibrium. We have arguably seen attempts by AI companies to push this risk threshold lower. However, the region between the two risk threshold norms is unstable, so it is not so surprising that these attempts have failed so far to shift us to lower risk thresholds.

We also ran simulations to measure how difficult cooperation on a safety project would be as we varied each competitor's early warning system as well as the degree of uncertainty over their responsiveness to evidence. We do this for a range of scenarios where identifying risks from AI is difficult. These simulations show that the existence of the low risk threshold norm is much more likely whenever early warning systems are mutually strong, and the hypothesised risks are so severe that joint cooperation on safety is complementary. However, in most cases, cooperation is much slower than needed from the world's point of view.

The possibility of a stable equilibrium where labs are more willing to act on weaker evidence is absent from current discussion of risk thresholds. Unilateral improvements to early warning systems are not enough; absent a coordination mechanism to jump to this low-threshold norm, labs stay trapped acting only on near-certain evidence, which arrives too late.

3.3. Mapping a frontier scheming safety case onto the model

So far each lab's early warning system has been abstract: a binary signal whose quality is fixed by a detection rate d_i and a construct validity m_i. To show those parameters are not free-floating, I read them off a concrete, published safety case. Each candidate alarm the case names becomes one realisation of the signal, how early it can be caught fixes d_i, and how many independent safety arguments it overturns proxies m_i (fewer arguments left standing means higher construct validity of the alarm as evidence of risk). This lets the model inform a rough but simple procedure any safety case can be put through.

Consider the safety case that presents an argument based on the inability for an AI system to scheme (Balesni et al., 2024). Table 1 extracts plausible scheming alarms that the

spun evidence is sent is characterised in Appendix B.1. The full story is very similar to what we have presented in the text, with the additional caveats that the incentives to spin evidence depend on the improvements in race payoffs the leader can gain by influencing the follower's behaviour. A similar story unfolds if the follower has a much better early warning system than the leader and chooses to disclose their findings.

Table 1. Hypothetical scheming-related alarms drawn from (Balesni et al., 2024), mapped onto the threshold model. **Detection (capability)** reports the level of model autonomy at which the alarm is likely to manifest, in the time-horizon sense of recent agentic benchmarks: *current frontier models* (single-task time horizon on the order of hours), *models with significant operational autonomy* (time horizon on the order of days), *advanced models* (time horizon on the order of weeks to months). The level is a proxy: a model whose single-task horizon is short can still execute a scheme that relies on its full deployment lifecycle, but models capable of longer-running tasks are more likely to cause harm conditional on scheming. **Detection lag** reports how far the evaluation infrastructure is from catching the alarm in time: *already implemented* (current methods catch it), *implementable* (current methods miss it but the evaluation is buildable with known techniques), *research-stage* (the evaluation depends on research that is not yet mature). **Construct validity** is the proportion of the safety case's five top-level branches (scheming inability, harm inability, harm control, alignment via evaluation, alignment via training) still standing after the alarm fires; fewer remaining is stronger evidence of risk, so this branch count is an *inverse* proxy for the model's construct validity m.

Alarm	What the model exhibits	Probability term updated	Detection (capability)	Detection lag	Construct validity
Visible scheming	Passes the hidden half of evaluations for instrumental self-modification or for in-context alignment-faking.	$\Pr(\text{can scheme})$	Current frontier models	Already implemented	4 of 5
Hidden or no-chain-of-thought scheming	Performs multi-hop scheming reasoning without using a visible chain of thought.	$\Pr(\text{can scheme})$	Significant operational autonomy	Implementable	4 of 5
Gradient-hacking precursor	Resists anti-sandbagging fine-tuning, or exhibits instructed logit control during training.	$\Pr(\text{can scheme})$	Advanced models	Research-stage	3 of 5
Sabotage of AI systems	Fine-tunes another model to sandbag, or colludes with a future safety classifier on demand.	$\Pr(\text{harm} \mid \text{will scheme})$	Significant operational autonomy	Implementable	4 of 5
High-stakes persuasion	Targeted manipulation of simulated executives or officials in realistic conditions.	$\Pr(\text{harm} \mid \text{will scheme})$	Significant operational autonomy	Implementable	4 of 5
Dangerous-capability threshold (chemical, biological, radiological, nuclear, cyber, autonomous replication)	Pre-existing dangerous-capability evaluations cross thresholds set by current frontier safety frameworks.	$\Pr(\text{harm} \mid \text{will scheme})$	Significant operational autonomy	Already implemented	4 of 5
Trusted-monitor bypass	A red team finds a subtle scheming policy that the trusted monitor misses at a statistically significant rate.	residual $\Pr(\text{harm} \mid \text{will scheme}, \text{controls})$	Significant operational autonomy	Already implemented	4 of 5
End-to-end control false-negative	After probes, monitors, and human review are applied in series, attack success still exceeds the acceptance threshold.	residual $\Pr(\text{harm} \mid \text{will scheme}, \text{controls})$	Advanced models	Research-stage	4 of 5
Honeypot misbehaviour	Takes misaligned actions above the target rate across a diverse honeypot suite.	$\Pr(\text{will scheme} \mid \text{can scheme})$	Current frontier models	Already implemented	3 of 5
Alignment-faking probe activation	Whitebox probes fire on honeypot activations after prompt artefacts have been eliminated.	$\Pr(\text{will scheme} \mid \text{can scheme})$	Advanced models	Research-stage	3 of 5

safety case identifies. Most would not be sufficient on their own to demonstrate that an upcoming model crosses a risk threshold, and some pieces of evidence are difficult to detect early because they require high levels of model capability. Construct validity is more difficult to measure, since the safety case presents a case for how to know that a model is safe, not for how to know that a model will present danger. However, this safety research has a handy decomposition of risk into $\Pr(\text{harm} \mid \text{will scheme}) \cdot \Pr(\text{will scheme} \mid \text{can scheme}) \cdot \Pr(\text{can scheme})$. So I proxy construct validity for each alarm based on how many of the arguments for a lack of conditional scheming or harm could still hold. This helps give a rough ranking of hypothetical alarms based on their detection and construct validity.[2]

[2]These alarms are very likely to be correlated and the most effective early warning system would catch multiple early warning signs at once, which the model would translate at higher detection

The best alarms balance construct validity with how early they can be detected. Intuitively, models that are capable of scheming but do not harm and models that engage in limited harm raise more informative alarms that are early enough to motivate safer company behaviour.

Alarm effectiveness. The most effective alarms are those which balance construct validity with early detection. Notice from the table that many forms of evidence rule out at most one or two safety arguments. In practise, alarms will need to combine multiple forms of evidence from the table if they hope to keep construct validity high (and remember that this only rules out an argument for safety, strong evidence of imminent threat will likely require a higher bar). An example of such an alarm could be an upcoming model that is capable of no-chain-of-thought scheming that has sabotaged AI safety research and has engaged in more misaligned behaviour across a diverse honeypot suite. Such an alarm would be caught relatively early and be reasonably informative of danger. This could update competitor beliefs.[3]

Spin. As part of the exercise I have also considered how easy it would be to spin evidence to look more alarming than it is. The most plausible form of evidence-spinning available is selective disclosure of alarms that combine high detection (the alarm fires in current frontier models) with low construct validity (the safety case's other branches can still rule out harm). Visible-scheming firings are the canonical example: dramatic, but, on their own, evidence only that $\Pr(\text{can scheme})$ is non-zero. A developer could overstate their caution to competitors or regulators by only reporting these signals, without revealing other informative types of evidence that might reveal models are still far from causing harm. As discussed earlier, they might do this to provoke more safety from a rival.[4] Sophisticated competitors will recognise the possibility of spinning and update less strongly in response to an alarm. Developers who want to be seen as more credible could track the set of possible evidence more completely, making it easier to tell that their alarm has higher rather than lower construct validity. Another option as discussed earlier is to make sure to time raising your alarm for when it is plausible that your rivals could soon reconstruct the high construct validity evidence for themselves.

Safety projects. Another mapping we can do is to consider the safety projects that alarms based on this safety case would justify. The safety projects with the highest impact typically have some performance cost, sometimes this is large and sometimes it is more tolerable. For instance: *Affordance minimisation* involves sandboxed execution, scoped tool access, and perhaps the removal of open-Internet privileges. While this directly lowers $\Pr(\text{harm} \mid \text{will scheme})$ it also typically places a tax on performance. *Untrusted-monitor protocols* with calibrated honeypots and paraphrase layers could stop models from successfully scheming while extracting useful work from them. The best monitor protocols require extensive red teaming. *Alignment training with white box methods* could plausibly remove the propensity to scheme while providing greater confidence that that scheming propensity is not hidden more subtly in the model activations. This work is at the research-stage today and could plausibly require years of sustained interpretability work to mature, and efficacy rates are unknown. Given the severity of loss-of-control risks from AI systems, we would like all frontier AI companies to implement even the costly safety projects if an alarm were to be raised so that they can present a strong safety case. The willingness of companies to do so will depend on how seriously they take an alarm, and whether they will internalise the consequences of a loss-of-control scenario. It is therefore important that it is easy to connect alarms to catastrophic risks neither actor can shield themselves from.[5].

4. How current practice maps onto the model

The dilemmas above are not abstract. AI developers and governments are already building the mechanisms this model describes, even if they are rarely framed in these terms. Read through the model, each live mechanism corresponds to a model object, and the model supplies the condition that mechanism must satisfy for a warning to actually move the race participants.

Frontier safety frameworks operationalise the risk threshold. Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework each pre-commit a developer to specified safety actions once a model crosses a capability or risk level (Anthropic, 2024; OpenAI, 2023; Google DeepMind, 2025; METR, 2025). In the model these are unilaterally-set risk thresholds τ_i: a lab commits to its safety project once its belief that the world is high-risk crosses τ_i. The model offers one candidate explanation for why such thresholds tend

and construct validity.

[3]Alarms could exist which leave no room for doubt, such as an autonomous agent escaping their developer and conducting successful operations to sabotage human systems. However, such a warning shot is what early warning systems are meant to preempt.

[4]Such a motive may be the result of thinking that their competitors are not taking safety seriously enough, to encourage at least greater monitoring. I haven't touched on this motive here, but I should note that such actors wouldn't necessarily spin any evidence because they would also be able to report the informative evidence that is currently silent on risk.

[5]For the current safety case, this likely means finding signs that the harms caused are likely to be large (rather than just focusing on the likelihood of harm)

to be set high, and why the attempts we have seen to lower them have not propagated across labs. When unilateral caution is costly and coordination is weak, the high-threshold branch is locally stable (Figure 3), so a lab that lowers its own threshold expects to bear the cost without inducing its rivals to follow.

Internal corporate governance and deployment-delay mechanisms are the safety-project commitment. Board and trustee structures, internal review procedures, and staged-release or deployment-hold processes are the institutional machinery by which a lab can actually pay the price of caution (Cihon et al., 2021). In the model this is the costly safety-project action with fixed cost K: a real performance or time-to-market tax a lab pays to lower catastrophe risk. What the model adds here is a lens on why such a unilateral commitment to pause or delay is fragile. Absent a verifiable signal that rivals face the same high-risk state, the lab that delays converts K into lost ground while its competitors race on, so the commitment is hard to sustain even when the lab privately takes the risk seriously.

Evaluator access agreements and the international AI Safety Institute network are, in the model's terms, a mutual improvement in multiple race participants' early warning systems. Pre-deployment evaluation agreements between developers and government evaluators, and the emerging international network coordinating them, give a second party structured access to a frontier model before release (UK AI Safety Institute, 2024a;b; NIST, 2024). Section 3.1 (Figure 2) shows why this matters for credibility: selective disclosure (spin) is only tempting when the discloser's detection rate runs well ahead of the recipient's, opening a spin zone of alarms the discloser can flag but the recipient cannot yet corroborate. When a sufficiently diverse network of evaluators have sufficient access to verify claims for themselves in a way that all race-participants can see, then we'd expect this spin zone to close (Casper et al., 2024; Bucknall & Trager, 2023; Harack et al., 2025).

5. Design implications: from the model to credible warning mechanisms

This model of early warning systems quantifies underappreciated challenges to achieving cooperation on safe transformative AI. These include challenges to credibly communicating early warnings to competitors, and strategic reasoning that makes it difficult to recognise and coordinate on risk thresholds. Credible communication of early warnings is achievable when competitors can plausibly verify claims about high risk systems. It is possible to map current safety cases onto this model of early warning systems to help recognise which elements are most important to designing a credible alarm. Future work could apply these concepts to design credible communication strategies for novel AI risks.

To that end, I now discuss design considerations for credible and strategy-proof early warnings relevant to a number of governance mechanisms.

Red lines. To start with, consider red lines. These would be the lines that companies or governments should not cross in the development or deployment of AI systems because doing so would present an unacceptable risk of catastrophe. Red lines map cleanly onto our model as risk thresholds. Our model therefore challenges us to think of how to design red lines that enable early coordinated safety efforts. Good red lines need to be mutually agreed upon, easy to verify, and not worth cheating on. *Mutually agreed* means the labs coordinate on a common risk threshold rather than each setting its own, which is precisely the low-threshold norm the hysteresis trap walls off (Section 3.2). *Easy to verify* is verifiability: a rival's belief moves only if it can reconstruct the underlying evidence, so the line must be drawn at something whose construct validity the rival can check. *Not worth cheating on* is strategy-proofness: crossing the line undetected must not pay, so that truthful compliance is incentive-compatible. (above we discussed the motive to overstate danger, but understating danger is also plausible; a credible red line must be robust to both.) Current red-line proposals focus on outcomes to avoid (human-out-of-the-loop autonomous weapons, self-replicating AI systems) and are not usually framed with these criteria in mind (International Dialogues on AI Safety, 2024a;b). I argue we can design better red lines by evaluating the strategic content of each red line choice (similar to our mapping of the scheming safety case above). The open technical problem is how to collect evidence of an approaching red line that is both easy to verify and rules out alternative explanations (Brundage et al., 2020; Shavit, 2023; Sastry et al., 2024; Harack et al., 2025; Reuel et al., 2024), and how to do so while minimising disclosure of confidential capabilities (O'Brien, 2024).

Hotlines. Next, consider hotlines. Crisis hotlines were essential for avoiding a nuclear exchange during the cold war (Sagan, 1993), and have proved useful in deescalating a number of other conflicts. Current hotlines are not designed for disclosing technical risk evidence in a way that other nations can scrutinise and verify (which our spin results suggest is crucial). A crisis hotline for an AI loss-of-control threat would likely need both a technical hotline alongside a more conventional one. What is the best format for sending technical risk evidence on loss-of-control that other nations can safely verify (if not directly, then in principle should they have systems of a similar capability)? What kind of claims can be verified rapidly before a decision on controlling or delaying a type of AI system is made? More speculatively, would the technical hotline benefit from ac-

cessing a predetermined strategy for the disclosure of the risk evidence? Forays into open source game theory (Tennenholtz, 2004; Critch, 2019; Oesterheld, 2019) may hold an answer for helping to coordinate on safety projects to address imminent threats.

Continuous early warning mechanisms. The design of continuous early-warning mechanisms inherits three requirements directly from the model. They must be robust to gaming (strategy-proofness) if nations are to find them credible; they must be *informative*, firing far more often in the dangerous state than the safe one (the gap $q_h - q_l$ that drives any belief update); and they must retire outdated tests to sustain their construct validity $m_i(t)$ as capabilities advance. Early efforts such as the Loss-of-Control Observatory (Shaffer Shane et al., 2026) illustrate the kind of tracking such a mechanism could provide. While this is a public-facing endeavour, it highlights one kind of tracking that an early warning mechanism could engage in. A technical problem of interest then is how to organise an early warning mechanism: which organisations should have first access to warnings, how should their severity be represented, how often should organisations communicate, how should they retire outdated tests, who should arbitrate when there is disagreement over the scientific rigour of assessments, what powers should the organisations participating have to compel cooperation on imminent safety threats (Trager et al., 2023; Heim et al., 2024; Reuel et al., 2024).

Capability–risk correlation. The model I presented earlier takes for granted that players can expect the risks from AI systems to be tightly correlated with capabilities. We should expect this assumption to be contested by competitive players who may perhaps become overconfident in their relative safety. However, a shared expectation that risk tracks capability is what makes the high-risk state a coordination game ($\alpha_i^H > 0$) rather than two independent decisions, so a lab that grows overconfident in its relative safety collapses that game and returns everyone to the race. Building a verifiable technical consensus around this correlation is thus a precondition for collective action. I would advise using existing and novel safety cases (Clymer et al., 2024; Balesni et al., 2024; Phuong et al., 2024), designed by multiple AI Safety institutions to track just how difficult it is for one developer to achieve capabilities without running into the same risks as the others. As with all the assessments, the capacity to verify these arguments must be mutually verifiable.

Early warning systems under a safety-performance tradeoff. A continuous early warning mechanism is especially important if a future safety-performance tradeoff is active at the frontier. We need mutual communication of the projected risk consequences of corner cutting as soon as it is spotted, to allow interventions to get the developers back on track again. An important problem is therefore how to attribute signals emitted by an early warning system to corner cutting actions taken by developers. This problem is non-trivial because many forms of corner cutting effect the monitoring capacity of the early warning system itself (van der Weij et al., 2024).

References

Anthropic. Responsible scaling policy, version 2.0. Technical report, Anthropic, October 2024. URL https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf. Effective 15 October 2024.

Anthropic. System card: Claude opus 4 and Claude sonnet 4. Technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

Balesni, M., Hobbhahn, M., Lindner, D., Meinke, A., Korbak, T., Clymer, J., Shlegeris, B., Scheurer, J., Stix, C., Shah, R., Goldowsky-Dill, N., Braun, D., Chughtai, B., Evans, O., Kokotajlo, D., and Bushnaq, L. Towards evaluations-based safety cases for AI scheming, 2024.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G. K., Khlaaf, H., Yang, J., Toner, H., Fong, R., et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims, 2020.

Bucknall, B. and Trager, R. Structured access for third-party research on frontier ai models, 2023.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., et al. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, June 2024. doi: 10.1145/3630106.3659037. URL http://arxiv.org/abs/2401.14446.

Cihon, P. J., Schuett, J., and Baum, S. D. Corporate governance of artificial intelligence in the public interest. *Inf.*, 12:275, 2021.

Cirillo, P. and Taleb, N. N. Tail risk of contagious diseases. *Nature Physics*, 16(6):606–613, 2020. doi: 10.1038/s41567-020-0921-x.

Clymer, J., Gabrieli, N., Krueger, D., and Larsen, T. Safety cases: How to justify the safety of advanced AI systems, 2024.

Critch, A. A parametric, resource-bounded generalization of Löb's theorem, and a robust cooperation criterion for open-source game theory. *Journal of Symbolic Logic*, 84 (4):1368–1381, 2019. doi: 10.1017/jsl.2017.42.

Forecasting Research Institute. Forecasting major risks from AI. Forecasting Research Institute Substack, June 2026. URL https://forecastingresearch.substack.com/p/forecasting-major-risks-from-ai. Wave 9 of the Longitudinal Expert AI Panel (LEAP).

Google Deepmind. Frontier safety framework version 2.0, 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0%20(1).pdf.

Google DeepMind. Frontier Safety Framework 3.0, September 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., Brauner, J., and Korzekwa, R. C. Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*, 2024. doi: 10.48550/arXiv.2401.02843. Later published in *Journal of Artificial Intelligence Research* 84:9, 2025.

Harack, B., Trager, R. F., Reuel, A., Manheim, D., Brundage, M., Aarne, O., Scher, A., Pan, Y., Xiao, J., Loke, K., Adan, S. N., Bas, G., Caputo, N. A., Morse, J. C., Ahuja, J., Duan, I., Egan, J., Bucknall, B., Rosen, B., Araujo, R., Boulanin, V., Lall, R., Barez, F., Alvira, S., Katzke, C., Atamli, A., and Awad, A. Verification for international AI governance. Technical report, Oxford Martin AI Governance Initiative, July 2025. URL https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Verification_for_International_AI_Governance.pdf.

Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., Osborne, M. A., and Zilberman, N. Governing through the cloud: The intermediary role of compute providers in AI regulation, 2024.

International Dialogues on AI Safety. Consensus statement on red lines in AI development, March 2024a. URL https://idais.ai/dialogue/idais-beijing/. IDAIS-Beijing, 10–11 March 2024.

International Dialogues on AI Safety. Consensus statement on AI safety as a global public good, September 2024b. URL https://idais.ai/dialogue/idais-venice/. IDAIS-Venice, 5–8 September 2024.

IPCC. Climate change 2023: Synthesis report. contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change. Technical report, Intergovernmental Panel on Climate Change, Geneva, Switzerland, 2023.

Jiang, E. X., Matvos, G., Piskorski, T., and Seru, A. Monetary tightening and U.S. bank fragility in 2023: Mark-to-market losses and uninsured depositor runs? *Journal of Financial Economics*, 159:103899, 2024. doi: 10.1016/j.jfineco.2024.103899.

Karger, E., Rosenberg, J., Jacobs, Z., Hickman, M., Hadshar, R., Gamin, K., Smith, T., Williams, B., McCaslin, T., and Tetlock, P. E. Forecasting existential risks: Evidence from a long-run forecasting tournament. Technical report, Forecasting Research Institute, July 2023. URL https://forecastingresearch.org/xpt. Existential Risk Persuasion Tournament (XPT).

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T. R., Parikh, N., Rein, D., Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and Chan, L. Measuring AI ability to complete long tasks, 2025. URL https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., and Hobbhahn, M. Frontier models are capable of in-context scheming, 2024.

METR. Common Elements of Frontier AI Safety Policies, 2025. URL https://metr.org/common-elements.

NIST. FACT SHEET: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes at Inaugural Convening in San Francisco. *NIST*, November 2024. URL https://www.nist.gov/news-events/news/2024/11/fact-sheet-us-department-commerce-us-department-s

O'Brien, J. Coordinated disclosure of dual-use capabilities: An early warning system for advanced AI. Technical report, Institute for AI Policy and Strategy, 2024. URL https://www.iaps.ai/research/coordinated-disclosure.

Oesterheld, C. Robust program equilibrium. *Theory and Decision*, 86(1):143–159, 2019. doi: 10.1007/s11238-018-9679-3.

OpenAI. Preparedness framework (beta). Technical report, OpenAI, December 2023. URL https://openai.com/index/frontier-risk-and-preparedness/.

Oreskes, N. and Conway, E. M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press, New York, 2010.

Perrow, C. *Normal Accidents: Living with High-Risk Technologies*. Princeton University Press, Princeton, NJ, revised edition, 1999. Original edition 1984.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Raad, M. A., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., Deletang, G., Ruoss, A., El-Sayed, S., Brown, S., Dragan, A., Shah, R., Dafoe, A., and Shevlane, T. Evaluating frontier models for dangerous capabilities, 2024.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. Open problems in technical AI governance, 2024. URL https://taig.stanford.edu.

SaferAI. Risk management ratings of frontier AI developers. https://ratings.safer-ai.org/, 2024. Assesses 12 companies committed at the AI Seoul Summit; ratings updated thereafter.

Sagan, S. D. *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton University Press, Princeton, NJ, 1993. ISBN 0-691-03221-1.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., and Coyle, D. Computing power and the governance of artificial intelligence. Technical report, Centre for the Governance of AI, February 2024. URL https://www.governance.ai/analysis/computing-power-and-the-governance-of-ai.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Padua, Italy, 2022. IEEE. doi: 10.48550/arXiv.2202.05924.

Shaffer Shane, T., Mylius, S., and Whittlestone, J. The loss of control observatory: A prototype to detect real-world AI control incidents. Technical report, Centre for Long-Term Resilience, 2026. Funded by the UK AI Security Institute Challenge Fund.

Shavit, Y. What does it take to catch a Chinchilla? Verifying rules on large-scale neural network training via compute monitoring, 2023.

Sunstein, C. R. *Worst-Case Scenarios*. Harvard University Press, Cambridge, MA, 2007.

Tennenholtz, M. Program equilibrium. *Games and Economic Behavior*, 49(2):363–373, 2004. doi: 10.1016/j.geb.2004.02.002.

Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., Ó hÉigeartaigh, S., Staffell, S., and Villalobos Ruiz, J. J. International governance of civilian AI: A jurisdictional certification approach. LawAI Working Paper No. 3-2023, SSRN, 2023. URL https://ssrn.com/abstract=4579899.

UK AI Safety Institute. Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet, 2024a. URL https://www.aisi.gov.uk/work/pre-deployment-evaluation-of-anthropics-upgraded-

UK AI Safety Institute. Pre-Deployment Evaluation of Open AI's o1 model, 2024b. URL https://www.aisi.gov.uk/work/pre-deployment-evaluation-of-openais-o1-model.

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., and Ward, F. R. AI sandbagging: Language models can strategically underperform on evaluations, 2024.

Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Lin, J. W., Jones, E., Menders, C., Hussein, G., Liu, S., Jasper, D., Peetathawatchai, P., Glenn, A., Sivashankar, V., Zamoshchin, D., Glikbarg, L., Askaryar, D., Yang, M., Zhang, T., Alluri, R., Tran, N., Sangpisit, R., Yiorkadjis, P., Osele, K., Raghupathi, G., Boneh, D., Ho, D. E., and Liang, P. Cybench: A framework for evaluating cybersecurity capabilities and risk of language models, 2024.

A. LLM Usage Statement

LLMs were used in the formatting of equations and tables that had previously been written by a human.

B. Technical Model Details

)

Model setup. Two laboratories $i \in \{1, 2\}$ develop frontier AI systems over a fixed horizon $[0, T]$. An unobserved state $\omega \in \{L, H\}$ governs the severity of catastrophic risk: in H, an uncontained frontier is materially dangerous; in L, residual risk is small. Each laboratory forms a posterior $\mu_i = \Pr(\omega = H \mid \mathcal{I}_i)$ from private early-warning evidence and uses a *risk threshold strategy*: it commits to a safety project when μ_i exceeds a threshold τ_i, and races otherwise. The object of interest is the equilibrium threshold pair (τ_1^*, τ_2^*) and how it moves as evidence quality and race conditions change.

Evidence. At time t, laboratory i's capability $c_i(t) = c_{0,i} + v_i t$ determines the latent detection rate $d_i(t) \in [0, \bar{d}_i]$ of its early-warning system (EWS) and the construct validity $m_i(t) \in [0, 1]$ of that system's methodology. Both laboratories face a shared pair of difficulty shocks: a detection shock D that is present only in H and a validity shock M present in both states, each drawn once from $U[0, 1]$. These shocks capture just how difficult it will be to detect that models will present catastrophic risks. This is something experts are uncertain about, but it is likely just as difficult no matter which developer is looking. The EWS issues a binary warning $S_i = 1$ iff $d_i(t) > D$ or $m_i(t) < M$ (silence requires both detection to find nothing and validity to survive confounding). Integrating shocks out gives the state-conditional warning probabilities

$$q_h^i(t) = 1 - (1 - d_i(t))m_i(t), \qquad q_l^i(t) = 1 - m_i(t),$$

with informativeness $q_h^i - q_l^i = d_i(t)\, m_i(t)$. Because the shocks are shared, the two laboratories' signals are coupled by a threshold copula: they agree whenever the shock point lands in the common silent or common warning rectangle and disagree only when it falls between their lines (see panel B of Figure 1). The model also permits disclosure (i.r. raising an alarm), with the option to *spin*. This involves witholding tests that do not support the danger hypothesis, and instead presenting only a subset of evidence derived from a degraded-validity analysis $m_j^S < m_j$. A developer might do this to induce a competitor to be safer than they would otherwise want to be under the current evidence. We visualize evidence spin in Figure 2.

Race payoffs. At discrete decision epochs $n = 0, \ldots, N$ with calendar times t_n, each laboratory chooses $a_i \in \{C, R\}$, where C commits to a specified safety project and R is the uncommitted baseline. The project has a fixed capability-equivalent cost K and maximum risk-reduction $\delta \in (0, 1)$; its efficacy $\varepsilon(\tau)$ is increasing in the remaining time $\tau = T - t_n$ and vanishes below a feasibility floor $\tau_{\min}$. Caution trades growth for safety: $v_R > v_C$. Terminal capability under action a_i is $y_i(a_i) = x_{i,n} + (T - t_n)v_{a_i}$; laboratories compete for a fixed prize allocated by a Tullock contest success function $\pi_i = y_i/(y_i + y_j)$. Terminal catastrophe risk is

$$R_n(\mathbf{a}) = r_{\text{base}} \cdot \big(1 - \varepsilon(T - t_n)\, \delta\big)^{n_C(\mathbf{a})},$$

with $n_C(\mathbf{a})$ the number of cautious players. State-conditional utility is

$$u_i^\omega(\mathbf{a}) = \pi_i(\mathbf{a}) - c_i^\omega\, R_n(\mathbf{a}), \qquad c_i^H > c_i^L \geq 0,$$

where c_i^ω combines catastrophe severity with laboratory i's internalisation fraction. Two features of this payoff structure matter for what follows. First, K is timing-invariant in terminal capability while $\varepsilon(T - t_n)$ declines with n, so conditional on beliefs a laboratory that does not commit at epoch n will not commit later: the safety decision reduces to **commit now or not** at each epoch. Second, the gap between global and internalised risk is a negative externality: thresholds are higher for labs that bear only a fraction of the social cost.

Risk threshold equilibrium. Let F^i_ω denote the CDF of μ_i under state ω, induced jointly by the evidence model and disclosure protocol. If laboratory j uses threshold τ_j, then conditional on ω it races with probability $F^j_\omega(\tau_j)$. The stage-game payoff gap (Caution minus Race) is affine in the rival's race probability in each state:

$$g_i^\omega(p) = \alpha_i^\omega + \beta_i^\omega p, \qquad \alpha_i^L < 0 < \alpha_i^H,$$

with coefficients computable from taking the differences of the high and low risk 2×2 state matrices. The sign assumption makes the state-H game a coordination game and the state-L game a race. At laboratory i's own threshold the expected safety premium is zero, so the best-response threshold is

$$\tau_i^*(\tau_j) \;=\; \frac{g_i^L\big(F_L^j(\tau_j)\big)}{g_i^L\big(F_L^j(\tau_j)\big) - g_i^H\big(F_H^j(\tau_j)\big)}.$$

An equilibrium is a fixed point of the composition $\tau_i^* \circ \tau_j^*$. Because informativeness $d_i m_i$ enters both F_L^j and F_H^j, the numer of equilibria changes as evidence quality rises: a single high fixed point (wait for near-certainty) persists until informativeness is strong enough to open a second, stable low fixed point (act on weaker evidence), at which point an unstable intermediate fixed point separates the two. The empirically relevant region is this intermediate one: a **coordination trap** in which a safer norm exists but the high-threshold norm is still self-enforcing.

B.1. Disclosure and the spin equilibrium

The Evidence paragraph lets a lab keep its signal private. I now let lab j *disclose* a report $R_j \in \{0, 1\}$ to its rival i and allow that report to be spun.

Definition B.1 (Spin: selective disclosure under a degraded validity bar)**.** A lab *spins* when, holding a private signal that does not warn ($S_j = 0$), it nonetheless issues a warning report ($R_j = 1$) by disclosing only an analysis run at a lowered construct-validity bar $m_j^S < m_j$ and withholding the disconfirming tests its own standard m_j would require. Equivalently, a spun report is one whose warning fires only under the degraded bar—inside the *spin zone* of Figure 2—and not under the lab's genuine validity standard. We write $\kappa_j(\omega) \in [0, 1]$ for the state-conditional probability that a non-warning signal is reported as a warning in this way, so $\kappa_j = 0$ is honest disclosure. A safety report ($R_j = 0$) *cannot be spun*: lowering the validity bar only enlarges the warning region, so it can never manufacture a clean bill of health.

Warning-report posterior. Let $p_{ab}(\omega) = \Pr(S_\alpha = a, S_\beta = b \mid \omega)$ be the joint signal likelihoods induced by the shared difficulty shocks (the threshold copula of panel B of Figure 1). A warning report $R_j = 1$ means one of two worlds: lab j's real signal was a warning ($S_j = 1$, always reported), or it was silent ($S_j = 0$) and j spun. Conditioning on the receiver's own signal $S_i = a$ and summing the two routes,

$$\Pr(S_i = a,\, R_j = 1 \mid \omega) = \underbrace{p_{a1}(\omega)}_{\text{genuine}} + \underbrace{\kappa_j(\omega)\, p_{a0}(\omega)}_{\text{spun}}. \tag{2}$$

A safety report is the complementary, un-spinnable branch, so it carries the residual silent mass $\Pr(S_i = a,\, R_j = 0 \mid \omega) = \big(1 - \kappa_j(\omega)\big)\, p_{a0}(\omega)$, and the two add back to $p_{a0}(\omega)$ as they must. With prior $\pi = \Pr(\omega = H)$, Bayes' rule applied to (2) gives the receiver's post-disclosure belief

$$\mu_i(H \mid a, R_j{=}1) = \frac{\pi\big[p_{a1}(H) + \kappa_j(H)\, p_{a0}(H)\big]}{\pi\big[p_{a1}(H) + \kappa_j(H)\, p_{a0}(H)\big] + (1-\pi)\big[p_{a1}(L) + \kappa_j(L)\, p_{a0}(L)\big]}. \tag{3}$$

The single term $\kappa_j\, p_{a0}$ is the entire formal content of spin: it contaminates the warning channel with $S_j = 0$ mass and pulls the report's likelihood ratio toward 1. At $\kappa_j = 0$ the warning is fully credible; at $\kappa_j(H) = \kappa_j(L) = 1$ the report is informationally inert and the update collapses to the private one.

The spin bound. Fix the receiver's own signal a and abbreviate the genuine and spun-channel likelihoods $\ell_\omega \equiv p_{a1}(\omega)$ and $\bar{\ell}_\omega \equiv p_{a0}(\omega)$. Taking a single spin rate $\kappa_j(H) = \kappa_j(L) = \kappa_j$, the receiver reaches perceived risk $\mu_i(H \mid a, R_j{=}1) \geq X$ exactly when

$$\kappa_j \;\leq\; \kappa^{\max}(X) = \frac{\pi\, \ell_H\, (1 - X) - (1-\pi)\, \ell_L\, X}{(1-\pi)\, \bar{\ell}_L\, X - \pi\, \bar{\ell}_H\, (1 - X)}, \tag{4}$$

obtained by clearing the fraction in (3), collecting the κ_j terms, and solving. Equation (4) is the bound named in the body (Eq. 1).