The Coordination Gap in Frontier AI Safety Policies

Isaak Mengesha
Informatics Institute
University of Amsterdam
Amsterdam, 1098XH
i.a.mengesha@uva.nl

Abstract

Frontier AI Safety Policies concentrate on prevention: capability evaluations, deployment gates, usage constraints, while neglecting institutional capacity to coordinate responses when prevention fails. We argue this coordination gap is structural: investments in ecosystem robustness yield diffuse benefits but concentrated costs, generating systematic underinvestment. Drawing on risk regimes in nuclear safety, pandemic preparedness, and critical infrastructure, we propose that similar mechanisms (precommitment, shared protocols, standing coordination venues) could be adapted to frontier AI governance. Closing the gap requires cross-actor "note-exchange" of ex ante if-then response logic, exposing not only triggers but the decision processes that convert signals into actions. Without such architecture, institutions cannot learn from failures at the pace of relevance.

1 Introduction

The central challenge of AI governance is to harness the economic and social benefits of widespread deployment while preventing the potential harms that such systems may generate. Yet the societal impacts of advanced AI systems remain difficult to anticipate [5], creating fundamental tensions in policy design. Regulators seek to hold developers accountable without driving innovation offshore, and to impose sufficient safety requirements without stifling the very technological progress they aim to govern. This is a delicate balancing act, and most research has naturally focused on optimizing preventative measures for development and deployment [70].

Given the scale of potential harms, societies are likely underinvesting in AI safety, and additional preventive effort plausibly yields positive marginal returns [35]. Yet even prevention with extremely high reliability, e.g. on the order of 99%, is unlikely to guarantee full containment in complex socio-technical systems [57]. Recognizing this, many researchers anticipate that harmful incidents or near-misses ("warning shots" or "focusing events") will be necessary to shift political feasibility toward more stringent and costly interventions [11]. In effect, this requires taking seriously the possibility of failures that generate large or even catastrophic societal harm. As AI capabilities and deployment accelerate, such failures become more likely.

Importantly, this does not weaken the case for prevention: preventive investment continues to dominate expected-value calculations for existential risk. The question is instead whether accounting for inevitable failures changes where additional effort has the highest marginal payoff. We argue that it does. When institutions lack capacity for shared learning, coordinated response, and adaptive adjustment, governance is forced to rely almost exclusively and increasingly on preventive control, even when prevention cannot eliminate residual risk. Under these conditions, the absence of resilience is what makes prevention the only available, yet ultimately insufficient, strategy.

Article presented at the Second IASEAI Conference (IASEAI'26).

Broader frameworks for societal adaptability have been proposed [10], yet they leave open concrete implementation pathways. Meanwhile, a large body of work has identified specific harms: cognitive overload at the societal level [39] and shifts in the offense-defense balance in cybersecurity [29], among others. More broadly, the literature identifies a "context gap" between technical prevention protocols and holistic societal impact assessments [70]. What matters for frontier-AI governance is whether systems and institutions can maintain acceptable performance under unexpected conditions and continue functioning when shocks occur. Given the costs of such shocks, the goal is to minimize their occurrence, maximize learning from them, and where possible learn even in their absence. Current policies address only the first [14].

Many proposals contribute frameworks for monitoring and reflection around such events [28, 64, 69]. Monitoring is indeed a necessary condition for effective mitigation. Yet the underlying challenge is a collective action problem: benefits are diffuse while costs are concentrated, creating incentives to free ride and a high risk of coordination failure. These dilemmas are well studied, and established strategies exist to address them; one such mechanism is precommitment [55].

The relevant actors span government regulators, critical infrastructure operators, private labs and cloud providers, and civil society. No single actor possesses authority or control across all domains, yet optimal responses require coordination among them. During high-pressure crisis events, this coordination capacity becomes critical, a point recognized by calls for interagency task forces, public-private partnerships, and treaty-like mechanisms. From other high-risk fields, we can extract best practices for such coordination efforts [67].

While multiple actors must account for robustness to improve coordination, labs as primary developers and operators carry particular responsibility. Labs are already producing Frontier AI Safety Policies (FASPs), yet these policies implicitly assume that preventative measures will succeed. Taking seriously the possibility that they will not—and planning accordingly—requires integrating robustness standards into these policies, making them the natural anchor point for multi-actor coordination.

2 The 'robustness gap' in the broader Ecosystem

Robustness in frontier-AI governance concerns whether policies and institutions continue to function when underlying assumptions fail, signals are ambiguous, or external conditions shift in ways developers did not anticipate [46]. The problem is that current approaches operate under the assumption of control and fail to take failure seriously.

2.1 Consequences of ignoring robustness

Traditional "predict-then-act" approaches assume accurate forecasts. Under conditions of deep uncertainty, this produces brittle policies that lead to one of two failure modes: gridlock, where actors cannot agree on underlying assumptions, or overconfidence, where plans fail when futures diverge from expectations [61, 40]. This brittleness is particularly consequential when actors operate under differing assumptions about baseline risk and responsibility [25, 24]. While actors may not be able to agree on underlying probability distributions—and therefore on adequate degrees of preventative measures—they can nevertheless disclose their post-hoc responses in light of impending harm or novel evidence of potential harm [18, 31].

The myopic focus on monitoring compliance of individual developers neglects socio-technical feedback loops. System-level harms such as cyber-risks or disinformation cascades cannot be mitigated by individual actors alone [49]. When policies meet internal compliance standards yet still fail under real-world shocks, trust in governance architectures declines. Risk management in low-trust environments poses major challenges, as documented in other high-risk contexts [51, 34, 50].

As Figure 1 illustrates, existing AI governance efforts are heavily concentrated on prevention, with substantial work on alignment and testing [4], usage constraints [12], compliance auditing [13], and regulatory norms [23, 71]. By contrast, mitigation and preparedness remain far less developed. Although monitoring and incident reporting have begun to receive attention [52], policy proposals addressing coordination and crisis management are sparse, leaving coordination capacity the most underexplored element of the governance landscape.

Robust Decision Making (RDM) offers an alternative framework. Rather than relying on contested forecasts, it stress-tests candidate strategies across hundreds of plausible futures to reveal vulnerabilities and adaptive levers. This shifts attention from prediction to preparation, and from "what will happen?" to "what actions hold across many futures?" and "what tradeoffs are we willing to make?" [61, 40]. Neglecting such robustness thinking in frontier AI governance means that Responsible Scaling Policies (RSPs) remain strong at local gating—controlling deployment decisions within a single organization—but weak at global adaptability. Without cross-domain triggers, external signposts, and pre-committed adaptation pathways, policies risk either paralysis or ad hoc responses once crises unfold.

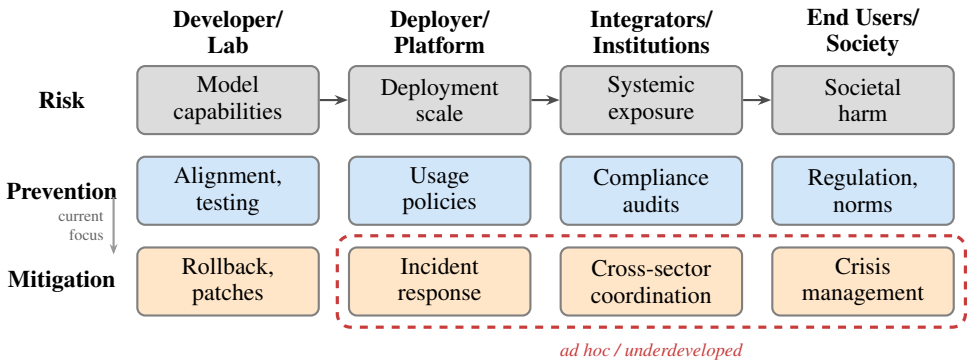

Figure 1: **The robustness gap in frontier AI governance.** Risk of harm accumulates downstream (top row). Opportunities for prevention concentrate upstream. Broader mitigation capacity remains underdeveloped. However, this determines society's capacity to adopt and learn.

2.2 Tentative steps towards robust AI governance

Current FASPs make effective use of adaptive triggers. Anthropic's AI Safety Level (ASL) upgrades, OpenAI's High/Critical risk gates, and DeepMind's Critical Capability Level (CCL) alerts provide clear internal triggers for escalating safety requirements [53, 6, 47]. Each lab has implemented some form of scalable or early-warning evaluation, along with post-deployment monitoring and red-team feedback loops. The policies contain firm pre-commitments to halt or withhold deployment absent required safeguards, with named decision bodies and defined documentation requirements such as capability reports and safety case assessments.

However, these mechanisms remain internally focused and siloed by hazard category. Consider the following cross-domain scenarios: cybersecurity disruptions that cascade into compromised biosafety protocols in research laboratories; AI agents deployed at scale to commit or assist in criminal activity; a model that successfully exfiltrates itself but fails to recursively self-improve, leaving the system in an ambiguous risk state; or significant psychological harms emerging in vulnerable subpopulations from sustained interaction with frontier models. Current frameworks lack explicit triggers and response protocols for such scenarios.

Recent proposals by OpenAI and Anthropic have begun to incorporate transparency and coordination as important criteria [54, 7]. The primary motivation appears to be ecosystem security, particularly regarding cybersecurity vulnerabilities. Anthropic proposes transparency requirements for frontier developers' safety policies and notes that required safeguards can be lowered if another actor passes a capability threshold without them. While this represents precisely the kind of robust, conditional, and transparent mechanism that is argued as necessary, it constitutes a relaxation of commitments rather than improving mitigation.

These developments acknowledge both the need for transparency and adaptive triggers, yet fall short of enabling coordination on optimal responses. In a crisis scenario, critical conversations will either be held for the first time or conducted without knowledge of other relevant actors' reasoning processes—precisely when coordination capacity matters most.

3 Introducing robustness thinking to FASP and other stakeholders

Integrating robustness into frontier AI governance requires three foundational shifts. First, reactions, tradeoffs, and reasoning in response to triggers must be made transparent—not merely the triggers themselves, but the decision logic that converts signals into actions. Second, triggers must be partially externalized, moving beyond purely technical metrics to incorporate consequentialist and external measures of societal impact. Third, this integration should build on existing institutional capacity: some government bodies already deploy robustness metrics in adjacent domains [43], providing templates for adaptation.

A growing body of work makes clear that evaluation alone is insufficient for managing high-stakes risk. Lukošiūtė and Swanda [42] show that even sophisticated capability evaluations fail to capture real-world risk dynamics, particularly under adversarial or rapidly shifting operational conditions. This limitation is not merely technical: it reflects a broader structural challenge in which uncertainty about failure modes and their thresholds undermines coordinated precaution. Experimental evidence from climate threshold-public-goods games illustrates this dynamic starkly. When the location or consequences of a catastrophic threshold are uncertain, cooperative investment collapses–even among actors who prefer the cooperative outcome and understand the stakes [9]. The mechanism is simple: uncertainty about others' responses lowers each actor's expectation that coordinated action will materialize, producing individually rational but systemically disastrous underinvestment. Frontier-AI safety faces the same structural vulnerability. As long as actors hold divergent beliefs about thresholds of unacceptable risk and lack visibility into others' intended responses, even well-designed preventive policies exhibit a robustness gap.

3.1 Learning from other domains

Risk management in other disciplines offers valuable lessons, though rarely at the scope and complexity of transformative AI [67]. The most relevant precedents come from nuclear safety, pandemic preparedness, and cybersecurity—domains that share characteristics of catastrophic tail risk, socio-technical complexity, and the need for coordinated response across organizational boundaries. These domains have converged on layered safeguards, rapid response capacity, early warning systems, pre-decided decision frameworks [27], and pre-positioned resources. Interdependence between critical infrastructure and their failure modes is well documented [59, 66, 3], and similar thinking will be necessary for AI crisis preparedness. Each domain has also developed a specific coordination device: Sector Risk Management Agencies (SRMAs) designate sector leads that convene owners and operators through standing information-sharing channels [21, 72]; the International Health Regulations (IHR) impose mandatory, time-bounded signal escalation that creates common knowledge under uncertainty [30]; and the IAEA provides standardized incident classification vocabularies enabling rapid cross-actor interpretation without renegotiating meanings mid-crisis [33].

Across these regimes, the shared institutional function is not "perfect prevention" but rather a set of coordination primitives: precommitment to notify and coordinate, interoperable procedures, and standing venues for joint situational awareness and response alignment [58]. In each case, coordination capacity is treated as a first-class governance objective—responsibilities for convening, escalation, and cross-actor alignment are assigned ex ante rather than improvised during crises [16]. What appears optimal from an individual actor's perspective may produce dramatically worse outcomes at the system level; individual organizations optimizing their own response plans in isolation may duplicate efforts inefficiently, create incompatible protocols, or leave critical gaps uncovered [27].

Advanced AI governance faces structural challenges that make these coordination problems more acute. First, the broad applicability of general-purpose technologies—particularly agentic systems—generates fuzzy domain boundaries, ever-diversifying taxonomies, and arbitrary thresholds that frustrate the standardization these regimes rely upon. Unlike nuclear incidents or disease outbreaks, AI-related harms often lack clear categorical boundaries: cybersecurity disruptions may cascade into compromised biosafety protocols; psychological harms may emerge gradually in vulnerable subpopulations; models may partially exfiltrate without achieving recursive self-improvement, leaving systems in ambiguous risk states. Second, increasing potential severity combined with shorter response times makes institutional learning-by-doing—the mechanism through which previous

regimes developed their coordination capacity—undesirable or infeasible. Third, unlike established regimes that built trust and interoperability over decades of practice, AI governance must develop coordination capacity ahead of time, before the focusing events that typically catalyze institutional development. This means coordination mechanisms cannot wait for consensus on threat vectors or accumulated incident data; they must function under conditions of deep uncertainty about failure modes themselves [65].

Transparency, rather than legal commitment, may be sufficient to substantially increase systemic coordination capacity. However, many interventions are costly, and private entities are understandably cautious about committing capital to contingent responses. This makes it valuable to first investigate the option space of possible interventions through structured scenario-response explorations, which can reveal synergies and complementarities in coordination capacity, or identify responses by one actor that are necessary to enable effective responses by others.

3.2 A tentative policy proposal

We propose a Scenario Response Registry (SRR) as an institutional mechanism to operationalize robustness thinking.[1] The core function is to make coordination capacity visible, comparable, and stress-testable before crises materialize. By requiring actors to articulate their intended responses to specified scenarios, the SRR enables red-teaming of governance frameworks, evaluation of policy under technical constraints, modelling of adversarial and cascading risk scenarios, and drafting of governance proposals anchored in real model behaviour. Where current approaches leave coordination to be improvised under crisis conditions, the SRR shifts learning and alignment upstream.

A public authority would maintain the registry, curating a library of salient technical, socio-economic, and security scenarios with specified potential harms. All relevant actors—frontier labs, cloud providers, critical infrastructure operators, and government agencies—would file ex ante "if-then" plans detailing thresholds that would trigger action, the specific actions to be taken, and resource commitments allocated to those responses. Submissions would be standardized, time-stamped, and comparable, creating common knowledge and curbing cheap talk—the costless announcement of commitments unlikely to be honored. The registry would review filings to identify gaps, flag overlaps, and suggest improvements. Access structures may be asymmetric: the scenario library could be broadly accessible while specific response filings and harmonization analyses remain restricted to relevant actors and the coordinating authority.

The SRR's incentive architecture would link disclosure quality to tangible consequences. Quality-weighted plans could determine regulatory forbearance, access to public compute resources, or procurement eligibility. Skin-in-the-game mechanisms—such as bonds or insurance instruments—would be forfeited upon failure to act when declared triggers are met. Dynamic scoring would update robustness assessments that shape audit intensity and capital requirements, creating ongoing incentives for plan quality and execution capacity. An independent technical panel with rotating membership would set the scenario library and auditing rubric, limiting regulatory capture while maintaining technical credibility.

[1] The SRR builds on the concept of risk registers, established in industrial risk management, which catalog known hazards alongside likelihood assessments, impact ratings, and designated responses [8]. The key difference is that the SRR requires prospective filing based on foresight about failure modes that have not yet materialized.

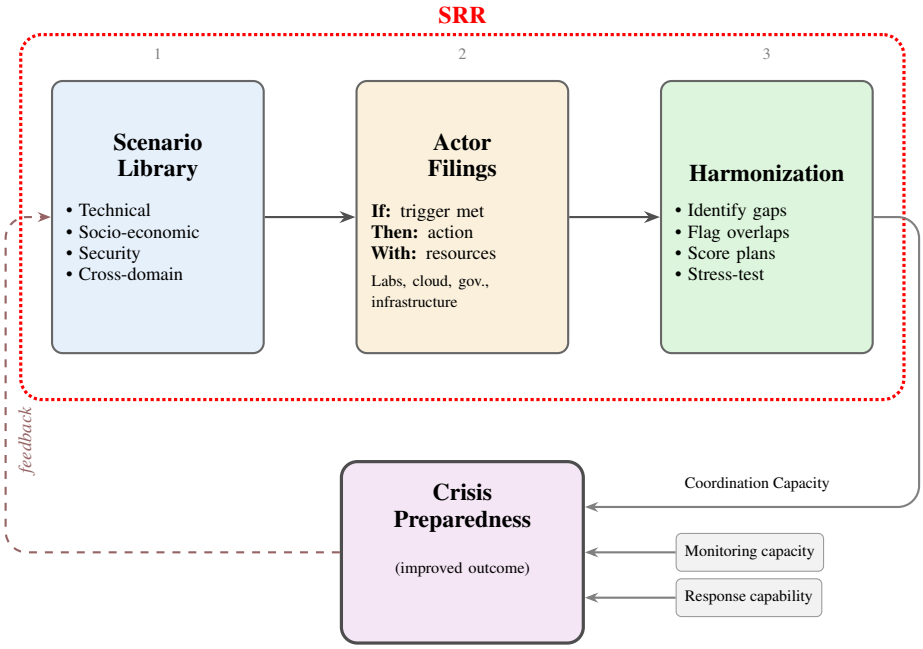

Figure 2: **The Scenario Response Registry (SRR) as a coordination mechanism.** Scenarios are curated by an independent panel (1). Relevant actors file standardized if–then response plans (2). The SRR harmonizes filings to identify gaps and coordination opportunities (3). Together with complementary efforts—incident monitoring and preparedness frameworks—the SRR contributes to improved crisis readiness. Feedback from drills and real incidents updates the scenario library.

4 Challenges and conditions for integration

The SRR does not emerge in a vacuum. Complementary efforts in the broader ecosystem are addressing other dimensions of the preparedness deficit. Various incident databases document realized harms or near-misses, providing empirical grounding for scenario identification [45, 48]. Recent work develops operational frameworks for emergency response and loss of control [65, 68]. These contributions advance incident learning and response planning—but neither addresses how actors coordinate credibly to each other's responses before crises materialize. The SRR targets this coordination gap directly, complementing calls for greater systematicity in frontier AI risk management [73].

4.1 Why robustness is undersupplied

The underinvestment in robustness thinking in frontier AI governance stems from a fundamental misalignment between private and social returns [26]. Private returns favor speed and scope—first-mover advantages, market share, and strategic positioning—while social returns hinge on resilience and the capacity to absorb shocks. Firms therefore systematically underinvest in robustness [5]. Strategic ambiguity compounds this problem by preserving bargaining power in negotiations with regulators and competitors. Codified triggers that expose tradeoffs and curtail managerial discretion are thus resisted, as they reduce flexibility that may prove valuable under competitive pressure [63].

Translating abstract commitments into operational thresholds, triggers, and rules requires substantial upfront investment in scenario mapping and the identification of credible signposts [44]. Miscalibration carries significant costs: false positives erode credibility and generate pressure to relax standards, while false negatives allow harm to accumulate undetected [62, 22]. This risk of calibration failure reinforces organizational preferences for implicit pathways that preserve interpretive flexibility.

Maintaining live indicators demands sustained resources, disciplined data collection, and specialized analytic capacity. Without continuous investment, these systems decay into periodic check-ins that

provide neither early warning nor actionable intelligence. This degradation further reduces the perceived payoff to robustness investments, creating a negative feedback loop.

Robustness exhibits the classic structure of a public good: benefits are diffuse across many stakeholders, while the costs are concentrated on those who must slow development, narrow deployment scope, or invest in standby capacity [32]. This predictably generates free-riding and holdout problems, and measures that impose real constraints face resistance unless governance rules are perceived as fair, technically justified, and applied consistently across competitors. In a first-best setting, these incentive failures require strong governance institutions capable of verifying commitments, enforcing reciprocity, and resisting capture by either regulated entities or bureaucratic interests [1]. In their absence, however, coordination mechanisms should not be evaluated by whether they offer guarantees, but by whether they measurably improve coordination relative to the status quo. Absent such institutions, coordination-enhancing arrangements that increase transparency, align expectations, and support preparedness may not fully resolve the commons problem, but they can still dominate uncoordinated responses and reduce the risk that robustness failures propagate unchecked.

4.2 Towards improving coordination

Several structural constraints limit the feasibility of global pre-coordination in frontier AI governance. Geopolitical dynamics and strategic rivalry reduce willingness to share assessments, thresholds, or vulnerabilities across jurisdictions [38]. Competitive pressures create incentives for both states and firms to conceal strategically relevant capabilities, undermining the credibility of ex ante commitments [41]. Because sensitive technical information is inherently dual-use, any viable coordination regime must incorporate tiered access, controlled-sharing infrastructures, or delayed-release protocols [19, 20].

These constraints suggest that global coordination—requiring convergence on shared threat models, uniform standards, and binding enforcement—is unrealistic under current conditions of strategic mistrust, asymmetric information, and divergent regulatory priorities. However, polycentric governance offers a viable alternative [56, 36]. Rather than awaiting comprehensive international agreement, coordination can begin locally through domestic SRRs, bilateral agreements such as US-UK arrangements or lab consortia, or sector-specific pilots.

A common objection holds that local or partial approaches risk irrelevance without broader participation. Yet durable coordination regimes often begin locally and expand through overlapping coalitions as trust accumulates [37]. The key design principle is interoperability over uniformity: compatible procedures that can later integrate as mutual confidence builds [15]. Interoperability represents a lower bar than full harmonization and is therefore more robust under political uncertainty [2]. Concrete bootstrapping pathways—whether anchored in lab consortia, cloud infrastructure providers, or bilateral regulatory arrangements—warrant dedicated analysis beyond the scope of this paper, though the interoperability principles would apply across configurations.

Any viable policy solution must function under realistic failure modes: partial participation, strategic under-specification, geopolitical fragmentation, and crisis-time friction. Critically, even partial coordination delivers value by reducing uncertainty about others' responses among willing actors [17]. Tiered access regimes and controlled-sharing infrastructures can reconcile transparency with security constraints; such mechanisms are already deployed for hardware access and can apply similarly to information crucial for coordination [60].

A final objection notes that without enforcement, no mechanism can guarantee compliance. This is true but misses the point: imperfect coordination mechanisms implemented sooner dominate perfect mechanisms implemented later, particularly when institutional learning is required. The experience of the International Health Regulations illustrates this dynamic—even with established frameworks, implementation capacity lags behind formal commitments, arguing for starting now to build institutional muscle [30]. For a minimum viable SRR to be impactful requires only a curated scenario set, willing filers, a reviewing body, and regular tabletop exercises. These modest requirements lower the barrier to entry while establishing the infrastructure for more robust coordination as the ecosystem matures.

The SRR itself is subject to failure modes that could undermine its coordination function. Scenario libraries may be misspecified—either too narrow to capture emergent cross-domain risks or too expansive to enable meaningful preparation. Harmonization processes could generate false confidence

if actors treat filed plans as sufficient preparation rather than inputs to ongoing learning. Performative compliance remains a risk: organizations may file detailed responses they lack capacity or intent to execute. Finally, the scenario-setting process is vulnerable to capture, whether by regulated entities seeking to narrow the scope of plausible threats or by bureaucratic interests seeking to expand jurisdiction. These failure modes argue for iterative design with built-in review mechanisms rather than a fixed institutional architecture.

5 Conclusion

Frontier AI Safety Policies are built on the assumption that prevention will hold. It mostly will. But these policies also implicitly assume that when prevention fails, the broader ecosystem will know what to do, and at present it does not. There is no shared vocabulary for cross-domain failures, no agreed escalation path, and no common knowledge of how major actors would behave once their internal triggers fire. This is the coordination gap. Closing it does not require global agreement, which is unlikely to arrive in time. It requires that the actors who already make these decisions, labs, infrastructure operators, and regulators, expose their if-then reasoning to one another ex ante. Such note-exchange is the cheapest form of coordination capacity available, and the form most likely to scale with the pace of capability progress, deployment, and incidents. Polycentric beginnings, domestic, bilateral, and sector-specific, are not a compromise on global ambitions; they are how durable regimes typically start. The case for this work is not that it replaces prevention. It is that the institutions capable of learning at the speed of frontier AI do not yet exist, and waiting for the focusing events that would otherwise produce them is not a coordination strategy that anyone should choose. Building those institutions now, while the cost of doing so is a cost that can be chosen rather than imposed, is the agenda this paper opens.

Acknowledgments and Disclosure of Funding

The author is especially grateful to Jan Pieter Snoeij (Simon Institute for Longterm Governance) for sustained and substantive discussions that shaped both the framing and the central arguments of this paper. The author also thanks the Centre for the Governance of AI and co-fellows for valuable input and discussions. The author declares no competing interests.

References

[1] Kenneth W Abbott. *International organizations as orchestrators*. Cambridge University Press, 2015.

[2] Kenneth W Abbott and Duncan Snidal. The governance triangle: Regulatory standards institutions and the shadow of the state. In *The spectrum of international institutions*, pages 52–91. Routledge, 2021.

[3] Cristina Alcaraz and Sherali Zeadally. Critical infrastructure protection: Requirements and challenges for the 21st century. *International journal of critical infrastructure protection*, 8: 53–66, 2015.

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[5] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

[6] Anthropic. Anthropic's responsible scaling policy. `https://www.anthropic.com/news/anthropics-responsible-scaling-policy`, September 2023. Accessed: 2025-08-11.

[7] Anthropic. Proposed frontier model transparency framework. Technical report, Anthropic, 2025. URL `https://www-cdn.anthropic.com/19cc4bf9eb6a94f9762ac67368f3322cf82b09fe.pdf`.

[8] Terje Aven. Risk assessment and risk management: Review of recent advances on their foundation. *European journal of operational research*, 253(1):1–13, 2016.

[9] Scott Barrett and Astrid Dannenberg. Sensitivity of collective action to uncertainty about climate tipping points. *Nature Climate Change*, 4(1):36–39, 2014.

[10] Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, and Markus Anderljung. Societal adaptation to advanced ai. *arXiv preprint arXiv:2405.10295*, 2024.

[11] Thomas A Birkland. *After disaster: Agenda setting, public policy, and focusing events*. Georgetown University Press, 1997.

[12] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[13] Miles Brundage, Noemi Dreksler, Aidan Homewood, Sean McGregor, Patricia Paskov, Conrad Stosz, Girish Sastry, A Feder Cooper, George Balston, Steven Adler, et al. Frontier ai auditing: Toward rigorous third-party assessment of safety and security practices at leading ai companies. *arXiv preprint arXiv:2601.11699*, 2026.

[14] Giliberto Capano and Jun Jie Woo. Resilience and robustness in policy design: A critical appraisal. *Policy Sciences*, 50(3):399–426, 2017.

[15] Yik Chan Chin, David A Raho, Hag-Min Kim, Chunli Bi, James Ong, Jingbo Huang, and Serge Stinckwich. Interoperability in ai safety governance: Ethics, regulations, and standards. *arXiv preprint arXiv:2601.06153*, 2026.

[16] Tom Christensen, OLE Andreas Danielsen, Per Lægreid, and LISE H. RYKKJA. Comparing coordination structures for crisis management in six countries. *Public Administration*, 94(2): 316–332, 2016.

[17] Russell Cooper, Douglas V DeJong, Robert Forsythe, and Thomas W Ross. Communication in coordination games. *The Quarterly Journal of Economics*, 107(2):739–771, 1992.

[18] National Research Council. *Informing Decisions in a Changing Climate*. The National Academies Press, 2009. ISBN 978-0-309-13799-0. doi: 10.17226/12626. URL https://nap.nationalacademies.org/catalog/12626/informing-decisions-in-a-changing-climate.

[19] National Research Council, Global Affairs, Security, Cooperation, Committee on Research Standards, and Practices to Prevent the Destructive Application of Biotechnology. Biotechnology research in an age of terrorism. 2004.

[20] National Research Council, Global Affairs, Committee on Science, Law, Committee on a New Government-University Partnership for Science, and Security. Science and security in a post 9/11 world: A report based on regional discussions between the science and security communities. 2007.

[21] Cybersecurity and Infrastructure Security Agency. Sector risk management agencies, n.d. URL https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors/sector-risk-management-agencies. Accessed 2026-01-25.

[22] European Commission. Understanding and applying the precautionary principle: A policy framework. Technical report, European Commission, 2023. URL https://reforms-investments.ec.europa.eu/system/files/2023-10/5b14362c-en.pdf. Accessed 2026-02-10.

[23] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 — Artificial Intelligence Act. Official Journal of the European Union, L 2024/1689, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[24] Severin Field. Why do experts disagree on existential risk and p (doom)? a survey of ai experts. *arXiv preprint arXiv:2502.14870*, 2025.

[25] FLI. Ai safety index: Summer 2025. https://futureoflife.org/ai-safety-index-summer-2025/, 2025. Accessed: May 4, 2026.

[26] Benjamin Gil Friedman. Shared residual liability for frontier ai firms. *Available at SSRN 5339887*, 2025.

[27] Prasanna Gai and Sujit Kapadia. Contagion in financial networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2120):2401–2423, 2010.

[28] Milan Gandhi, Peter Cihon, Owen Larter, and Rebecca Anselmetti. Societal capacity assessment framework: Measuring resilience to inform advanced ai risk management. *arXiv preprint arXiv:2509.22742*, 2025.

[29] Ben Garfinkel and Allan Dafoe. How does the offense-defense balance scale? In *Emerging Technologies and International Stability*, pages 247–274. Routledge, 2021.

[30] Lawrence O Gostin and Rebecca Katz. The international health regulations: the governing framework for global health security. *The Milbank Quarterly*, 94(2):264–313, 2016.

[31] Marjolijn Haasnoot, Jan H. Kwakkel, Warren E. Walker, and Judith ter Maat. Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change*, 23(2):485–498, 2013. doi: 10.1016/j.gloenvcha.2012.12.006. URL https://doi.org/10.1016/j.gloenvcha.2012.12.006.

[32] Russell Hardin. *Collective action*. Rff Press, 2015.

[33] International Atomic Energy Agency and Nuclear Energy Agency. The international nuclear event scale (ines) user's manual. Feb 2001.

[34] IRGC. Introduction to the irgc risk governance framework, revised version, 2017. URL https://irgc.org/wp-content/uploads/2018/09/IRGC.-2017.-An-introduction-to-the-IRGC-Risk-Governance-Framework.-Revised-version..pdf. Accessed: May 4, 2026.

[35] Charles I Jones. How much should we spend to reduce ai's existential risk? Technical report, National Bureau of Economic Research, 2025.

[36] Emily Jones. *The political economy of bank regulation in developing countries: risk and reputation*. Oxford University Press, 2020.

[37] Robert O Keohane and David G Victor. The regime complex for climate change. *Perspectives on politics*, 9(1):7–23, 2011.

[38] Andrew H Kydd. *Trust and mistrust in international relations*. Princeton University Press, 2005.

[39] Salem Lahlou. Mitigating societal cognitive overload in the age of ai: Challenges and directions. *arXiv preprint arXiv:2504.19990*, 2025.

[40] Robert J Lempert, Benjamin P Bryant, Myles T Collins, Andrew Hackbarth, Tom LaTourrette, Robert T Reville, Steven W Popper, Christophe Mijere, David G Groves, Klaus Keller, et al. Making good decisions without predictions: Robust decision making for planning under deep uncertainty. 2013.

[41] Keir A Lieber and Daryl G Press. The new era of counterforce: Technological change and the future of nuclear deterrence. *International Security*, 41(4):9–49, 2017.

[42] Kamilė Lukošiūtė and Adam Swanda. Llm cyber evaluations don't capture real-world risk. *arXiv preprint arXiv:2502.00072*, 2025.

[43] Jeremy Mangin, Michael Bluck, Geoff Darch, Diarmid Roberts, Solomon Brown, and Mark Workman. A robust decision-making approach to evaluating state support for nuclear programmes under uncertainty: A uk case study. *Energy Policy*, 205:114653, 2025.

[44] Vincent AWJ Marchau, Warren E Walker, Pieter JTM Bloemen, and Steven W Popper. *Decision making under deep uncertainty: from theory to practice*. Springer Nature, 2019.

[45] Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15458–15463, 2021.

[46] Marjolein JP Mens, Frans Klijn, Karin M De Bruijn, and Eelco Van Beek. The meaning of system robustness for flood risk management. *Environmental science & policy*, 14(8): 1121–1131, 2011.

[47] Meta. Our approach to frontier ai, February 2025. URL `https://about.fb.com/news/2025/02/meta-approach-frontier-ai/`.

[48] MITRE Corporation. MITRE ATT&CK. `https://attack.mitre.org/`, 2025. Accessed: 2026-01-12.

[49] José Moran, Frank P Pijpers, Utz Weitzel, Jean-Philippe Bouchaud, and Debabrata Panja. Critical fragility in socio-technical systems. *arXiv preprint arXiv:2307.03546*, 2023.

[50] NAIIC. *The Official Report of the Fukushima Nuclear Accident Independent Investigation Commission*. The National Diet of Japan, 2012. URL `https://www.nirs.org/wp-content/uploads/fukushima/naiic_report.pdf`. Accessed: May 4, 2026.

[51] OECD. *Trust and Public Policy: How Better Governance Can Help Rebuild Public Trust*. OECD Publishing, Paris, 2017. ISBN 9789264268913. doi: 10.1787/9789264268920-en. URL `https://doi.org/10.1787/9789264268920-en`.

[52] OECD. Towards a common reporting framework for AI incidents. OECD Artificial Intelligence Papers 34, OECD Publishing, Paris, 2025. URL `https://doi.org/10.1787/f326d4ac-en`.

[53] OpenAI. Preparedness framework (v2), April 2025. URL `https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf`.

[54] OpenAI. Scaling coordinated vulnerability disclosure, June 2025. URL `https://openai.com/index/scaling-coordinated-vulnerability-disclosure/`.

[55] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

[56] Elinor Ostrom. Beyond markets and states: polycentric governance of complex economic systems. *American economic review*, 100(3):641–672, 2010.

[57] Charles Perrow. Normal accidents: Living with high risk technologies-updated edition. 2011.

[58] Ronald W Perry and Michael K Lindell. Preparedness for emergency response: guidelines for the emergency planning process. *Disasters*, 27(4):336–350, 2003.

[59] Shannon Prier, Aaron M Strong, and Jonathan W Welburn. Interdependence across the national critical functions. 2023.

[60] RAND Corporation. Understanding the artificial intelligence diffusion framework: Can export controls create a u.s.-led global ai ecosystem? Perspective on a Timely Policy Issue PEA3776-1, RAND Corporation, 2025. URL `https://www.rand.org/content/dam/rand/pubs/perspectives/PEA3700/PEA3776-1/RAND_PEA3776-1.pdf`.

[61] RAND Corporation. Robust decision making. RAND Tool Page, n.d. URL `https://www.rand.org/pubs/tools/TL320/tool/robust-decision-making.html`. Accessed 14 August 2025.

[62] Yohei Sawada, Rin Kanai, and Hitomu Kotani. Impact of cry wolf effects on social preparedness and the efficiency of flood early warning systems. *Hydrology and Earth System Sciences*, 26 (16):4265–4278, 2022.

[63] Louisa Selivanovskikh, Pier Luigi Giardino, Matteo Cristofaro, Yongjian Bao, Wenlong Yuan, and Luming Wang. Strategic ambiguity: a systematic review, a typology and a dynamic capability view. *Management Decision*, 63(13):123–145, 2025.

[64] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*, 2023.

[65] Elika Somani, Anjay Friedman, Henry Wu, Marianne Lu, Christopher Byrd, Henri van Soest, and Sana Zakaria. Strengthening emergency preparedness and response for ai loss of control incidents. Research Report RR-A3847-1, RAND Corporation, 2025. URL `https://www.rand.org/pubs/research_reports/RRA3847-1.html`. Accessed: 2026-01-12.

[66] Tove Rydén Sonesson, Jonas Johansson, and Alexander Cedergren. Governance and interdependencies of critical infrastructures: Exploring mechanisms for cross-sector resilience. *Safety science*, 142:105383, 2021.

[67] Nicholas Stetler. Reinsuring ai: Energy, agriculture, finance & medicine as precedents for scalable governance of frontier artificial intelligence. *arXiv preprint arXiv:2504.02127*, 2025.

[68] Charlotte Stix, Annika Hallensleben, Alejandro Ortega, and Matteo Pistillo. The loss of control playbook: Degrees, dynamics, and preparedness. *arXiv preprint arXiv:2511.15846*, 2025.

[69] UKAISI. The uk ai security institute's research agenda. `https://www.aisi.gov.uk/research-agenda`, 2025. Accessed: 2025-10-10.

[70] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.

[71] Scott Wiener. Senate bill 53: Transparency in frontier artificial intelligence act. Chapter 138, Statutes of 2025, California Legislature, 2025–2026 Regular Session, 2025. URL `https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53`. Approved by Governor September 29, 2025.

[72] Jose M Yusta, Gabriel J Correa, and Roberto Lacal-Arántegui. Methodologies and applications for critical infrastructure protection: State-of-the-art. *Energy policy*, 39(10):6100–6119, 2011.

[73] Marta Ziosi, James Gealy, Miro Plueckebaum, Daniel Kossack, Simeon Campos, Lama Saouma, Uzma Chaudhry, Lisa Soder, Merlin Stein, Nicholas Caputo, et al. Safety frameworks and standards: A comparative analysis to advance risk management of frontier ai. 2025.