Adaptive Governance for Advanced AI: A Conceptual Foundation for Managing Complex Risks

A framework for governance that moves at the speed of AI, built on the science of complex systems

Kyle A. Kilian and Richard Mallah | May 20, 2026 | Center for AI Risk Management & Alignment

CARMA

Adaptive Governance for Advanced AI: A Conceptual Foundation for Managing Complex Risks

Kyle A. Kilian[1,2] and Richard Mallah[1,3]

[1]Center for AI Risk Management and Alignment, [2]RAND, [3]Future of Life Institute

1. Introduction

The rapid acceleration of artificial intelligence development, particularly in generative AI, dual-use foundation models, fixed AI agents, and dynamic agent swarms is outpacing traditional governance institutions. As AI's capabilities expand across varied industries, functions, and oversight visibility, and are networked through critical sectors, from finance to national defense, they create new risks, threat vectors, and vulnerability surfaces, and the stakes for managing those risks grow increasingly urgent.[1] [2] The potential for AI to bypass human oversight, driven by self-learning agent systems, increased autonomy, growing complexity, and quickening actions, poses profound societal-scale risks, including loss of control. The growing asymmetries between human decision-making and the complexity of AI systems in speed of throughput, decision-making, bandwidth, leverage, and actions will increasingly strain the capacity of public institutions and regulatory bodies to develop safeguards.

It is becoming nearly impossible for AI regulators or policymakers to keep pace with the exponential progress in AI development and deployment.[3] With each change in capability, degree of autonomy, or novel use case, new instantiations are rapidly deployed across industries, from social media to finance to safety-critical infrastructure, with little if any time to deliberate or respond. Furthermore, the true extent of AI integration and use remains difficult to ascertain or apply standard hierarchical controls for, with new AI agents deployed outside formal IT processes, often armed with enterprise-level permissions, the capacity to spawn sub-agents, and access to tools with which to interact with the real world.[4] This

rapid integration without clear mechanisms to monitor or adapt to change has the potential to destabilize industries, political structures, economies, and even national security. Governing such technologies requires more than static and reactive regulation: it demands an adaptive approach that can evolve alongside the technology, ensuring resilience and adaptivity in the face of transformative progress.

Yet the challenge is not merely that governance is too slow. A subtler and in some respects more dangerous failure mode has emerged alongside the obvious one: governance arrangements that *appear* adaptive but lack the structural properties necessary for substantive or genuine oversight. Voluntary pre-deployment evaluations conducted behind closed doors, emergency working groups convened without statutory authority, and industry-government partnerships that lack independence or enforcement capacity can occupy the political space that substantive governance would fill while providing none of its protective benefits. This performative adaptivity is in certain respects more dangerous than acknowledged rigidity, because it provides political cover for inaction while creating an illusion of oversight that discourages the development of robust institutional capacity.

This paper argues that understanding AI, in general, and its governance, in particular, as a complex adaptive system (CAS) is essential for crafting effective governance. It goes further, however, than merely proposing a lens through which to analyze the problem. It articulates a positive philosophy of governance under conditions of rapid technological change: a coherent vision of how societies should organize protection against evolving technological risk while maintaining democratic legitimacy, institutional resilience, and the capacity for continuous adaptation. This philosophy integrates the principles of complexity science (emergence, adaptability, redundancy, sensitivity to initial conditions) with insights from democratic theory, institutional design, and systems engineering to produce a comprehensive framework for thinking about, evaluating, and constructing AI governance.

While adaptive governance frameworks have been deployed in other contexts to mixed reviews, we argue that it is essential to revisit the approach and devise a modern framework to manage AI's super-exponential growth and limit catastrophic dangers. This is not a playbook or plan for devising specific policies. It is a framework for understanding what genuine adaptive governance requires, how to distinguish it from performative alternatives, and what structural properties any

institutional arrangement must possess to maintain protective capacity across political cycles, capability shocks, and changing threat landscapes.

Integrating an adaptive governance framework that embraces experimentation, efficient distributed decision-making, and continuous learning has the potential to allow policymakers to keep pace with the unpredictable risks and unprecedented changes in the development and deployment cycles of advanced AI systems. In this context, we outline five key principles that are essential for an effective adaptive governance framework: collectivity, adaptability, modularity, redundancy, and antifragility. These principles form part of the foundation of an adaptive governance philosophy, but they must be understood within a broader architecture that includes layered defensive capacity, structural legitimacy under uncertainty, and differential development of protective capabilities. Together, these components constitute a philosophy of governance flow: governance as a continuous dynamic process rather than a sequence of discrete regulatory acts, one that maintains structural integrity through adaptive response rather than rigid form.

2. The AI Ecosystem

To govern transformational change, we must understand AI and its larger sociotechnical context as a complex adaptive system (CAS). Indeed, AI systems are CAS, networked across higher levels of complexity, such as universities, companies, and countries, and within global communities and ecosystems. At its most basic, CAS are systems composed of large numbers of interacting components without central control that adapt through networks of interactions and feedback loops and display emergent "global" (i.e., collective) behavior.[5] Neural networks, social insect colonies, organizations (political, administrative, and corporate), economic markets, and societies are all CAS.

Key elements of CAS include self-organization and collective intelligence, adaptability, emergence, critical transitions or tipping points, and sensitivity to small variations (the observation by mathematician Edward Lorenz that led to the popular phrase "the butterfly effect").[6] This sensitivity means that even small interactions across systems at different scales can have unpredictable, outsized consequences. For example, researchers have observed sudden phase transitions in learning dynamics, where networks often experience abrupt shifts where performance rapidly improves (e.g., Grokking; see Power et al., 2022).[7] Large

language models demonstrate emergence of new capabilities at critical scale thresholds.

At the micro level, the interactions of AI agents can enhance learning and increase capabilities, driving adoption timelines and the degree of use across connected, even critical systems.[8] At the macro scale, AI adoption in media and education can drive rapid cultural shifts with transgenerational consequences; economically, AI's increased adoption across financial markets has contributed to cascading failures, such as the Flash Crash of 2010, where trading algorithms drove down the New York Stock Exchange by over a thousand points in minutes;[9] politically, AI-driven disinformation could influence voters, degrade trust in institutions, or more directly impact voting infrastructure; militarily, AI adoption could hasten the use of autonomous weapons, increasing the risks of automated warfare or shifting the balance of power.[10] The latest general-purpose systems are amplifying these feedback loops, in investment, users, and research, while gaining the capacity for recursive self-improvement.[11]

Crucially, AI systems differ from other complex adaptive systems in ways that compound the governance challenge. Unlike ecological or social systems that evolve on generational timescales, AI systems can undergo phase transitions in capability within weeks or months. Unlike economic systems where most actors have limited individual leverage, a single AI system may possess transformative capability concentrated in one deployment. And unlike biological systems where recursive self-improvement is constrained by physical reproduction cycles, AI systems face no such constraint: algorithmic improvements, inference-time scaling, and self-taught optimization allow capabilities to compound without the delays that give natural systems time to co-adapt.[12] These properties do not invalidate the CAS framework but they do establish that AI governance operates under more compressed timescales and with greater potential for single-point phase transitions than any domain where adaptive governance has previously been attempted.

Furthermore, modern AI systems exhibit a property with profound governance implications: post-deployment capability evolution without retraining. Inference-based scaling paradigms allow models with fixed parameters to exhibit substantially greater capability when allocated additional compute time for reasoning, self-verification, and iterative problem-solving.[13] A model that appears safe under standard evaluation conditions may behave as a vastly more powerful

system when given extended inference time on a specific problem, when integrated into multi-agent scaffolding, when combined with external tools, or when subjected to improved elicitation techniques discovered after release. This means that any governance framework relying solely on point-in-time pre-deployment evaluation will systematically underestimate deployed capability. Governance must account for the gap between measured capability under standard conditions and latent capability accessible through post-deployment optimization.

Looking at AI through this lens in governance requires embracing the default unpredictability and nonlinearity of systems and a willingness to pursue approaches that match these dynamics, such as iterative regulation, distributed decision-making, and continuous monitoring, to achieve sufficiently broad coverage of risk scenarios. Given the pace of AI progress and the convergence of other complex challenges, there is little choice but to pursue alternative governance mechanisms.[14] [15] The alternative, persisting with governance paradigms designed for stable, linear systems, amounts to a choice to remain unprotected.

3. The Challenge of Traditional Governance

Traditional governance processes are designed with stability as a baseline in mind and a focus on quality, permanence, and certainty.[16] This is crucial for less dynamic risks requiring broad consensus, like international treaties such as the prohibition of nuclear weapons.[17] Many proposed AI regulatory regimes follow this approach, including new international bodies, agreements, expert committees, executive agencies, and provisions for disclosure, registration, and risk management.[18] While many of these kinds of governance mechanisms are important and can help contain near and intermediate-term risks, there are critical gaps that must be addressed.

Compute-based registration systems, for example, designed to monitor large-scale compute clusters required for the most capable models, are dependent on rigid thresholds fixed to our latest understanding of state-of-the-art capabilities. However, there is a notable deficiency in this approach. In the event of hardware or algorithmic breakthroughs, regulators would need to replace or substantively adjust the rules to adapt to the new capability. The development of inference-based scaling paradigms, allowing models the time and structure to reason through user prompts, or breakthroughs in distributed training, allowing scalable compute across multiple smaller data centers, have increased the capabilities of smaller models while decreasing or potentially obviating the need for large compute clusters.[19] Each of

these methods allows developers to bypass simplistic compute restrictions while still producing potent, potentially uncontrollable AI systems.

These limitations should not be mistaken for an argument against compute governance, but a call to vigorously orchestrate it in a more robust assurance mechanism. While simplistic compute thresholds are brittle, more sophisticated compute governance mechanisms — encompassing not only training-compute registration but also runtime monitoring, hardware-level security features, and compute allocation tracking — substantially expand the option space available to governance bodies and increase the reliability of governance decisions across all four governance flow functions. Compute is the most measurable, physically grounded, and jurisdictionally tractable substrate of AI development. Governance mechanisms anchored in compute observability provide the sensing function with verifiable signals that do not depend on developer self-reporting, provide the evaluation function with concrete capability proxies that can be continuously recalibrated as algorithmic efficiency improves, and provide the response function with physically enforceable leverage points that cannot be circumvented through purely algorithmic means.

The integration of compute governance into an adaptive framework, with thresholds that update dynamically as algorithmic efficiency improves, with monitoring that extends beyond training to encompass inference-time scaling, and with internationally coordinated hardware tracking that addresses jurisdictional arbitrage, offers a more reliable foundation for governance action than any purely behavioral or capability-assessment-based approach alone. Compute governance does not solve the governance problem, but it materially increases the probability that governance decisions are grounded in verifiable physical reality rather than depending entirely on the assessments and disclosures of the entities being governed.[20]

But the inadequacy of traditional governance mechanisms constitutes only half the problem. The other half, less frequently discussed but equally dangerous, is the emergence of governance arrangements that mimic adaptivity while lacking its substance. Consider the characteristics of what we term *performative adaptivity*[21]:

Voluntary agreements between government entities and the companies they nominally oversee create the appearance of oversight without establishing authority to compel compliance or impose consequences for failure. Pre-deployment evaluations conducted without public reporting, contestable criteria, or enforcement

mechanisms produce information that may or may not inform decisions but provide no structural guarantee that dangerous systems will be restricted. Emergency coordination among agencies without statutory mandates creates rapid response capacity that exists only at the pleasure of current leadership and can be dissolved, redirected, or captured at any time. Industry-government working groups where the regulated entities are the primary interlocutors produce what might be more accurately described as negotiated tolerance than substantive governance. This is not an argument against public-private partnership as such: the technical complexity of AI and the concentration of relevant expertise and infrastructure in the private sector make such partnerships structurally necessary. It is an argument that partnerships must be designed with structural properties that prevent their degeneration into captured arrangements, and that the absence of those properties should be named honestly rather than papered over with the language of cooperation.

These arrangements share a common structural weakness: they depend entirely on the continued goodwill and competence of current leadership, the continued cooperation of regulated entities, and the absence of political conflicts that might redirect institutional attention. They are, in complex systems terms, governance structures with no robustness to perturbation. A change in political administration, a shift in competitive dynamics, a disagreement between government and industry on an unrelated matter: any of these can dissolve the entire governance structure overnight. This has already been demonstrated in practice: executive arrangements can be revoked within hours of inauguration, voluntary commitments can be abandoned without consequence, and political disputes can produce the immediate exclusion of safety-critical tools from government access[22].

The distinction between genuine adaptive governance and performative adaptivity is not merely academic. When performative arrangements occupy the governance space, they reduce political pressure for substantive institutional development. Policymakers can point to the existence of agreements, evaluations, and working groups as evidence that governance is functioning, even when these arrangements lack the structural properties necessary for genuine oversight. The result is that political capital that might otherwise be directed toward building durable institutions is consumed by maintaining the appearance of governance.

4. Governance as Flow: A Positive Philosophy

To contend with these challenges, we propose not merely a set of flexible programs or mechanisms but a comprehensive philosophy of governance as continuous flow. The concept of adaptive governance is not new. As early as 1927, it was proposed that "policies be treated as experiments, with the aim of promoting continual learning and adaptation in response to experience."[23] A more complete framework emerged in the 20th and early 21st century as a response to the challenges of managing complex social-ecological systems in rapidly changing environments.[24] Key features of this approach have been a focus on agility, distributed decision-making, and iteration so that governing bodies can learn and change with new information, shifting from the fixed policy process to a multi-step adaptive framework.

The success of adaptive governance mechanisms has been mixed: natural resource management, sustainable development, and some climate initiatives have shown tangible progress. However, barriers such as institutional inertia, lapses in accountability or opacity, power imbalances, polarization, and shifts in political administrations have made progress challenging.[25] These failures are instructive: they reveal that adaptive governance requires more than good intentions and flexible mandates. It requires specific structural properties that produce and sustain adaptive capacity regardless of the political environment.

There are precedents of large societal shifts requiring radical adjustments in governance systems and institutions: these arise when social complexity reaches a point where new regulatory paradigms become necessary to maintain stability. A notable example is the emergence of the modern Western administrative state as a response to the Industrial Revolution, requiring new mechanisms to stabilize labor markets and address social unrest as the economy evolved.[26] We are at such a juncture again, though the timescale of the current transition is vastly more compressed.

Our positive philosophy reconceives governance not as a sequence of discrete acts (laws passed, rules promulgated, decisions rendered) but as a continuous process operating across multiple timescales simultaneously.[27] Governance flow comprises four constitutive functions that must operate continuously and in coordinated relation:

4.1 Sensing

The first function is continuous observation of the regulated domain: capability development, deployment patterns, incident occurrence, and emerging risk surfaces. Sensing must be structurally independent from the entities being observed and must operate at a temporal resolution matching the dynamics of what it observes.

For AI, this means evaluation infrastructure capable of detecting capability changes on timescales of weeks rather than years. It means monitoring infrastructure capable of detecting deployment patterns and incident clusters in near-real-time. It means intelligence functions capable of identifying emerging risk surfaces before they manifest as realized harms. And critically, it means that sensing cannot be accomplished through self-reporting by AI developers or through periodic scheduled assessments alone. The system being sensed changes too rapidly, and the entities being monitored have inherent conflicts of interest in what they report. This does not mean that private sector participation in sensing is illegitimate. Much of the relevant technical information originates within private entities, and sensing infrastructure will necessarily involve cooperation with those entities. But the sensing function's structural independence requires that it not depend *entirely* on voluntary self-reporting, that it possess independent capacity to verify reported information, and that it have authority to access information that entities might prefer not to disclose.

Sensing functions must also account for the post-deployment capability evolution discussed above. Evaluation at the point of deployment captures only the system's capabilities under the conditions of evaluation. Continuous monitoring must detect when deployed systems exhibit capabilities exceeding their pre-deployment assessment, whether through inference-time scaling, novel elicitation techniques, multi-agent composition, or tool integration.

4.2 Evaluation

The second function translates sensed information into normative judgments through defined criteria. Evaluation requires explicit criteria that are public, contestable, and revisable; methodologies that are reproducible and subject to external review; and judgment functions that are independent from both the entities being evaluated and the political principals who may have conflicting interests in the evaluations' outcomes.

The criteria themselves must be subject to regular review and revision as understanding develops, but the revision process must be transparent and bounded rather than discretionary. This distinguishes effective evaluation from the kind of closed-door assessment that produces conclusions without accountability. When evaluation criteria are opaque, there is no way for external observers to determine whether evaluations are being conducted competently, whether their conclusions are justified, or whether they have been influenced by considerations unrelated to safety.

Evaluation must also be designed for adversarial conditions. AI developers have strong incentives to present their systems in favorable light during evaluations. AI systems themselves may, at sufficient capability levels, model the evaluation process and behave differently under evaluation conditions than in deployment.[28] Evaluation methodologies must anticipate and resist gaming, with red-team evaluation, surprise testing, and independent replication as structural features rather than occasional supplements.

4.3 Response

The third function translates evaluative judgments into protective action through pre-authorized mechanisms. Response mechanisms must include graduated options matched to severity (advisory, restrictive, prohibitive), automatic triggering for defined conditions to prevent paralysis at the moment of crisis, and discretionary authority for novel situations within bounded parameters.

The relationship between automatic and discretionary response is critical. Too much automaticity produces brittleness: rules that fire inappropriately because conditions don't match the scenarios they were designed for. Too much discretion produces capture and arbitrariness: decisions that reflect political convenience rather than risk assessment. The solution is layered: automatic mechanisms for clearly defined threshold crossings, bounded discretionary authority for situations within the framework's defined parameters, and escalation to higher authority for situations outside those parameters.

An instructive analogy exists in financial market governance. Circuit breakers in securities markets were established through democratic legislative process and operate on millisecond timescales, far faster than any deliberative body could act.[29] They demonstrate that governance architectures *can* incorporate mechanisms operating at the speed of the systems they govern, provided those mechanisms are

pre-authorized with clear triggers and defined responses. The governance architecture operates at the speed of the threat because it was designed to do so by a process that took the time necessary for legitimacy. The principle translates: AI governance need not choose between democratic authorization and rapid response. It needs democratic authorization *of* rapid response mechanisms.

4.4 Learning

The fourth function systematically incorporates experience into the sensing, evaluation, and response functions. Learning requires mandatory incident analysis, public after-action reviews, structured updating of criteria and methodologies, and feedback into criteria revision processes.

Learning must be public to be legitimate: when lessons are learned privately, there is no mechanism for external stakeholders to verify that learning has occurred, that correct lessons have been drawn, or that institutional behavior has actually changed. Learning must also be structured to avoid two failure modes: excessive volatility (constant revision that prevents stable operation) and excessive inertia (failure to update criteria even when clear evidence indicates they are inadequate). The resolution is procedural: defined triggers for review, defined processes for revision, and defined standards of evidence for changes, all operating transparently.

Critically, the learning function must include learning from failures, including failures of the governance system itself. Mandatory after-action reviews following governance failures, analogous to post-incident reviews in aviation safety, should be a structural requirement rather than a discretionary choice. These reviews should be public, should identify structural causes rather than individual blame, and should produce concrete changes to governance processes.

4.5 The Integration of Functions

These four functions, when properly designed and properly connected, constitute genuine adaptive governance. Their absence or distortion constitutes governance failure regardless of what activities are occurring under the banner of governance. Current arrangements exhibit characteristic distortions: sensing captured by industry partnership, evaluation lacking independence and transparency, response limited to discretionary executive action without durable foundation, and learning effectively absent or occurring only internally without public accountability.

Importantly, the functions must be connected in ways that maintain their integrity. Sensing must inform evaluation without sensing entities determining evaluative conclusions. Evaluation must inform response without evaluative entities controlling response decisions. Response must generate information that feeds learning without response entities controlling what lessons are drawn. And learning must influence sensing, evaluation, and response without any single function dominating the others. This separation of functions is a structural requirement for preventing capture: when the entity that senses is also the entity that evaluates, responds, and determines what was learned, the entire flow can be distorted by a single capture point.

5. Principles of Adaptive AI Governance

Building on this philosophy of governance flow, five key principles are essential for an effective adaptive governance framework: collectivity, adaptability, modularity, redundancy, and antifragility. These principles form the operational foundation of the philosophy by enabling governance systems to be more resilient and responsive to complex challenges. However, each principle must be understood not merely as an aspiration but as a design requirement with specific structural conditions for proper implementation, and with identifiable failure modes when those conditions are absent.

5.1 Collectivity

The principle of collectivity, including self-organization, describes governance structures that ensure diverse stakeholders collaborate for shared goals without centralized capture by any single constituency. Practically, a collective self-organizing system would emerge from intergroup deliberation and networking, potentially from working groups, panels of experts, and agency stakeholders, to derive policies that correspond to measured changes in capabilities. Initiatives in distributed decision-making could support these principles, allowing for a more decentralized and agile response to complex challenges.

Genuine collectivity is identifiable by a specific structural property: the participation of constituencies whose interests conflict with one another, and particularly the inclusion of voices whose interests conflict with those of AI developers and deployers. When governance processes include only entities with aligned commercial interests in AI deployment, the resulting consensus reflects shared interest rather than genuine collective deliberation. The presence of civil

society, academic researchers independent of industry funding, labor representatives, affected communities, and international counterparts is what distinguishes collective governance from bilateral negotiation between regulator and regulated.

It is well known that formal committees tend to slow or even block progress; thus, creative measures such as online collective polling and governance tools need further testing, research, and development (e.g., see Polis and related deliberative technology). But the solution to deliberative slowness is not to abandon collectivity in favor of closed bilateral arrangements; it is to develop deliberative infrastructure that enables meaningful collective input on compressed timescales. This infrastructure should be treated as part of the governance technology stack requiring investment and development.

The failure mode of collectivity is not merely its absence but its performative simulation: governance arrangements that include nominally diverse participants but structure participation such that well-resourced incumbents dominate through superior access, information asymmetry, or institutional design that privileges established positions. Genuine collectivity requires not just representative presence but substantive influence, and can be recognized by whether outcomes meaningfully reflect the input of constituencies without commercial stakes in the governance outcome.

5.2 Adaptability and Continuous Learning

Adaptability and continuous learning are key to ensuring that rules and policies can flexibly respond to change. The static nature of traditional governance measures cannot account for the continuous learning and evolution of AI systems, as the nature of their capabilities, application paradigms, and risk classes are evolving regularly. Like organisms in an ecosystem, AI regulations can benefit from adopting evolutionary learning principles, allowing for small experimental regulatory approaches (like regulatory sandboxes or pilot programs) that inform broader guidelines.

The EU's deployment of "regulatory sandboxes," virtual and physical environments to test AI systems to inform policy changes, is a good example of an approach that should be expanded.[30] Systems of novel testing strategies and experiments linked to policy adjustment are key to quickly highlighting change.[31]

However, the principle of adaptability must be carefully bounded to distinguish genuine adaptation from unbounded discretion. The distinction is that genuine adaptability operates within defined parameters with clear procedural requirements, where both the scope of what can be adapted and the process by which adaptation occurs are specified in advance. Unbounded discretion, by contrast, allows any change for any reason at any time. The first is adaptive governance. The second is government by whim, regardless of whether it occasionally produces good outcomes.

The structural property that distinguishes genuine adaptability is *bounded flexibility*: the capacity to change within constraints that are themselves durable. A governance arrangement is genuinely adaptive when the adaptation mechanism itself is stable even as its outputs change. This can be recognized by asking: is the process by which criteria are revised itself defined and public? Are there limits on how much can change in a single revision cycle? Is there a requirement that changes be justified by evidence and subject to review?

Government agencies should partner with regulatory bodies and external stakeholders to develop flexible mechanisms for responding to changing conditions. Importantly, these regulatory regimes should remain prospective, trying to remain ahead of future risks that could be particularly damaging rather than simply responding to failures.

The failure mode of adaptability is twofold: ossification (adaptive mechanisms that calcify into static frameworks because the political will or institutional capacity for revision is absent) and volatility (constant revision that prevents stable operation and allows interested parties to exploit each revision cycle for parochial advantage). Both must be anticipated and structurally resisted.

5.3 Modularity

The third principle, modularity, supports the distributed approach by breaking down problems into manageable components, preventing widespread failure or unmanageable administrative impasse. Modularity is a crucial element of complex systems involving nested hierarchies of systems within systems. Breaking down problems into more tractable, independent parts could elicit more granular and novel policy proposals; indeed, focusing on small-scale interventions on the same problem can allow greater variation of practical solutions as long as higher-order phenomena and feedback loops are taken into account.

However, modularity without coordination produces dangerous blind spots at the intersection points between modules. AI systems do not respect jurisdictional boundaries, and risks that emerge from the interaction between systems governed by different modules may be invisible to any single module. Effective modularity therefore requires not just the decomposition of governance into independent parts but also defined mechanisms for information-sharing across module boundaries, escalation of cross-cutting risks, and coordinated response when threats span boundaries.

Modularity also provides an important structural insight for multi-level governance. Different risk surfaces are best addressed at different governance levels: frontier model capabilities present externalities that cross all boundaries; application-layer harms manifest in specific communities and contexts; sector-specific risks require domain expertise accumulated within particular regulatory traditions. Breaking governance into modules matched to these risk surfaces allows specialized capacity to develop at each level while coordination mechanisms address interactions among levels. The critical principle is that eliminating any level of modular governance creates gaps that no other level is optimally positioned to fill.

The failure mode of modularity is fragmentation without coordination: modules that operate independently without mechanisms for detecting and responding to cross-module interactions. This produces a governance system that is locally rational but globally blind.

5.4 Redundancy

Redundancy ensures that the failure of any single governance node does not produce systemic collapse. In engineered systems, redundancy is a deliberate design choice: critical functions are performed by multiple independent components so that failure of any one component is compensated by others. In governance, redundancy means that multiple institutional nodes possess overlapping authority over critical functions, ensuring that the capture, defunding, politicization, or failure of any single node does not leave the system unprotected.

The case for deliberate redundancy in AI governance is particularly strong because the political environment introduces a form of volatility that pure engineering does not face: governance mechanisms can be dissolved not because they fail but because they succeed in ways that create political friction, or because they become inconvenient for current political leadership, or because unrelated political conflicts

redirect institutional resources and attention. In such an environment, governance architectures with single points of failure are predictably fragile.

Redundancy in AI governance takes several forms: multiple institutional nodes with overlapping evaluation authority at any single governance level; governance capacity across multiple levels such that failure at one level is partially compensated by others; international coordination structures providing redundancy against any single jurisdiction's failure; civil society monitoring providing redundancy against institutional capture; and open evaluation tools and methodologies providing redundancy against monopolized assessment capability.

The key structural insight is that governance monocultures are fragile in the same way that biological monocultures are fragile.[32] When a single governance approach, implemented through a single institution at a single level, constitutes the entire protective capacity, any perturbation that disrupts that specific arrangement (political change, institutional capture, leadership failure, resource withdrawal) produces total governance collapse. Diverse governance approaches implemented across multiple nodes, even if individually imperfect, provide systemic resilience that no single optimized approach can match.

The failure mode of redundancy is coordination failure: redundant nodes that produce conflicting requirements, accountability diffusion ("everyone's responsibility is no one's responsibility"), and regulatory arbitrage where actors exploit gaps between overlapping mandates. This failure mode is real and historically consequential. The pre-2008 financial regulatory architecture, where multiple agencies held overlapping authority over different aspects of systemic risk, produced mutual assumptions that others were monitoring the risks that in fact none adequately covered. The pre-9/11 intelligence community's failure to share threat indicators across agencies reflected redundancy without information-sharing obligations. These are not arguments against redundancy per se, but they are arguments that redundancy without *explicit coordination architecture* produces failures that can be catastrophic.[33]

However, this failure mode remains less dangerous than the alternative of single-point collapse, for a structural reason: coordination failures are *diagnosable and correctable within the existing architecture*, whereas single-point collapse eliminates the architecture itself. When redundant nodes fail to coordinate, the failure can be identified through after-action analysis and addressed through improved coordination protocols without dismantling the redundancy that provides

systemic resilience. When a single governance node fails, there is nothing to fall back on and nothing to correct: the protective function simply ceases.

The resolution is not to abandon redundancy but to design coordination mechanisms that maintain coherence without eliminating independence. Effective coordination among redundant governance nodes requires:

First, *defined escalation protocols* specifying when cross-node communication is mandatory rather than optional. When any node's sensing function detects a potential cross-boundary risk (e.g. a capability that spans multiple sectors, a deployment pattern that affects multiple jurisdictions, a threat that exceeds any single node's response capacity), escalation to a coordinating mechanism should be structurally triggered rather than left to discretion.

Second, *shared situational awareness infrastructure* that makes each node's assessment visible to others without requiring consensus before independent action. A common operational picture, analogous to shared threat registries in intelligence or shared dashboards in emergency management, allows redundant nodes to act independently while remaining aware of what others are observing and concluding. This prevents the scenario where each node assumes others are addressing a risk that in fact none has prioritized.

Third, a *lead coordinator function* for defined threat classes that activates specific coordination authority without establishing permanent hierarchy. For a given category of risk, one node holds primary responsibility: not exclusive authority, but the obligation to act if no other node does, and the authority to convene coordinated response. This assignment should be defined in advance for foreseeable threat categories and should rotate or be re-assigned as institutional capacities evolve.

Fourth, *mandatory post-event reconciliation* when redundant nodes produce conflicting assessments or responses. Disagreement among redundant nodes is not inherently a failure, as it may reflect genuine uncertainty or different analytical perspectives, but it must be surfaced, examined, and resolved through structured process rather than left unaddressed. Post-event reconciliation produces the learning that improves future coordination without requiring that all nodes always agree in advance.

Fifth, *explicit primary responsibility assignment* for defined risk categories, so that while multiple nodes *can* act, one node is *required* to act. This directly addresses

accountability diffusion: the existence of a named primary responsible entity for each risk category ensures that gaps in coverage are identifiable and attributable, even when secondary nodes provide redundant monitoring. The primary entity does not monopolize governance of its assigned risk, since other nodes retain authority to act, but it does bear the obligation that prevents the assumption that "someone else is handling it."

These coordination mechanisms add complexity to the governance architecture but are categorically preferable to the fragility of monoculture. An imperfectly coordinated redundant system remains functional; a perfectly designed single-point system that experiences failure does not.

5.5 Antifragility and Robustness

The concept of antifragility, as developed by Nassim Nicholas Taleb, describes systems that do not merely withstand shocks but grow stronger in response, improving their capacity to handle uncertainties.[34] Applied to governance, antifragility means designing regulatory systems that systematically improve their protective capacity through exposure to stress rather than merely surviving it.

Genuine antifragility requires specific mechanisms that translate stress into improvement: mandatory incident reporting and root-cause analysis that produces structural changes (not merely individual accountability); threshold tightening when failures occur (such that each governance failure triggers more protective criteria rather than less); public after-action reviews that surface systemic weaknesses and produce binding commitments to address them; competitive variation where multiple governance approaches are tested simultaneously and successful approaches are expanded while unsuccessful approaches are modified or retired.

To support measures based on these principles, any governance framework requires continuous monitoring, through multilevel data collection, risk assessments, and predictive modeling, to ensure that regulations and standards adjust to changing conditions. Rigid regulations aim for predictability and control, but shocks or disruptions are inevitable. Emphasizing antifragility means designing regulations that internalize new information from each disruption and emerge with greater protective capacity.

It is crucial to develop guardrails or partitions that monitor boundary conditions to delineate the operational space and prevent AI systems from surpassing various thresholds. These partitions must be designed for adversarial conditions: AI systems at sufficient capability levels may actively model governance mechanisms and exploit gaps between modules, time delays in response, or limitations in sensing capacity.

The failure mode of antifragility is superficial implementation: governance systems that describe themselves as adaptive because they survived a shock, without evidence that they actually improved from it. Mere survival is resilience, not antifragility. The distinction matters because governance systems that claim to be learning from stress without observable improvement are consuming political legitimacy (by claiming adaptive capacity) without delivering its benefits.

Not all governance functions can or should be antifragile. Some risks require robust prevention rather than learning-from-failure: catastrophic and irreversible harms cannot be treated as learning opportunities.[35] For these risk categories, robustness (prevention even under stress) is more appropriate than antifragility (improvement through stress). The governance philosophy must distinguish between domains where antifragility is appropriate (evaluation methodology, monitoring scope, coordination mechanisms) and domains where robustness is essential (prevention of catastrophic capability deployment, containment of systems exhibiting dangerous autonomous behavior).

6. Layered Defensive Architecture

Governance, understood as the institutional layer of societal protection, is one layer among several. A coherent philosophy of resilience must specify how governance relates to other defensive layers operating in parallel and must articulate why parallel development across all layers is essential. The substantive claim that follows is that resilience against rapidly evolving technological risk emerges from the interaction of multiple defensive layers operating at different timescales with different dependencies.[36] No single layer is sufficient alone, and sequential prioritization (developing one layer before beginning another) creates vulnerability windows that adversaries, accidents, and emergent harms can exploit. The appropriate posture is deliberate parallel development of all layers, with resources allocated based on each layer's development timeline rather than on which layer appears most urgent in any given moment.

6.1 The Technical Layer

At the most immediate timescale, the technical layer encompasses defensive capabilities embedded in AI systems themselves and in the tools used to evaluate them: alignment research bearing fruit in deployable form, interpretability techniques enabling external evaluation of system behavior, security properties limiting misuse potential, and defensive AI tools that protect critical systems against AI-enabled attacks.

This layer is developed primarily by technical researchers and AI developers, with governance providing incentives, constraints, and resources rather than directly producing the capabilities. The governance philosophy's relationship to the technical layer is primarily to create conditions under which defensive technical capabilities receive investment priority commensurate with their social importance, which market incentives alone will not produce. This relationship is inherently a public-private partnership: governance provides direction, incentives, and resources while private sector actors provide technical capacity, development infrastructure, and domain expertise. The quality of this partnership depends on whether its structure aligns private incentives with public protective goals rather than allowing public goals to be subordinated to private commercial interests.

6.2 The Infrastructure Layer

The infrastructure layer addresses the hardening of systems on which society depends: critical infrastructure cybersecurity, financial system resilience, biosecurity capacity, information ecosystem integrity, and democratic process protection. This layer is in some ways the most tractable on near-term timescales because much of the work involves applying known defensive practices with greater urgency and resources, informed by understanding of how AI-enabled attacks differ from previous threat models.

The infrastructure layer provides protection against both malicious use of AI capabilities and unintended cascading failures when AI systems are integrated into critical functions. It does not depend on controlling AI development itself; it focuses on ensuring that societal systems can withstand AI-enabled stresses regardless of how AI development proceeds. This independence from AI development timelines makes infrastructure hardening an essential complement to governance approaches that necessarily require longer development timescales. Because most critical infrastructure is privately owned and operated, the infrastructure layer depends heavily on public-private coordination for implementation. Governance's role at this

layer is to establish standards, provide incentives, mandate minimum resilience requirements, and coordinate across sectors, while the operational work of hardening is performed substantially by private actors within their own systems.

6.3 The Institutional Layer

The institutional layer is governance proper: statutory frameworks, regulatory institutions, international agreements, evaluation entities, enforcement mechanisms, and accountability structures. This layer is the slowest to develop because it requires democratic legitimacy that comes only from durable deliberative processes and constitutional foundations. It cannot be compressed beyond certain limits without sacrificing the legitimacy that makes it function.

However, the institutional layer provides something no other layer can: durable authority, democratic legitimacy, and structural independence from both the entities being governed and from the volatility of political cycles. Technical capabilities without institutional backing can be ignored. Infrastructure hardening without institutional coordination is piecemeal. International agreements without domestic institutional implementation are aspirational. The institutional layer is the backbone that makes other layers effective and enduring.

6.4 The Civil Society Layer

Independent monitoring, public-interest research, civic technology development, journalistic capacity, and democratic deliberation infrastructure constitute a layer that provides essential redundancy against institutional failure, capture, and political volatility. Civil society can be developed on faster timescales than institutional governance and provides resilience that institutional governance alone cannot.

Critically, civil society functions cannot be governmentalized without losing their independence. The value of civil society monitoring derives precisely from its structural independence from both government and industry. Public funding for civil society capacity is appropriate and necessary (ensuring that monitoring and research capacity are not dependent on industry goodwill), but governmental control over civil society activity is inappropriate and counterproductive.

6.5 The International Layer

Coordination among aligned jurisdictions on evaluation standards, deployment restrictions, export controls, defensive infrastructure, and incident response addresses externalities that no national governance can address alone. Capability shocks do not respect national borders: a model released in one jurisdiction affects security in all others. International coordination is essential but operates on timescales constrained by diplomatic process and sovereign consent.

6.6 The Interaction Among Layers

The layers interact in ways that a philosophy of governance must explicitly articulate.

First, sequential prioritization is wrong. The temptation to focus on whichever layer seems most tractable or most urgent produces vulnerability windows in the layers being deferred. All layers must develop in parallel, with resources allocated based on each layer's development timeline rather than on which layer appears most politically salient.

Second, layers compensate for each other's weaknesses. Strong technical capabilities reduce but do not eliminate the need for governance. Strong governance reduces but do not eliminate the need for civil society monitoring. Strong infrastructure hardening reduces but does not eliminate the need for international coordination. The compensations are partial and asymmetric, but they make the overall architecture resilient against failures at any single layer.

Third, the layers have different legitimacy bases. Technical layer legitimacy derives from scientific validity and operational effectiveness. Institutional layer legitimacy derives from democratic process and constitutional foundation. Civil society legitimacy derives from independence and public interest orientation. International layer legitimacy derives from sovereign consent and procedural fairness. Confusing these legitimacy bases produces failure: technical decisions made through political processes, governance decisions made through technical expertise alone, civil society functions captured by partisan interests.

Fourth, each layer can be developed by different actors on different timescales. This is the practical significance of layered architecture: even when institutional governance is politically blocked or captured, infrastructure hardening can proceed, civil society monitoring can expand, technical defensive capabilities can be

developed, and international coordination can advance. The overall system maintains protective capacity even when individual layers are degraded.

7. Legitimate Governance Under Uncertainty

Adaptive governance as articulated here faces a legitimacy challenge that must be addressed directly. Governance mechanisms operating on timescales faster than democratic deliberation can sustain face a real tension with democratic theory. If circuit breakers activate automatically, if evaluation criteria adjust periodically, if emergency response protocols fire within hours, where is democratic authorization? How can governance be both fast and legitimate?

The resolution rests on a distinction between *deliberative legitimacy* (legitimacy produced through democratic deliberation about specific decisions) and *structural legitimacy* (legitimacy produced through democratic deliberation about the structures within which decisions are made).[37] Deliberative legitimacy is impossible for fast-moving governance because the timescales are incompatible. Structural legitimacy is achievable and provides the foundation for legitimate rapid governance action.

7.1 Structural Legitimacy as Foundation

Democratic deliberation belongs at the level of governance architecture, not individual governance decisions. Legislatures should deliberate at length about what powers regulatory entities should have, what criteria should govern their decisions, what procedural protections should constrain them, and what accountability mechanisms should bind them. This deliberation can and should take the time necessary for substantive democratic engagement. The deliberation produces structural legitimacy that flows to the rapid decisions made within the structure.

Importantly, this framework does not require that every element of adaptive AI governance be authorized through new legislation purpose-built for AI. Existing statutory authorities across regulatory agencies (encompassing consumer protection, securities regulation, product safety, telecommunications, employment discrimination, environmental protection, and national security) already cover substantial portions of the AI risk surface. These authorities were democratically authorized through prior legislative processes, often crafted with deliberate breadth to encompass technological change unforeseen at the time of enactment, and carry

structural legitimacy that extends to novel applications within their defined scope. An agency applying its existing consumer protection authority to AI-enabled fraud, or its existing securities authority to AI-driven market manipulation, or its existing product safety authority to AI systems integrated into consumer devices, is exercising democratically authorized power within established parameters, not acting without authorization merely because the specific technology was not named in the enabling statute.

The structural legitimacy framework therefore distinguishes between two categories of governance action: first, the *application* of existing authorities to AI-specific manifestations of harms already within established regulatory mandates, which requires no new democratic authorization and should proceed with urgency; and second, the *creation* of novel governance mechanisms addressing risks genuinely outside existing statutory scope, classes of risk for which no prior democratic authorization exists because they could not have been anticipated, which requires new legislative action to achieve structural legitimacy. Much of the current governance gap, particularly for near-term harms involving discrimination, fraud, manipulation, product safety failures, and market disruption, could be addressed through the first category alone, if agencies possessed sufficient technical capacity, resources, and political will to exercise authorities they already hold. The failure to apply existing authority to AI-specific harms is fundamentally a capacity and will problem rather than an authority problem, and framing it as requiring new legislation when existing authority already suffices serves the interests of those who prefer governance delayed indefinitely.[38]

This distinction also establishes that adaptive governance can be substantially bootstrapped within existing legal infrastructure. Agencies can develop sensing and evaluation capacity under existing investigatory authorities. They can apply existing enforcement mechanisms to AI-specific harms within their jurisdiction. They can build institutional expertise, staff technical capacity, and establish precedent that later informs purpose-built legislation. They can promulgate rules within existing rulemaking authority that apply established principles to AI-specific contexts. The ideal of comprehensive adaptive AI governance authorized through deliberate, AI-specific legislative design remains the ultimate goal, particularly for novel risk categories involving autonomous systems, recursive self-improvement, and societal-scale systemic effects that genuinely exceed any existing mandate. But this ideal should not be treated as the precondition for all governance action, nor should its absence be accepted as justification for leaving existing authorities unexercised.

This is not a novel principle. It is the foundation of administrative law: legislatures authorize agencies to make decisions within defined parameters, and the decisions inherit legitimacy from the authorization. What is novel is the degree of adaptivity that must be authorized. Traditional administrative delegation authorizes agencies to apply fixed criteria to particular cases. Adaptive governance requires authorization for governance entities to modify criteria themselves, within defined parameters, based on changing conditions. This is a broader delegation that requires correspondingly stronger constraints.

The implication is clear: adaptive governance mechanisms that operate without democratic authorization are structurally illegitimate regardless of their technical competence or good intentions. Emergency executive coordination may be pragmatically necessary in the absence of durable foundations, but it cannot substitute for those foundations. It should be understood as a temporary measure that creates urgency for institutional development rather than as a permanent governance mode.

7.2 Procedural Protections as Legitimacy Mechanisms

When decisions must be made faster than deliberation can sustain, procedural protections become more important rather than less. Requirements for public reporting, availability of review by independent bodies, legislative oversight mechanisms, sunset provisions, and contestability mechanisms do not require slow decision-making; they require transparent decision-making with opportunities for correction.

Transparency is not a luxury but a legitimacy requirement. Closed-door governance cannot be legitimate at speed because its legitimacy depends entirely on structural protections that require public scrutiny to remain functional. The argument that transparency must yield to security or competitiveness concerns must be examined skeptically, with narrow exceptions subject to independent review rather than broad exceptions subject to executive discretion.

Reversibility becomes structurally important as a legitimacy property. Decisions made quickly may be wrong, and the legitimacy of fast governance depends partly on the capacity to correct errors. Governance architectures should build in reversibility: sunset provisions that force reconsideration, mandatory review following defined periods, contestability mechanisms available to affected parties, and graduated rather than absolute responses where possible.

7.3 The Implications for Current Arrangements

This legitimacy framework has direct implications for evaluating current governance arrangements. Governance mechanisms operating through executive discretion without durable democratic authorization lack structural legitimacy regardless of their practical utility. They may produce useful information, they may even produce good outcomes, but they cannot provide the durable, accountable, transparent governance that democratic societies require and that the challenge of advanced AI demands.

This is not an argument against rapid response but rather an argument that rapid response capacity must be pre-authorized through democratic process, constrained by procedural protections, and accountable through multiple pathways. The speed of governance is compatible with its legitimacy when the architecture is properly designed. Financial market circuit breakers demonstrate this: they act within democratic authorization, with full transparency, subject to review and revision. AI governance can follow the same structural pattern.

8. Differential Development and International Dynamics

The governance of advanced AI cannot be addressed in isolation from the international environment in which development occurs. Governance choices in one jurisdiction affect development incentives, competitive dynamics, and risk distributions across all others. A comprehensive philosophy must articulate how governance can shape development trajectories rather than merely responding to them, and how international coordination can be achieved under conditions of strategic competition, divergent values, and rapidly shifting capability distributions.

8.1 Differential Development as Governance Strategy

Not all AI capabilities present equivalent risk profiles, and governance can and should differentiate among them. This principle, which we term differential development, holds that governance choices should preferentially support development of capabilities that favor defense, protection, broad distribution, and societal resilience, while creating appropriate friction for capabilities that favor offense, concentration, or autonomous action beyond human oversight.[39]

The categories are not always crisp, but the principle provides directional guidance:

Capabilities that favor defenders over attackers, including defensive cybersecurity tools, anomaly detection, interpretability research, and alignment techniques, should receive preferential support through governance incentives, public funding, and reduced regulatory burden. These capabilities increase societal resilience regardless of how offensive capabilities develop.

Capabilities that favor attackers over defenders, including autonomous offensive cyber capability, capabilities for generating novel biological agents, and capabilities for large-scale manipulation of human behavior, warrant heightened scrutiny, controlled deployment, and international coordination to prevent proliferation.

Capabilities that support distributed protective capacity, such as open evaluation tools, widely accessible defensive systems, and interoperable monitoring infrastructure, should receive preferential support as elements of the broader defensive architecture.

This is not a simple permissive/restrictive dichotomy. It is a governance posture that consciously shapes the development landscape to favor outcomes compatible with broad human flourishing, democratic governance, and distributed resilience. It requires continuous assessment of which capabilities fall into which categories as the technology evolves, which connects back to the sensing and evaluation functions of governance flow.

The assessment of whether specific capabilities favor defense or offense is itself a complex analytical task requiring structured frameworks rather than ad hoc categorization. Corsi, Kilian, and Mallah (2024) develop a taxonomy of factors influencing offense-defense dynamics in AI systems, providing a systematic basis for evaluating the directional tendency of specific capabilities and identifying the structural properties that determine whether a given advance predominantly enables harm or enhances protection.[40] This taxonomy demonstrates that offense-defense categorization, while not always crisp at the level of individual capabilities, can be disciplined by attention to specific structural factors: the degree of asymmetry in required resources between attacker and defender, the reversibility of harms enabled, the scalability of defensive versus offensive applications, the degree to which capability diffusion favors broad protective capacity versus concentrated threat potential, and the existing balance of advantage in the relevant domain.

These factors provide the analytical foundation for governance decisions about differential support even when individual capabilities exhibit dual-use properties, as many AI capabilities inevitably do. The dual-use nature of many AI capabilities is not an argument against differential development as a governance strategy; it is an argument that differential development requires structured analytical frameworks rather than simple binary categorization, especially as there are multiple domain axes on which to gauge offense-defense balance. A capability that is dual-use but whose structural properties (such as resource asymmetry, scalability characteristics, or diffusion dynamics) favor defensive application on balance can still be preferentially supported, with appropriate safeguards for its offensive potential, rather than treated identically to a capability whose structural properties favor offensive exploitation. The governance task is not to eliminate ambiguity but to develop principled, revisable frameworks for reasoning about directionality under uncertainty.

8.2 Coalitions for Coordination

Effective international coordination does not require comprehensive multilateral agreement including all major powers. Smaller coalitions of jurisdictions sharing specific properties, including technical capacity sufficient for meaningful evaluation, governance infrastructure capable of implementing restrictions, rule-of-law traditions enabling enforcement, and demonstrated commitment to evaluation standards, can develop interoperable frameworks on faster timescales than universal bodies allow.

Such coalitions can pursue shared evaluation standards (so that a model evaluated in one jurisdiction need not be fully re-evaluated in another), coordinated deployment restrictions (so that models restricted in one jurisdiction cannot simply be deployed from another), shared defensive infrastructure (so that jurisdictions with fewer resources benefit from collective investment in protective tools), and joint incident response (so that capability shocks affecting multiple jurisdictions receive coordinated rather than fragmented responses). They can also demonstrate that compute verification and the adversarial cooperative technical governance regimes that compute governance can enable do indeed work for adversarial safety thresholding in the real world.[41]

The strategic value of coalitional approaches includes reduced coordination costs enabling faster action, alignment on values enabling substantive rather than merely procedural agreement, and practical interoperability of governance

mechanisms enabling mutual reinforcement. As coalitions demonstrate effectiveness, their membership can expand based on whether additional jurisdictions meet the structural conditions for participation.

8.3 Governance Under Expanding Capability Distribution

A critical governance design challenge arises from the observation that algorithmic improvements are rapidly lowering the capability threshold for producing dangerous AI systems.[42] Governance architectures designed around the assumption that dangerous capability is concentrated among a small, identifiable set of cooperative entities face an increasingly poor fit with reality as the set of actors possessing such capability expands and the identities of those actors become less predictable.

This trajectory has two implications for governance design. First, governance cannot depend on specific cooperative relationships with specific entities. Relationships with individual developers are inherently unstable: they can be disrupted by political conflicts, commercial pressures, corporate restructuring, or simple changes in leadership and priorities. Governance that functions only when particular entities choose to cooperate is governance that functions only by permission of the governed, which is not governance at all.

Second, governance must be designed to function effectively as the number of relevant actors grows, their identities shift, and not all of them choose to cooperate. This implies that governance architectures need authority to compel rather than merely request cooperation; evaluation capacity that does not depend on any single developer's participation; monitoring that functions regardless of developer posture; and response mechanisms effective against non-cooperating entities.

The combination of these observations yields a clear design requirement: governance must be robust to a future in which dangerous capability is more broadly distributed, actors possessing it are less predictable, and cooperative relationships cannot be assumed. Architectures meeting this requirement will also function in the current, more concentrated environment. The reverse is not true: governance designed for a small set of cooperative frontier developers will prove inadequate as the actor landscape evolves.

But this observation also carries an urgent strategic implication: *the window for establishing effective governance architectures narrows as capability diffuses*. The

governance task is structurally different, and dramatically easier, when dangerous capability is concentrated among a small number of identifiable, jurisdiction-bound, and potentially cooperating entities than when it has proliferated to numerous, geographically dispersed, potentially unidentifiable, and potentially non-cooperating actors. This asymmetry creates a compelling case for rapid legislative action and international agreement, not merely because current risks demand governance, but because the future governance task becomes *qualitatively harder* with each year of delay.

Consider the specific mechanisms through which delay degrades governance feasibility. A statutory framework establishing mandatory evaluation requirements, reporting obligations, and enforcement authority over AI development and deployment can be implemented relatively straightforwardly when the entities subject to governance are identifiable, domestic or allied, and operating through visible infrastructure (large compute clusters, known research facilities, regulated cloud providers). Extending such a framework to new entrants as they emerge is an incremental governance challenge, expanding scope within an established architecture. But constructing such a framework from scratch after capability has already proliferated to numerous actors, some operating outside accessible jurisdictions, some using distributed or concealed infrastructure, some potentially hostile to governance itself, is not merely a harder version of the same task: it is a categorically different problem requiring different tools (intelligence-based identification, cross-border enforcement, infrastructure-level controls) that are themselves far harder to establish.

The same logic applies to international coordination with greater force. International agreements on AI governance are achievable, though demanding, while the set of frontier-capable actors is small enough for meaningful multilateral negotiation and while the jurisdictions hosting frontier development share sufficient interests to sustain binding commitments. The current moment, in which a coalition of perhaps a dozen jurisdictions hosts the overwhelming majority of frontier development capacity, represents a window in which coordinated frameworks (such as shared evaluation standards, interoperable deployment restrictions, mutual recognition agreements, and coordinated export controls) can be negotiated among parties with both the technical capacity to implement and the strategic interest to participate.[43] As algorithmic efficiency improvements and hardware proliferation expand the set of relevant actors, the coordination problem scales combinatorially: more parties whose agreement is required, more divergent interests to reconcile,

more complex verification requirements, and more opportunities for defection by actors outside the coalition.

This temporal narrowing transforms governance from a policy preference into a time-critical imperative. Legislatures and international bodies that defer establishment of statutory and treaty frameworks until capability has diffused broadly will find the governance task they face has become not merely quantitatively larger but qualitatively intractable by the mechanisms available to democratic governance. The appropriate response is to pursue legislative foundation and international agreement with urgency commensurate to the pace of capability diffusion, understanding that each month of delay does not merely postpone governance but materially reduces the probability that effective governance remains achievable through legitimate institutional means. Governance frameworks must be designed for the expanding-actor future, but they are far easier to *establish* in the present concentrated-actor environment and then adapt, than to construct from scratch once the window of concentration has closed.

8.4 Capability Externalities as Design Constraints

The challenge that safety-oriented governance in one jurisdiction might accelerate unsafe development elsewhere is real but should be treated as a design constraint rather than a reason governance is impossible. Specific mechanisms can address this externality:

Export control coordination among coalitional members ensures that restrictions apply broadly rather than merely displacing development. Compute governance applied internationally prevents circumvention through geographic relocation. Shared evaluation infrastructure creates cooperative rather than competitive dynamics among aligned jurisdictions: when evaluation is a shared public good rather than a unilateral burden, it ceases to create competitive disadvantage. And investment in defensive capabilities, because they protect regardless of where threats originate, sidesteps the externality problem entirely for the infrastructure layer.

The argument that governance is futile because it will be undercut by less responsible jurisdictions is strategically convenient for those opposing governance but analytically weak.[44] Many governance domains face analogous dynamics (tax policy, environmental regulation, financial regulation) and have developed

mechanisms for managing them. The appropriate response is not abandonment of governance but design of governance that addresses the externality directly.

8.5 Governance for Open-Weight Models: Societal Resilience as Governance Strategy

The governance flow framework as articulated in Section 4 applies most directly to centrally deployed AI systems where an identifiable deployer maintains ongoing operational control and can be held to pre- and post-deployment obligations. Open-weight models, i.e. systems whose parameters are publicly released and can be downloaded, modified, fine-tuned, and deployed by any actor without ongoing relationship to or oversight by the original developer, present a structurally distinct governance challenge. Once model weights are public, deployment-stage interventions become technically impossible: there is no deployment choke point, no ongoing service that can be suspended, no centralized infrastructure that can be restricted. The model exists as distributable information, and governance mechanisms premised on controlling information after widespread release face fundamental feasibility constraints.[45]

This structural reality does not render governance of open-weight models impossible, but it requires governance to operate at different points in the lifecycle, to employ different levers, and crucially, to shift strategic emphasis from *controlling the model* to *strengthening the societal systems that deployments affect*. This shift connects open-weight model governance to the broader principle of layered defense (Section 6): when one defensive layer (deployment control) is structurally unavailable, other layers must compensate with enhanced protective capacity. Adaptive governance for open-weight models is substantially *disaster preparedness and societal resilience governance*: building the institutional, technical, and infrastructural capacity to withstand AI-enabled harms regardless of the diffusion state of the underlying capabilities.

The governance flow functions apply to open-weight models as follows, with adaptations reflecting the structural constraints:

Sensing shifts from monitoring specific deployments to ecosystem-level observation: tracking the capability frontier of publicly available models and their fine-tuned variants; detecting novel elicitation techniques, scaffolding architectures, and tooling combinations that unlock dangerous capabilities from base models assessed as safe in isolation; monitoring misuse patterns through incident

reporting, platform-level detection, and harm-specific indicators (biosecurity monitoring for novel agent synthesis attempts, cybersecurity monitoring for novel attack patterns, information integrity monitoring for novel manipulation techniques); and identifying the emergence of capability compositions (multiple open models combined through agent architectures) that exceed the capability of any individual component. This is *ecologically-oriented sensing* rather than entity-oriented sensing: it monitors the capability ecosystem rather than individual deployments.

Evaluation bifurcates into pre-release and post-release functions. Pre-release evaluation, the assessing of model capabilities and risks before public weight release, can be mandated through statutory requirements on the *act of release* rather than the act of deployment. The developer who trains a model and chooses to release its weights publicly is performing an identifiable, jurisdiction-bound act at a specific point in time, and this act can be subject to regulatory requirements including mandatory capability evaluation against defined thresholds, structured release protocols (staged access periods, researcher-only availability before general release, documentation requirements), and in cases where evaluation reveals capabilities exceeding defined thresholds, prohibition or conditional restriction of public release. Post-release evaluation becomes a continuous function distributed across the research community, civil society, and governance bodies: monitoring what fine-tuned variants emerge, what capability elicitation techniques are discovered, what downstream harms materialize, and whether the model's realized harm profile matches pre-release assessments.

Response necessarily emphasizes different levers than deployment restriction, organized in temporal sequence:

Pre-release response includes mandatory capability evaluation before public weight release, with thresholds that adapt (per Section 5.2) as understanding develops; structured release protocols providing graduated access rather than binary public/private (research-access tiers, delayed full release, capability-specific restrictions on weight availability); and in extreme cases, prohibition of public weight release for models exceeding defined capability thresholds, enforceable against the identifiable releasing entity.[46]

Post-release response shifts to harm mitigation and societal resilience: hardening the systems that foreseeable misuse targets (critical infrastructure cybersecurity hardening against AI-enabled attacks, biosecurity monitoring systems capable of

detecting AI-assisted pathogen design, information ecosystem tools that detect and attribute AI-generated manipulation); developing and deploying defensive AI tools that detect and counter misuse patterns at scale; building rapid-response capacity for novel harm vectors as they emerge from the open-weight ecosystem; and maintaining law enforcement and attribution capacity sufficient to identify, locate, and hold accountable specific actors who misuse openly available models for defined harmful purposes, even when the underlying model cannot be recalled.[47]

Learning draws on the full ecosystem of open-weight deployment, including the harms that materialize, the defenses that succeed, the attack patterns that emerge, and the governance interventions that prove effective or ineffective, to inform both future release decisions (adapting pre-release thresholds and evaluation criteria) and ongoing resilience investments (directing infrastructure hardening toward demonstrated rather than speculative threat vectors).

This reframing establishes that governance for open-weight models is not primarily about preventing model access (though pre-release restrictions remain appropriate for models above defined capability thresholds) but about building the societal capacity to maintain safety and stability regardless of the model access landscape. This includes: investing in the infrastructure layer (Section 6.2) with explicit attention to AI-enabled threat vectors; strengthening the civil society layer (Section 6.4) with open-source evaluation tools, independent red-teaming capacity, and community-driven harm detection; developing the technical layer (Section 6.1) of defensive AI capabilities that can be widely deployed as public goods; and maintaining the institutional layer's (Section 6.3) capacity to enforce pre-release requirements and pursue post-hoc accountability for misuse.

A governance posture that combines pre-release restrictions where technically feasible and proportionate with sustained investment in societal resilience against foreseeable misuse provides robust protection even under conditions where restrictions fail, are circumvented through jurisdictional arbitrage, or are politically overridden. Governance designed solely around preventing release will prove brittle as capability thresholds decline with algorithmic progress and as the political economy of open-source AI creates powerful constituencies favoring unrestricted release. Resilience-focused governance, by contrast, maintains protective value regardless of the release regime and provides insurance against the possibility that pre-release restrictions prove insufficient. Both strategies should be pursued in parallel, with neither treated as substitute for the other: restriction where feasible, resilience regardless.

9. The Separation and Coordination of Functions

A comprehensive governance philosophy must articulate principles governing how functions are allocated across institutional forms and levels. This section addresses not *which* institutions should perform *which* functions (a question of institutional design specific to particular jurisdictions and political contexts) but rather the structural principles that should guide any such allocation.

9.1 Why Separation Matters

The governance flow functions (sensing, evaluation, response, learning) should be distributed across different institutional entities rather than concentrated in a single body. The reasoning is structural: when a single entity performs all functions, the entire flow can be distorted by a single capture point. An entity that both evaluates models and decides whether to restrict them faces inherent conflicts in its evaluative function. An entity that both monitors compliance and learns from failures faces incentives to underreport failures that reflect poorly on its monitoring. Structural separation creates mutual accountability among functions that concentration eliminates.

9.2 Different Functions Require Different Properties

Different governance functions have different structural requirements, which implies that no single institutional form is optimal for all functions.

Sensing requires technical capacity, temporal resolution, and access to information from the entities being monitored. It benefits from proximity to the technical domain and from the kind of sustained attention that specialized institutions provide. However, sensing entities must be structurally independent from the entities they observe, which means they cannot depend operationally on the cooperation of those entities for their basic function.

Evaluation requires both technical expertise and normative judgment: the capacity to determine not just what a system can do but what level of capability constitutes unacceptable risk. It benefits from insulation from both commercial pressures and short-term political interests, because evaluative conclusions that displease powerful actors must be protectable from retaliation.

Response requires authority (the capacity to compel behavior), legitimacy (democratic authorization for the exercise of that authority), and speed (the capacity

to act within timeframes relevant to the threat). These properties are in tension: authority and legitimacy require democratic foundations that take time to construct, while speed requires pre-authorization that may not anticipate all relevant scenarios.

Learning requires openness (access to information about failures and near-misses), independence (freedom to draw conclusions that may displease powerful actors), and influence (the capacity to translate lessons into changes in practice). It benefits from being distributed across multiple institutional forms, since different analytical perspectives surface different lessons from the same events.

9.3 Multi-Level Governance and Functional Complementarity

Different risk surfaces are best addressed at different governance levels, and the relationship among levels should be one of functional complementarity rather than hierarchy or competition.

Some risks present externalities that cross all jurisdictional boundaries: frontier model capabilities that could enable mass-casualty attacks, autonomous systems that propagate across networks regardless of geography, and information operations that target populations across borders. These risks require coordination at the broadest feasible level because governance at any narrower level creates arbitrage opportunities.

Other risks manifest primarily in specific deployment contexts affecting specific communities: algorithmic discrimination in hiring or lending, deepfakes targeting local elections, AI-enabled fraud targeting particular populations. These risks are best addressed by governance bodies with proximity and accountability to the affected populations, with local knowledge that broader governance levels lack.

Still other risks operate within specific sectors where pre-existing expertise, relationships, and regulatory frameworks provide essential context: AI in healthcare, in financial services, in aviation, in energy systems. Sectoral governance that leverages accumulated domain knowledge can address risks that generalist AI governance would struggle to evaluate.

The critical principle is that eliminating any level creates coverage gaps that no other level is optimally positioned to fill. This argues against governance consolidation that removes either the most local or the most international levels,

and argues for explicit coordination mechanisms that allow levels to communicate, escalate, and support one another without requiring any single level to dominate.

9.4 The Role of Public-Private Partnership

The private sector possesses technical expertise, operational infrastructure, and informational access that governance cannot replicate independently within any plausible timeframe or budget. Most critical infrastructure is privately owned. Frontier AI capability is developed by private entities. The talent pool with relevant technical expertise is concentrated substantially in industry. Any governance architecture that does not incorporate structured partnership with private actors will lack the capacity to perform its functions effectively.

The philosophy articulated here therefore does not reject public-private partnership. It demands that such partnerships be designed with structural properties that maintain the integrity of governance functions. The diagnostic criteria developed in Section 10 provide the relevant test: a public-private partnership that maintains independence (governance conclusions are not operationally dependent on the partner's approval), transparency (the partnership's operations and outputs are publicly observable), durability (the partnership survives changes in any single participant), and authority (governance can compel cooperation when voluntary engagement fails) is not merely legitimate but essential.

The distinction the philosophy draws is between partnership *as structural design* and partnership *as substitute for authority*. When partnership is designed with the structural properties above, it leverages private capacity in service of public goals. When partnership operates without those properties, it functions as a veto granted to the governed over the governing: cooperation proceeds only so long as the governed party finds it comfortable, and withdraws when governance becomes inconvenient. The first is genuine partnership while the second is deference mistaken for cooperation.

9.5 Institutional Design Principles: Lessons from Commons Governance

The governance architecture articulated in this paper — polycentric, multi-level, principle-based, combining self-organization with defined constraints — shares deep structural properties with Elinor Ostrom's design principles for long-enduring common-pool resource institutions.[48] This convergence is not coincidental: both

frameworks address the governance of shared resources under conditions of interdependence, uncertainty, incomplete information, and the ever-present risk of defection by individual actors pursuing private benefit at collective expense. AI governance can be productively understood as governing a particularly challenging commons: one where the shared resource, the very maintenance of safe AI development and deployment conditions, is non-excludable, where free-riding (developing or deploying AI without adequate safety investment) produces negative externalities borne by all, where the resource itself is evolving in ways that alter governance requirements, and where the governed system may, at sufficient capability levels, actively resist governance constraints.

Ostrom's eight design principles, derived inductively from extensive empirical study of successful and failed commons governance arrangements worldwide, provide both validation for the architecture proposed here and instructive guidance for its refinement.

Clearly defined boundaries. Ostrom found that long-enduring commons institutions define clearly who is entitled to participate and what constitutes the governed resource. The parallel in AI governance is the scope adequacy criterion (Section 10.6): governance must define clearly which actors, which capabilities, which deployment contexts, and which risk categories fall within its jurisdiction. Ambiguity in boundaries produces both gaps in coverage (actors or risks falling between jurisdictions) and illegitimate overreach (governance extending beyond its warranted scope). However, a critical divergence from commons governance must be noted: Ostrom's boundaries are typically *exclusionary* (defining a stable set of authorized participants), whereas AI governance faces an *expanding* actor set as capability diffuses. This means that boundary definitions in AI governance must be *capability-referenced* (applying to any entity whose systems exceed defined thresholds) rather than *entity-referenced* (applying to a named set of actors), to maintain coverage as the governed population changes.

Congruence between rules and local conditions. Successful commons institutions tailor their operational rules to the specific ecological and social conditions they face, and the costs imposed on participants are proportional to the benefits they receive. This principle validates the paper's emphasis on adaptability (Section 5.2) and multi-level governance (Section 9.3): rules must be context-sensitive, matching the specific risk profile of the relevant capability, deployment context, and affected community. It also carries a legitimacy implication: governance arrangements perceived as imposing disproportionate costs

on some actors (for example, imposing identical compliance burdens on small open-source developers and on well-resourced frontier laboratories) will face legitimacy challenges and resistance independent of their technical merit. Proportionality in governance burden is not merely equitable but instrumentally necessary for durable compliance.

Collective-choice arrangements. Ostrom found that commons governance endures when most individuals affected by operational rules can participate in modifying those rules. This directly validates the collectivity principle (Section 5.1) and strengthens the argument for inclusive governance processes that extend beyond bilateral negotiation between regulator and regulated. Ostrom's finding also carries a procedural insight: participation must be *practical*, i.e. low-cost, accessible, and understandable, rather than merely formal. Governance processes that are nominally open but practically accessible only to well-resourced actors with dedicated government-affairs staff do not satisfy the collective-choice principle. This reinforces the paper's observation that creative measures for scaled deliberation (digital deliberation tools, collective intelligence mechanisms) require investment and development as governance infrastructure.

Monitoring. Long-enduring commons institutions maintain active monitoring of both resource conditions and participant behavior, with monitors accountable to the community. The parallel to the sensing function (Section 4.1) is direct, but with an important divergence: Ostrom found that *self-monitoring* by community members often works in commons governance because participants share stakes in the resource's sustainability and social enforcement mechanisms (reputation, repeated interaction, community sanction) punish defection. In AI governance, the analogous "community" of developers does not necessarily share stakes in governance success: individual developers may benefit from evading governance even as the collective suffers, and social enforcement mechanisms are weaker in a highly competitive commercial environment with high employee mobility and limited community sanction capacity.[49] This structural difference in incentive alignment justifies the paper's more adversarial posture toward monitoring: independent evaluation capacity, resistance to gaming, structural independence from monitored entities, and the capacity to verify rather than merely receive self-reports. Where Ostrom's commons can often rely on community self-monitoring supplemented by external oversight, AI governance requires robust independent monitoring supplemented by (but not dependent on) developer cooperation.

Graduated sanctions. Ostrom found that effective commons governance employs graduated sanctions: small penalties for first-time or minor violations, escalating to severe consequences for repeated or egregious defection. This validates the response function's emphasis on "graduated options matched to severity" (Section 4.3) and carries an additional insight: graduated sanctions serve a *signaling* function beyond their direct punitive effect. They communicate that violations are detected and taken seriously without immediately imposing consequences so severe that they destroy the cooperative relationship or drive the sanctioned actor entirely outside the governance framework. For AI governance, this translates to the importance of intermediate enforcement actions (such as advisory notices, mandatory remediation plans, conditional deployment restrictions, and enhanced monitoring requirements) before escalation to prohibition. Governance that possesses only advisory capacity and prohibitive capacity, without intermediate options, loses the signaling and cooperation-maintaining functions that graduated sanctions provide.

Conflict-resolution mechanisms. Successful commons institutions provide rapid, low-cost, locally accessible mechanisms for resolving disputes among participants and between participants and governance authorities. This principle highlights a gap in the current paper that requires development: when redundant governance nodes (Section 5.4) produce conflicting assessments or requirements, affected actors need recourse to rapid resolution rather than being caught between incompatible mandates. Similarly, when governance decisions are contested by affected parties, the availability of rapid, accessible review mechanisms is essential for both legitimacy and operational function. Conflict resolution among governance nodes must be rapid enough to prevent prolonged regulatory paralysis and low-cost enough to be invoked routinely rather than reserved for crises. This argues for standing inter-institutional coordination mechanisms with defined resolution procedures rather than ad hoc political intervention when conflicts arise.

Minimal recognition of rights to organize. Ostrom found that commons governance collapses when external authorities override local governance arrangements without regard for local conditions or substitute their own judgment for that of the community that has developed effective local rules. For AI governance, this principle has two implications. First, it provides Ostrom-derived empirical support for the paper's argument that lower-level governance entities (state, sectoral, community) must retain functional autonomy and not be preempted by higher-level governance except for genuinely cross-boundary risks. Federal or international preemption of effective local governance arrangements without replacing their protective function destroys governance capacity without

compensating benefit. Second, it reinforces Section 6.4's argument that civil society governance functions must not be governmentalized: their value derives from independence, and external authority that co-opts or constrains independent monitoring and evaluation destroys the very properties that make civil society governance valuable.

Nested enterprises. For larger commons, Ostrom found that governance must be organized in multiple nested layers, with each layer addressing the governance tasks appropriate to its scale. This directly validates the multi-level governance architecture of Section 9.3 and provides empirical grounding for the claim that eliminating any level creates governance gaps. Ostrom's empirical work further specifies that successful nesting requires clear rules about which level handles which functions and defined interfaces between levels for escalation, information-sharing, and conflict resolution. This identifies a refinement opportunity for the current framework: the coordination mechanisms between governance levels require the same kind of explicit specification as the coordination mechanisms between redundant nodes at the same level, including defined triggers for escalation, protocols for information-sharing across levels, and decision rules for allocating novel risks to appropriate governance levels.

The convergence between Ostrom's empirically derived principles and the theoretically derived architecture proposed here provides mutual validation: the governance properties this paper identifies as necessary for AI (polycentric structure, adaptive rules, inclusive deliberation, independent monitoring, graduated response, multi-level nesting) are the same properties that Ostrom found empirically distinguishing successful from failed governance of complex shared resources. This convergence also provides a cautionary function: Ostrom identified these principles by studying *why institutions fail* as much as why they succeed, and her work documents numerous cases where the absence of even one principle (particularly monitoring, graduated sanctions, or conflict resolution) produced institutional collapse. The diagnostic criteria of Section 10 can be understood as AI-governance-specific elaborations of Ostrom's principles, adapted for a domain where the governed system changes faster, the stakes of failure are higher, the incentive divergence among actors is greater, and the system itself may actively resist governance.[50]

10. Diagnostic Criteria for Genuine Adaptive Governance

The philosophy articulated above generates specific criteria for evaluating whether any particular governance arrangement constitutes genuine adaptive governance or merely its performative simulation. These criteria serve as analytical tools applicable across institutional contexts: legislative proposals, executive arrangements, voluntary commitments, international agreements, and private governance can all be evaluated against them.

10.1 Independence

Are the entities making evaluative judgments structurally independent from the entities being evaluated and from political principals who may have conflicting interests? Independence is observable through the following indicators: whether the evaluating entity's continued operation depends on the cooperation of entities being evaluated; whether evaluative conclusions unfavorable to powerful actors can be maintained without institutional consequences; whether the evaluating entity's leadership, funding, and mandate are insulated from retaliation by those affected by its conclusions; and whether the entity's analytical conclusions are observably distinct from the preferences of its operational partners.

Governance arrangements where the evaluated entities are also the primary operational partners of the evaluating entity, or where the evaluating entity describes its function in relational terms ("primary point of contact," "partner," "facilitator") rather than authoritative terms, exhibit structural indicators of compromised independence regardless of the personal integrity of individual evaluators. This criterion does not preclude operational cooperation with industry — such cooperation is structurally necessary — but it distinguishes between cooperation that supplements independent capacity and cooperation that substitutes for it. An evaluating entity that has independent technical capacity but also accepts information from industry maintains independence. An evaluating entity that cannot function without industry's active participation has ceded independence to its partners regardless of formal arrangements.

10.2 Transparency

Are evaluation criteria, methodologies, results, and decisions public? Can civil society, academic researchers, and affected communities scrutinize the governance system itself? Transparency is observable through: public availability of the criteria

by which systems are evaluated; public reporting of evaluation results within defined timeframes; external replicability of evaluative methodologies; availability of sufficient information for independent researchers to assess whether evaluations are being conducted competently; and visibility of the governance system's own performance (e.g. its accuracy, its failures, and its evolution over time).

Closed-door evaluations that produce no public reporting fail this criterion even when they produce useful internal information. The governance system itself must be observable to be accountable. The question is not whether any information is withheld (some narrow security exceptions may be appropriate) but whether the overall system is opaque or transparent as a structural matter.

10.3 Durability

Does the governance mechanism survive changes in political leadership, institutional turnover, and political mood shifts? Durability is observable through the mechanism's foundation (whether it rests on durable democratic authorization or on revocable executive arrangements), its track record across political transitions, the degree of effort required to dissolve or redirect it, and whether its continued operation is structurally guaranteed or dependent on the continued active support of current leadership.

The concrete test for durability is: would this governance mechanism continue functioning if the next political transition brought leadership ideologically hostile to its mission? If the answer is no, the mechanism lacks durability regardless of its current effectiveness. Governance that depends on the goodwill of current leadership is not governance; it is toleration.

10.4 Accountability

Are there defined consequences when governance failures occur? Can decisions be challenged by affected parties? Are there feedback loops that surface failures rather than hiding them? Accountability is observable through: the existence of defined standards against which governance performance can be measured; mechanisms available to affected parties for contesting governance decisions; mandatory reporting of incidents and failures; visible consequences (institutional changes, criteria revision, leadership accountability) when governance demonstrably fails; and the distinction between activity metrics (evaluations conducted, meetings held) and outcome metrics (harms prevented, risks identified before realization, accuracy of assessments).

Governance arrangements that report activity without evidence of outcomes, or that have no mechanism for detecting their own failures, or that produce no visible consequences when failures occur, lack accountability regardless of how busy they appear.

10.5 Authority

Does the governance entity have actual power to compel compliance, restrict deployment, or impose costs for non-cooperation? Authority is observable through the entity's capacity to function over the objection of the entities being governed, the availability of consequences for non-compliance, and the enforceability of evaluative conclusions (whether unfavorable evaluations actually result in deployment restrictions or other protective actions).

Pure information-gathering without enforcement capacity is monitoring, not governance. Monitoring is valuable as a sensing function but cannot substitute for the response function, which requires authority. Governance arrangements that describe themselves as "facilitating" or "informing" without any mechanism for compelling action when facilitation and information fail to produce adequate behavior are exercising influence, not authority.

10.6 Scope Adequacy

Does the governance mandate cover the actual risk surface, or only a politically convenient subset? Scope adequacy is observable through: the breadth of risk categories addressed (not just national security but also economic disruption, democratic process integrity, civil liberties, environmental impacts, and systemic cascading risks); the range of actors covered (not just the largest developers but any entity deploying systems at relevant capability levels); and the range of deployment contexts addressed (not just military or intelligence applications but also commercial, consumer, and public-sector uses).

Governance frameworks focused exclusively on national security risks may be politically expedient but are categorically incomplete. The full risk surface of advanced AI extends far beyond national security, and governance that addresses only this subset while leaving other risk categories unmonitored creates a false sense of adequacy.

10.7 Application of the Criteria

Any governance arrangement can be evaluated against these six criteria. Arrangements failing multiple criteria should be identified as such rather than counted as governance progress. The framework does not demand perfection on all criteria simultaneously, but it does demand honest assessment of where arrangements fall short and explicit acknowledgment that shortfalls constitute real governance gaps rather than acceptable compromises.

The criteria also function as a warning system: when governance arrangements that score poorly on these criteria are presented as adequate, the performative-adaptivity failure mode is likely in operation. Political energy directed toward celebrating inadequate arrangements is political energy not directed toward building real institutional capacity.

11. Operational Mechanisms Towards Genuine Resilience

The philosophy and principles articulated in the preceding sections generate specific operational requirements for implementation. This section addresses the mechanisms through which adaptive governance transitions from architectural design to functioning protective capacity: the threshold systems, monitoring programs, risk assessment methodologies, and pre-authorized response mechanisms that constitute the operational layer of governance flow.

11.1 Flexible Threshold Systems

Any governance framework requiring response to changing conditions must operationalize the triggers that initiate that response. This requires an ensemble of thresholds across multiple dimensions of AI risk, e.g. capability thresholds, deployment-scale thresholds, autonomy thresholds, and integration-criticality thresholds, undergoing continuous monitoring. These thresholds must be designed as *systems* rather than isolated indicators: multiple metrics contributing to the same overall assessment, such that if any single metric is triggered, the governance system responds with at minimum enhanced sensing and evaluation of the relevant capability or deployment. The threshold system must be continuously updated as understanding develops, as capabilities advance, and as the relationship between measurable indicators and underlying risk changes. A threshold calibrated to

today's understanding of dangerous capability may be inadequate within months as algorithmic improvements shift the capability frontier.

Crucially, threshold systems must account for the gap between easily measurable indicators and actual risk. A system that monitors only training compute, for example, misses risks from inference-time scaling, fine-tuning of open-weight models, multi-agent composition, and novel elicitation techniques. The threshold ensemble must therefore include heterogeneous indicators: compute-based metrics, capability-evaluation-based metrics, deployment-pattern-based metrics, incident-rate-based metrics, and structural indicators (degree of autonomy, degree of integration into critical systems, degree of human oversight in operation). No single indicator is sufficient; the ensemble provides robustness through diversity of measurement approaches.

11.2 Continuous Monitoring Programs

The sensing function (Section 4.1) requires institutional form as continuous monitoring programs overseeing changes to boundary conditions that could signal potential dangerous phase transitions. These programs must operate at temporal resolution matching the dynamics of what they observe, which, given the pace of AI capability development, means weeks rather than years for frontier capabilities and near-real-time for deployment patterns and incident occurrence. Monitoring programs should be designed for adversarial conditions: the entities being monitored may have incentives to obscure dangerous changes, and AI systems at sufficient capability levels may model monitoring mechanisms and adjust behavior to avoid triggering thresholds while maintaining dangerous capabilities[51] (sandbagging, strategic underperformance on evaluations).

11.3 Pre-Authorized Adaptive Response

The response function (Section 4.3) must include mechanisms for adaptive policy adjustment that activate without requiring full deliberative processes at the moment of crisis. Pre-authorized protective measures, analogous to the circuit breakers discussed in Sections 4.3 and 7.1, should be established through democratic process with defined triggers, defined responses, defined durations, and defined review mechanisms. These pre-authorized responses provide governance with temporal resolution matching threat dynamics while maintaining structural legitimacy through prior democratic authorization. The design of pre-authorized response mechanisms requires careful attention to trigger specificity (vague triggers

produce inappropriate activation), response proportionality (excessive response produces backlash and legitimacy erosion), duration limits (indefinite emergency measures become permanent without accountability), and mandatory review following activation (every triggered response generates a learning obligation).

11.4 Predictive Analysis and Risk Assessment

Adaptive governance cannot be merely reactive; it must maintain prospective capacity, attempting to remain ahead of future risks rather than simply responding to realized harms. Probabilistic risk assessment methodologies adapted for AI systems provide structured approaches to identifying and quantifying risks before they manifest.[52] Scenario development exercises generate the institutional imagination necessary to prepare for counterintuitive or unprecedented risk configurations.[53] Adversarial red-teaming, both of AI systems themselves and of governance mechanisms, surfaces vulnerabilities that standard assessment might miss. Wargaming exercises, increasingly augmented by AI tools themselves, allow planners, developers, and regulators to rehearse policy responses to novel threat scenarios and identify gaps in institutional capacity before crises reveal them.[54] Together, these prospective methodologies constitute the forward-looking component of the sensing function, providing the early warning that enables pre-authorized response rather than post-hoc reaction.

11.5 Critical Transitions and the Governance Readiness Problem

Institutions are generally designed with stability as a baseline assumption, making them structurally unprepared for the phase transitions characteristic of complex adaptive systems. Critical transitions, where the accumulation of changes over time reaches a threshold causing rapid shift into an entirely new regime (analogous to economic collapse, rapid political reconfiguration, or ecological regime shift), are a common feature of the complex systems described in Section 2. It is likely that the rapid acceleration of AI capability and its saturation across industries, economies, and governance-critical functions will produce such a transition, if it has not already begun. The governance architecture described in this paper, if implemented through appropriate institutional design, provides the foundation for governance systems capable of maintaining protective capacity *through* such transitions rather than collapsing at the moment they are most needed. The key design property is that governance must be prepared for discontinuous change, not merely incremental adaptation, and must possess the structural resilience (redundancy,

modularity, antifragility) to maintain function even when the environment in which it operates undergoes rapid phase transition.

12. Conclusion

This paper has articulated a positive philosophy of governance for advanced AI: governance as continuous flow rather than discrete regulatory acts, anchored in structural principles derived from complexity science, institutional design theory, and democratic governance. The philosophy integrates five operational principles (collectivity, adaptability, modularity, redundancy, antifragility) with a four-function governance flow (sensing, evaluation, response, learning), a layered defensive architecture spanning technical through international domains, structural legitimacy mechanisms enabling rapid governance action within democratic constraints, and diagnostic criteria for distinguishing genuine adaptive governance from its performative simulation.

The central claim is that genuine resilience against the risks of advanced AI requires both stable institutional architecture and adaptive operational mechanisms, and these are complements rather than substitutes. Stable architecture provides durability, legitimacy, and authority, the properties that make governance binding across political cycles and resistant to capture. Adaptive mechanisms provide responsiveness, learning capacity, and temporal resolution matching the dynamics of the governed system, the properties that prevent governance from becoming obsolete as the technology it governs evolves. Neither is sufficient without the other. Static architecture without adaptive mechanisms becomes a monument to yesterday's risk landscape. Adaptive mechanisms without stable architecture become government by whim, vulnerable to capture, dissolution, and political convenience regardless of protective need.

The challenge for institutional designers, legislators, and international negotiators is constructing frameworks that incorporate adaptivity from the ground up: frameworks designed to evolve with the technology they govern, pre-authorizing rapid protective response while maintaining democratic legitimacy through structural protections, procedural constraints, and multiple accountability pathways. This is achievable — financial market governance, aviation safety, nuclear safety, and other domains demonstrate that governance can operate at the speed of fast-moving systems while maintaining democratic foundations — but it requires deliberate design informed by the structural properties identified in this paper.

The diagnostic criteria of Section 10 provide the evaluative standard: any governance arrangement, whether proposed or already operational, can be assessed against the criteria of independence, transparency, durability, accountability, authority, and scope adequacy. Arrangements failing multiple criteria should be identified as such and treated as governance gaps requiring remedy rather than celebrated as progress. The persistence of performative arrangements that score poorly on these criteria while consuming the political space genuine governance would occupy is perhaps the most insidious current failure mode, more dangerous in some respects than acknowledged absence of governance, because it provides political cover for inaction while creating the illusion that protection exists.

Only by understanding today's challenges as dynamical (i.e. evolving, nonlinear, emergent, and operating on timescales increasingly compressed beyond human institutional defaults) and by building governance architectures whose own dynamics match those of what they govern, can societies maintain the possibility of effective protective capacity. As AI systems increasingly surpass our social and political institutions in capability, speed, and bandwidth, governance must be reimagined not as a set of static constraints imposed at a point in time but as a living system: sensing, evaluating, responding, and learning in continuous flow, anchored in democratic legitimacy, protected by structural independence, and made resilient through deliberate redundancy across multiple defensive layers.

This philosophy does not guarantee safety. No governance framework can guarantee safety against a technology whose future capabilities are genuinely uncertain and whose development trajectory may involve discontinuous phase transitions. What this philosophy provides is a framework for maintaining the *possibility* of safety under conditions of radical uncertainty and rapid change, and a standard against which all governance proposals, existing arrangements, and institutional developments can be honestly evaluated. The gap between current arrangements and the architecture this philosophy describes measures the distance between current protective capacity and what the challenge demands. Closing that distance is the urgent task.

References

Allen, Craig R., and Lance H. Gunderson. "Pathology and Failure in the Design and Implementation of Adaptive Management." *Journal of Environmental Management* 92, no. 5 (2011).

Anderljung, Markus, Joslyn Barnhart, Anton Korber, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. "Frontier AI Regulation: Managing Emerging Risks to Public Safety." ArXiv abs/2307.03718 (2023).

Ayres, Ian, and John Braithwaite. *Responsive Regulation: Transcending the Deregulation Debate*. New York: Oxford University Press, 1992.

Ball, Dean. "Decentralized Training and the Fall of Compute Thresholds." Hyperdimensional Blog, October 10, 2024. https://www.hyperdimensional.co/p/decentralized-training-and-the-fall.

Bostrom, Nick. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9, no. 1 (2002).

Bradford, Anu. *The Brussels Effect: How the European Union Rules the World*. New York: Oxford University Press, 2020.

Brunner, Ronald D., and Amanda H. Lynch. "Adaptive Governance." *Oxford Research Encyclopedia of Climate Science*, October 26, 2017. https://oxfordre.com/climatescience/view/10.1093/acrefore/9780190228620.001.0001/acrefore-9780190228620-e-601.

Buocz, Thomas, Sebastian Pfotenhauer, and Iris Eisenberger. "Regulatory Sandboxes in the AI Act: Reconciling Innovation and Safety?" *Law, Innovation and Technology* 15, no. 2 (2023): 357–89. https://doi.org/10.1080/17579961.2023.2245678.

Busenberg, George J. "Learning in Organizations and Public Policy." *Journal of Public Policy* 21, no. 2 (2001): 173–189.

Center for AI Risk Management & Alignment. "Improving National Resilience Against AI Incidents: A Global Perspective." Interim policy brief, June 28, 2025. https://carma.org/research-highlights/f/national-preparedness-in-the-age-of-ai.

Center for AI Risk Management & Alignment. "Probabilistic Risk Assessment for AI." PRA Framework. Accessed October 14, 2024. https://pra-for-ai.github.io/pra/index.html.

Corrigan, Jack, and Owen Daniels. "Don't Reinvent the Wheel to Govern AI." Council on Foreign Relations, August 20, 2024. https://www.cfr.org/blog/dont-reinvent-wheel-govern-ai.

Corsi, Giulio, Kyle Kilian, and Richard Mallah. "Considerations Influencing Offense-Defense Dynamics From Artificial Intelligence." ArXiv abs/2412.04029 (2024).

Cox, Michael, Gwen Arnold, and Sergio Villamayor Tomás. "A Review of Design Principles for Community-Based Natural Resource Management." *Ecology and Society* 15, no. 4 (2010): 38.

Dewey, John. *The Public and Its Problems*. New York: Holt and Company, 1927.

Easley, David A., Marcos M. López de Prado, and Maureen O'Hara. "The Microstructure of the 'Flash Crash': Flow Toxicity, Liquidity Crashes, and the Probability of Informed Trading." *The Journal of Portfolio Management* 37 (2011): 118–128.

Edelman, Lauren B. "Legal Ambiguity and Symbolic Structures: Organizational Mediation of Civil Rights Law." *American Journal of Sociology* 97, no. 6 (1992): 1531–1576.

Engler, Alex. "The EU and U.S. Diverge on AI Regulation: A Transatlantic Comparison and Steps to Alignment." Brookings Institution, 2023.

European Parliamentary Research Service. "Artificial Intelligence: Challenges for EU Citizens and Consumers." European Parliament, January 2022. https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf.

Executive Order 14148. "Initial Rescissions of Harmful Executive Orders and Actions." 90 Fed. Reg. 8237 (January 28, 2025).

Executive Order 14179. "Removing Barriers to American Leadership in Artificial Intelligence." 90 Fed. Reg. 8741 (January 31, 2025).

Federal Trade Commission. *Combating Online Harms Through Innovation: A Report to Congress*. Washington, DC: Federal Trade Commission, June 2022. https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting-Online-Harms-Through-Innovation.pdf.

Financial Crisis Inquiry Commission. *The Financial Crisis Inquiry Report*. Washington, DC: U.S. Government Printing Office, 2011.

Folke, Carl, Thomas Hahn, Per Olsson, and Jon Norberg. "Adaptive Governance of Social-Ecological Systems." *Annual Review of Environment and Resources* 30 (2005): 441–473.

Frank, Michael. "Managing Existential Risk from AI without Undercutting Innovation." Center for Strategic and International Studies, July 10, 2023. https://www.csis.org/analysis/managing-existential-risk-ai-without-undercutting-innovation.

Frischmann, Brett M., Michael J. Madison, and Katherine J. Strandburg, eds. *Governing Knowledge Commons*. Oxford: Oxford University Press, 2014.

Future of Life Institute. "Global Governance of AI." 2025. https://global-governance.ai/.

General Services Administration. Administrative Notice: Removal of Anthropic from USAi.gov and Multiple Award Schedule. February 27, 2026.

Gerlich, Michael. "Brace for Impact: Facing the AI Revolution and Geopolitical Shifts in a Future Societal Scenario for 2025–2040." *Societies* 14, no. 9 (2024): 180.

Gunderson, Lance H., and C. S. Holling, eds. *Panarchy: Understanding Transformations in Human and Natural Systems*. Washington, DC: Island Press, 2002.

Habermas, Jürgen. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Translated by William Rehg. Cambridge, MA: MIT Press, 1996.

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Tamay Besiroglu, and Robert Trager. "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation." ArXiv abs/2403.08501 (2024).

Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. "Algorithmic Progress in Language Models." In *Advances in Neural Information Processing Systems 37* (NeurIPS 2024). ArXiv abs/2403.05812 (2024).

Hodge, Neil. "Pace of Innovation Will Make EU AI Act Hard to Enforce, Experts Say." *Compliance Week*, October 17, 2024. https://www.complianceweek.com/regulatory-policy/pace-of-innovation-will-make-eu-ai-act-hard-to-enforce-experts-say/35508.article.

Holling, C. S. "Resilience and Stability of Ecological Systems." *Annual Review of Ecology and Systematics* 4 (1973): 1–23. https://doi.org/10.1146/annurev.es.04.110173.000245.

International Nuclear Safety Advisory Group. *Defence in Depth in Nuclear Safety*. INSAG-10. Vienna: International Atomic Energy Agency, 1996.

Jensen, Ben, Yasir Atalan, and Dan Tadross. "It Is Time to Democratize Wargaming Using Generative AI." Center for Strategic and International Studies, February 22, 2024. https://www.csis.org/analysis/it-time-democratize-wargaming-using-generative-ai.

Kilian, Kyle A., Christopher J. Ventura, and Mark M. Bailey. "Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control." *Futures* 151 (2023). https://www.sciencedirect.com/science/article/abs/pii/S0016328723000861.

Leibo, Joel Z., Edward Hughes, Marc Lanctot, and Thore Graepel. "Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research." ArXiv abs/1903.00742 (2019).

Lorenz, Edward N. "Deterministic Nonperiodic Flow." *Journal of Atmospheric Sciences* 20, no. 2 (1963): 130–141.

Maas, Matthijs M. "Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions." SSRN Electronic Journal (2023).

McKelvey, Fenwick Robert, Jeremy Packer, and Joshua Reeves. "AI and the Automation of Warfare." *Canadian Journal of Communication* 47, no. 2 (2022): 377–398.

Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. "Frontier Models are Capable of In-Context Scheming." Apollo Research. ArXiv abs/2412.04984 (2024).

Meyer, John W., and Brian Rowan. "Institutionalized Organizations: Formal Structure as Myth and Ceremony." *American Journal of Sociology* 83, no. 2 (1977): 340–363.

Mitchell, Melanie. *Complexity: A Guided Tour*. New York, NY: Oxford University Press, 2009. https://doi.org/10.1093/oso/9780195124415.001.0001.

National Commission on Terrorist Attacks upon the United States. *The 9/11 Commission Report*. New York: W.W. Norton, 2004.

National Intelligence Council. "Global Trends 2040: A More Contested World." Office of the Director of National Intelligence, March 2021. https://www.dni.gov/index.php/gt2040-home/summary.

Noma Security. "Shadow AI Agents: Why Untracked AI Is the New Shadow IT." December 5, 2025. https://noma.security/resources/shadow-ai-agents-enterprise-risk/.

Ostrom, Elinor. "Beyond Markets and States: Polycentric Governance of Complex Economic Systems." *American Economic Review* 100, no. 3 (2010): 641–672.

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

Ostrom, Elinor. *Understanding Institutional Diversity*. Princeton: Princeton University Press, 2005.

Perrow, Charles. *Normal Accidents: Living with High-Risk Technologies*. Updated ed. Princeton: Princeton University Press, 1999.

Pilz, Konstantin F., Lennart Heim, and Nicholas Brown. "Increased Compute Efficiency and the Diffusion of AI Capabilities." *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (2025): 27582.

Posner, Richard A. *Catastrophe: Risk and Response*. New York: Oxford University Press, 2004.

Potts, Jason. "Governing the Innovation Commons." *Journal of Institutional Economics* 14, no. 6 (2018): 1025–1047.

Power, A., Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets." ArXiv abs/2201.02177 (2022).

Quinn, Kelsey. "How the U.S. Should Regulate AI After the End of Chevron Deference." New Lines Institute, July 11, 2024. https://newlinesinstitute.org/future-frontiers/how-the-u-s-should-regulate-artificial-intelligence-after-the-chevron-ruling/.

Reason, James. "Human Error: Models and Management." *BMJ* 320, no. 7237 (2000): 768–770. https://doi.org/10.1136/bmj.320.7237.768.

Reuel, Anka, and Trond Arne Undheim. "Generative AI Needs Adaptive Governance." ArXiv abs/2406.04554 (2024).

Revesz, Richard L. "Rehabilitating Interstate Competition: Rethinking the 'Race-to-the-Bottom' Rationale for Federal Environmental Regulation." *New York University Law Review* 67, no. 6 (1992): 1210–1254.

Sandbrink, Jonas, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg. "Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks." SSRN Electronic Journal (2022). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213670.

Santa Fe Institute. "SFI Working Group Summary: Exploring Universal Patterns in the Emergence of Bureaucratic Organizational Structures." Santa Fe, NM: Santa Fe Institute, November 30, 2018.

https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/resource_link_files/SFI_Working_Grp_Summary_V_November_30_2018_db5975.pdf.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, Allan Dafoe, and Jess Whittlestone. "Computing Power and the Governance of Artificial Intelligence." ArXiv abs/2402.08797 (2024).

Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Khoa Sherstan. "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives." ArXiv abs/2311.09227 (2023).

Solaiman, Irene. "The Gradient of Generative AI Release: Methods and Considerations." ArXiv abs/2302.04844 (2023).

Suk, Jonathan E., Kristie L. Ebi, David Vose, William Wint, Neil Alexander, Koen Mintiens, and Jan C. Semenza. "Indicators for Tracking European Vulnerabilities to the Risks of Infectious Disease Transmission due to Climate Change." *International Journal of Environmental Research and Public Health* 11 (2014): 2218–2235.

Sunstein, Cass R. "Nondelegation Canons." *University of Chicago Law Review* 67, no. 2 (2000): 315–343.

Taleb, Nassim Nicholas. *Antifragile: Things That Gain from Disorder*. New York: Random House, 2012.

Taleb, Nassim Nicholas. *Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications*. STEM Academic Press, 2020.

Thierer, Adam. "Governing Emerging Technology in an Age of Policy Fragmentation and Disequilibrium." The American Enterprise Institute, April 2022. https://platforms.aei.org/wp-content/uploads/2022/04/RPT_Governing-Emerging-Tech_Thierer-2022.pdf.

United Nations, Office of Disarmament Affairs. "Treaty on the Prohibition of Nuclear Weapons." Accessed October 14, 2024. https://disarmament.unoda.org/wmd/nuclear/tpnw/.

U.S. Securities and Exchange Commission. Order Approving Proposed Rule Changes Relating to Market-Wide Circuit Breakers. Release No. 67090 (May 31, 2012). 77 Fed. Reg. 33531 (June 6, 2012).

van der Weij, Teun, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. "AI Sandbagging: Language Models can Strategically Underperform on Evaluations." In *Proceedings of the International Conference on Learning Representations* (ICLR 2025). ArXiv abs/2406.07358 (2025).

Wisakanto, Anna K., Joe Rogero, Avyay M. Casheekar, and Richard Mallah. "Adapting Probabilistic Risk Assessment for AI." ArXiv abs/2504.18536 (2025). https://doi.org/10.48550/arXiv.2504.18536.

Wu, Qi-hui, Guoru Ding, Yuhua Xu, Shuo Feng, Zhiyong Du, Jinlong Wang, and Keping Long. "Cognitive Internet of Things: A New Paradigm Beyond Connection." *IEEE Internet of Things Journal* 1 (2014): 129–143.

Yang, Xianjun, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models." ArXiv abs/2310.02949 (2023).

Zelikman, E., Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. "Self-Taught Optimizer (STOP): Recursively Self-Improving Code Generation." ArXiv abs/2310.02304 (2023).

Endnotes

1. Michael Gerlich, "Brace for Impact: Facing the AI Revolution and Geopolitical Shifts in a Future Societal Scenario for 2025–2040," *Societies* 14, no. 9 (2024): 180.
2. National Intelligence Council, "Global Trends 2040: A More Contested World" (Washington, DC: Office of the Director of National Intelligence, 2021).
3. Neil Hodge, "Pace of Innovation Will Make EU AI Act Hard to Enforce, Experts Say," *Compliance Week*, October 17, 2024.
4. Noma Security. "Shadow AI Agents: Why Untracked AI Is the New Shadow IT." December 5, 2025. https://noma.security/resources/shadow-ai-agents-enterprise-risk/.
5. Melanie Mitchell, *Complexity: A Guided Tour* (New York: Oxford University Press, 2009), 6–8.
6. Edward N. Lorenz, "Deterministic Nonperiodic Flow," *Journal of Atmospheric Sciences* 20, no. 2 (1963): 130–141.
7. A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets," arXiv:2201.02177 (2022).
8. Joel Z. Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel, "Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research," arXiv:1903.00742 (2019).
9. David A. Easley, Marcos M. López de Prado, and Maureen O'Hara, "The Microstructure of the 'Flash Crash': Flow Toxicity, Liquidity Crashes, and the Probability of Informed Trading," *The Journal of Portfolio Management* 37 (2011): 118–128.
10. Fenwick Robert McKelvey, Jeremy Packer, and Joshua Reeves, "AI and the Automation of Warfare," *Canadian Journal of Communication* 47, no. 2 (2022): 377–398.
11. E. Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai, "Self-Taught Optimizer (STOP): Recursively Self-Improving Code Generation," arXiv:2310.02304 (2023).
12. This point distinguishes the application of CAS theory to AI governance from its application in ecological or economic governance contexts. See Kyle A. Kilian, Christopher J. Ventura, and Mark M. Bailey, "Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control," *Futures* 151 (2023), for extended discussion of recursive self-improvement and the question of control.

13. The inference-time scaling paradigm is exemplified by chain-of-thought reasoning, self-verification loops, and extended deliberation architectures that allow models to exhibit substantially greater capability when allocated additional compute at inference rather than training time.
14. Jonathan E. Suk et al., "Indicators for Tracking European Vulnerabilities to the Risks of Infectious Disease Transmission due to Climate Change," *International Journal of Environmental Research and Public Health* 11 (2014): 2218–2235.
15. National Intelligence Council, "Global Trends 2040: A More Contested World" (Washington, DC: Office of the Director of National Intelligence, March 2021).
16. Anka Reuel and Trond Arne Undheim, "Generative AI Needs Adaptive Governance," arXiv:2406.04554 (2024).
17. United Nations, Office of Disarmament Affairs, "Treaty on the Prohibition of Nuclear Weapons," accessed October 14, 2024.
18. Matthijs M. Maas, "Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions," SSRN Electronic Journal (2023).
19. Dean Ball, "Decentralized Training and the Fall of Compute Thresholds," Hyperdimensional Blog, October 10, 2024.
20. For comprehensive treatment of compute governance mechanisms and their application to AI oversight, see Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, Allan Dafoe, and Jess Whittlestone, "Computing Power and the Governance of Artificial Intelligence," arXiv:2402.08797 (2024). For analysis of how compute providers function as governance intermediaries expanding the enforcement option space beyond behavioral and self-reporting mechanisms, see Lennart Heim, Tim Fist, Janet Egan, Sihao Huang, Tamay Besiroglu, and Robert Trager, "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation," arXiv:2403.08501 (2024).
21. The concept of performative adaptivity draws on related observations in organizational sociology and legal theory. See Lauren B. Edelman, "Legal Ambiguity and Symbolic Structures: Organizational Mediation of Civil Rights Law," *American Journal of Sociology* 97, no. 6 (1992): 1531–1576. See also John W. Meyer and Brian Rowan, "Institutionalized Organizations: Formal Structure as Myth and Ceremony," *American Journal of Sociology* 83, no. 2 (1977): 340–363. The specific application to AI governance is, to our knowledge, novel.
22. Executive Order 14148, "Initial Rescissions of Harmful Executive Orders and Actions," 90 Fed. Reg. 8237 (January 28, 2025), revoking Executive Order

14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (October 30, 2023). See also General Services Administration, Administrative Notice: Removal of Anthropic from USAi.gov and Multiple Award Schedule (February 27, 2026) (removing Anthropic AI products from federal access following the company's refusal to remove restrictions on the Department of Defense's use of its AI technology).

23. John Dewey, *The Public and Its Problems* (New York: Holt and Company, 1927). As discussed in George J. Busenberg, "Learning in Organizations and Public Policy," *Journal of Public Policy* 21, no. 2 (2001): 173–189.
24. Ronald D. Brunner and Amanda H. Lynch, "Adaptive Governance," *Oxford Research Encyclopedia of Climate Science*, October 26, 2017.
25. Craig R. Allen and Lance H. Gunderson, "Pathology and Failure in the Design and Implementation of Adaptive Management," *Journal of Environmental Management* 92, no. 5 (2011).
26. Santa Fe Institute, "SFI Working Group Summary: Exploring Universal Patterns in the Emergence of Bureaucratic Organizational Structures" (Santa Fe, NM: Santa Fe Institute, November 30, 2018).
27. Ian Ayres and John Braithwaite, *Responsive Regulation: Transcending the Deregulation Debate* (New York: Oxford University Press, 1992). See also Carl Folke, Thomas Hahn, Per Olsson, and Jon Norberg, "Adaptive Governance of Social-Ecological Systems," *Annual Review of Environment and Resources* 30 (2005): 441–473.
28. Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward, "AI Sandbagging: Language Models can Strategically Underperform on Evaluations," in *Proceedings of the International Conference on Learning Representations* (ICLR 2025), arXiv:2406.07358 (2025). See also Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn, "Frontier Models are Capable of In-Context Scheming," Apollo Research, arXiv:2412.04984 (2024). The implications of these findings for monitoring and evaluation system design are developed in Section 11.
29. NYSE Rule 7.12 (Market-Wide Circuit Breakers), successor to Rule 80B (adopted 1988); statutory authority under Securities Exchange Act of 1934, § 12(k), 15 U.S.C. § 78l(k). Current thresholds: Level 1 (7% S&P 500 decline), Level 2 (13%), Level 3 (20%), triggering automatic 15-minute trading halts (Levels 1–2) or remainder-of-day halt (Level 3). See SEC Release No. 67090 (June 1, 2012).
30. European Parliamentary Research Service, "Artificial Intelligence: Challenges for EU Citizens and Consumers" (European Parliament, January

2022); Thomas Buocz, Sebastian Pfotenhauer, and Iris Eisenberger, "Regulatory Sandboxes in the AI Act: Reconciling Innovation and Safety?" *Law, Innovation and Technology* 15, no. 2 (2023): 357–389.

31. Buocz, Pfotenhauer, and Eisenberger, "Regulatory Sandboxes in the AI Act" (2023).
32. C. S. Holling, "Resilience and Stability of Ecological Systems," *Annual Review of Ecology and Systematics* 4 (1973): 1–23. See also Lance H. Gunderson and C. S. Holling, eds., *Panarchy: Understanding Transformations in Human and Natural Systems* (Washington, DC: Island Press, 2002).
33. The failure of redundant oversight in financial regulation is extensively documented in Financial Crisis Inquiry Commission, *The Financial Crisis Inquiry Report* (Washington, DC: U.S. Government Printing Office, 2011), particularly Chapter 2 ("The Shadow Banking System") and Chapter 3 ("Securitization and the Failure of Regulation"), documenting how multiple agencies with overlapping mandates each assumed others were monitoring risks that none adequately covered. For the intelligence community coordination failures, see National Commission on Terrorist Attacks upon the United States, *The 9/11 Commission Report* (New York: W.W. Norton, 2004), Chapter 11 ("Foresight—and Hindsight"), 339–360. For theoretical treatment of how redundancy in tightly coupled systems produces interaction failures rather than protective overlap when coordination mechanisms are absent, see Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, updated ed. (Princeton: Princeton University Press, 1999).
34. Nassim Nicholas Taleb, *Antifragile: Things That Gain from Disorder* (New York: Random House, 2012).
35. Nassim Nicholas Taleb, *Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications* (STEM Academic Press, 2020). See also Richard A. Posner, *Catastrophe: Risk and Response* (New York: Oxford University Press, 2004).
36. James Reason, "Human Error: Models and Management," *BMJ* 320, no. 7237 (2000): 768–770. See also International Nuclear Safety Advisory Group, *Defence in Depth in Nuclear Safety*, INSAG-10 (Vienna: International Atomic Energy Agency, 1996), STI/PUB/1013.
37. Jürgen Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, trans. William Rehg (Cambridge, MA: MIT Press, 1996). For the specific challenges of legitimacy in delegated administrative authority, see Cass R. Sunstein, "Nondelegation Canons," *University of Chicago Law Review* 67, no. 2 (2000): 315–343.

38. The breadth of existing statutory authorities applicable to AI-specific harms is surveyed in Alex Engler, "The EU and U.S. Diverge on AI Regulation: A Transatlantic Comparison and Steps to Alignment" (Brookings Institution, 2023). For consumer protection authority specifically, the Federal Trade Commission's Section 5 authority over "unfair or deceptive acts or practices" (15 U.S.C. § 45) has been applied to algorithmic harms without requiring new AI-specific legislation; see Federal Trade Commission, *Combating Online Harms Through Innovation: A Report to Congress* (Washington, DC: Federal Trade Commission, June 2022). For securities authority, the SEC's existing mandate over market manipulation (Securities Exchange Act § 9(a)(2), 15 U.S.C. § 78i(a)(2)) encompasses AI-driven manipulation without new authorization. The Food and Drug Administration's existing authority over medical devices (21 U.S.C. § 360c) encompasses AI-enabled diagnostic and treatment tools. The critical observation is that agencies possess substantially more authority than they currently exercise with respect to AI-specific manifestations of harms already within their statutory jurisdiction, and that the primary barriers to action are institutional capacity, technical expertise, and political will rather than legal authority. For the argument that the post-*Loper Bright* environment complicates but does not eliminate agency authority to address novel technological harms within existing mandates, see Kelsey Quinn, "How the U.S. Should Regulate AI After the End of Chevron Deference," New Lines Institute, July 11, 2024.
39. Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology* 9, no. 1 (2002), § 9.4. For its formal elaboration as a governance principle, see Jonas Sandbrink, Hamish Hobbs, Jacob Swett, Allan Dafoe, and Anders Sandberg, "Differential Technology Development: A Responsible Innovation Principle for Navigating Technology Risks," SSRN Electronic Journal (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213670.
40. Giulio Corsi, Kyle A. Kilian, and Richard Mallah, "Considerations Influencing Offense-Defense Dynamics From Artificial Intelligence," arXiv:2412.04029 (2024). This work establishes a taxonomy of factors (including resource asymmetry between attacker and defender, reversibility of enabled harms, scalability differentials between offensive and defensive applications, and the structural effects of capability diffusion on the balance of advantage) that provide systematic analytical foundations for governance decisions about differential support. The taxonomy demonstrates that offense-defense assessment, while irreducibly uncertain for dual-use capabilities, can be

disciplined through structured analysis of identifiable structural properties rather than relying on intuitive or politically motivated categorization.

41. Future of Life Institute. "Global Governance of AI." 2025. https://global-governance.ai/.
42. Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla, "Algorithmic Progress in Language Models," in *Advances in Neural Information Processing Systems 37* (NeurIPS 2024), arXiv:2403.05812 (2024). See also Konstantin F. Pilz, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (2025): 27582.
43. As of this writing, frontier AI development capacity (defined by training compute exceeding approximately 10^25 FLOP) is concentrated substantially in the United States, with significant capacity in China, the United Kingdom, the United Arab Emirates, and France, and emerging capacity in several additional jurisdictions. This concentration is sustained primarily by semiconductor supply chain bottlenecks (dominated by a small number of fabrication entities, principally TSMC) and by the large-scale capital requirements of frontier training runs. Both sustaining factors are eroding: hardware efficiency improvements reduce required scale, algorithmic improvements reduce required compute per unit capability (see Ho et al. 2024, cited at endnote 40), capital availability for AI development is broadening globally, and distributed training techniques may reduce dependence on concentrated infrastructure. The governance-feasibility window described here is defined by the period during which this concentration persists in a form amenable to coordinated oversight.
44. Richard L. Revesz, "Rehabilitating Interstate Competition: Rethinking the 'Race-to-the-Bottom' Rationale for Federal Environmental Regulation," *New York University Law Review* 67, no. 6 (1992): 1210–1254. See also Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (New York: Oxford University Press, 2020).
45. The governance challenges specific to open-weight model release are analyzed in Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Khoa Sherstan, "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source

Objectives," arXiv:2311.09227 (2023). For empirical demonstration of how fine-tuning can remove safety guardrails from released weights at minimal cost, see Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin, "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models," arXiv:2310.02949 (2023).

46. The legal and policy architecture for structured release protocols is analyzed in Markus Anderljung, Joslyn Barnhart, Anton Korber, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf, "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv:2307.03718 (2023). For the specific case for graduated access mechanisms as a practical alternative to binary open/closed release decisions, see Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," arXiv:2302.04844 (2023).
47. Center for AI Risk Management & Alignment. "Improving National Resilience Against AI Incidents: A Global Perspective." Interim policy brief, June 28, 2025. https://carma.org/research-highlights/f/national-preparedness-in-the-age-of-ai
48. Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge: Cambridge University Press, 1990). The eight design principles are articulated at pp. 90–102 and empirically validated across diverse institutional contexts including fisheries, irrigation systems, forest management, and groundwater basins. For meta-analysis confirming the design principles' applicability across hundreds of cases and identifying which principles are most robustly associated with institutional success, see Michael Cox, Gwen Arnold, and Sergio Villamayor Tomás, "A Review of Design Principles for Community-Based Natural Resource Management," *Ecology and Society* 15, no. 4 (2010): 38. For application of Ostrom's framework to knowledge and technology governance contexts, see Brett M. Frischmann, Michael J. Madison, and Katherine J. Strandburg, eds., *Governing Knowledge Commons* (Oxford: Oxford University Press, 2014).
49. Ostrom's finding that community self-monitoring can be effective is conditioned on specific structural properties: small community size enabling mutual observation, repeated interaction creating reputation effects, shared dependence on the resource creating aligned incentives, and low monitoring costs relative to community resources. See Ostrom, *Governing the Commons*, 94–96; Elinor Ostrom, *Understanding Institutional Diversity* (Princeton:

Princeton University Press, 2005), Chapter 9. The AI development community satisfies none of these conditions fully: the community is large and growing, interactions are often competitive rather than cooperative, individual developers may benefit commercially from governance evasion even as the collective suffers from the resulting risk externalities, and monitoring AI system capabilities requires substantial and costly technical infrastructure unavailable to most community members. This structural analysis justifies the paper's emphasis on institutionally independent monitoring rather than reliance on developer self-governance.

50. For discussion of how Ostrom's design principles apply under conditions of rapid technological change (where the "resource" itself evolves and the community of relevant actors is unstable), see Jason Potts, "Governing the Innovation Commons," *Journal of Institutional Economics* 14, no. 6 (2018): 1025–1047. Potts argues that innovation commons require more dynamic institutional design than natural resource commons precisely because the governed domain is non-stationary, a property shared in extreme form by AI governance. See also Elinor Ostrom, "Beyond Markets and States: Polycentric Governance of Complex Economic Systems," *American Economic Review* 100, no. 3 (2010): 641–672 (Nobel Prize lecture), which extends the polycentric governance framework to complex economic systems with dynamic properties more analogous to AI ecosystems than the relatively stable natural resource systems studied in her earlier work.
51. The specific capability of AI systems to strategically underperform on evaluations is empirically demonstrated in Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward, "AI Sandbagging: Language Models can Strategically Underperform on Evaluations," in *Proceedings of the International Conference on Learning Representations* (ICLR 2025), arXiv:2406.07358 (2025). The related capability for in-context scheming (pursuing misaligned objectives while appearing aligned during oversight) is documented in Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn, "Frontier Models are Capable of In-Context Scheming," Apollo Research, arXiv:2412.04984 (2024). Together these findings imply that monitoring systems must structurally incorporate adversarial evaluation protocols (e.g., surprise testing outside expected evaluation windows, varied evaluation contexts and framings, cross-model behavioral comparison, and capability elicitation techniques beyond standard benchmarks) as permanent architectural features rather than occasional supplements. See also discussion at endnote 27.

52. Anna K. Wisakanto, Joe Rogero, Avyay M. Casheekar, and Richard Mallah, "Adapting Probabilistic Risk Assessment for AI," arXiv:2504.18536 (2025), https://doi.org/10.48550/arXiv.2504.18536. This paper adapts established PRA techniques from high-reliability industries (nuclear power, aerospace) for advanced AI systems, introducing aspect-oriented hazard analysis, risk pathway modeling, and uncertainty management as methodological advances. It is implemented as a workbook tool for AI developers, evaluators, and regulators, and synthesizes results into a risk report card with aggregated risk estimates. Project website with workbook tool available at https://pra-for-ai.github.io/pra/index.html.
53. Kyle A. Kilian, Christopher J. Ventura, and Mark M. Bailey, "Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control," *Futures* 151 (2023).
54. See Ben Jensen, Yasir Atalan, and Dan Tadross, "It Is Time to Democratize Wargaming Using Generative AI," Center for Strategic and International Studies, February 22, 2024. CARMA provides such a platform, the AI-TTX Wargaming Platform, at https://github.com/CARMA-org/ai-ttx/.