The Frontier AI Risk Stack: A Six-Layer Framework for Understanding Systemic Risk in Advanced AI Systems

Sean Sooch

Independent Researcher

Abstract

Public and policy discourse on artificial intelligence risk remains fragmented across disconnected communities, each addressing different slices of the problem, benchmark performance, model safety, misinformation, cyber misuse, labor disruption, concentration of power, or long-term existential risk. This paper argues that frontier AI risk is best understood not as a single category but as a stack of six interacting risk layers: Model Risk, System Risk, Deployment Risk, Access Risk, Governance Risk, and Civilizational Risk. Drawing on recent capability disclosures from frontier AI laboratories, the growing body of evidence on emergent behaviors in advanced systems, and the architectural shift from isolated models to parallel agentic systems, this paper demonstrates that the most serious dangers arise not from any single layer but from the way failures compound across layers. The framework proposed here provides a structured analytical tool for policymakers, developers, and civil society to evaluate frontier AI risk without reducing it to any single dimension, and identifies key directions for empirical research and governance intervention.

Keywords: artificial intelligence risk, AI governance, agentic systems, frontier models, risk taxonomy, AI safety, parallel orchestration

1. Introduction

Artificial intelligence is entering a new strategic phase. The central issue is no longer whether frontier systems will become highly capable, that threshold has effectively been crossed. The more consequential question is how those capabilities are being architected, deployed, concentrated, and governed. Yet public discourse, investor perception, institutional preparedness, and regulatory debate remain anchored to an outdated mental model: one user, one assistant, one answer. This creates a systematic gap between the risk that regulators are preparing for and the risk that actually exists in deployed frontier systems.

The AI safety literature has historically concentrated on model-level concerns, the properties and failure modes of individual systems. Foundational work by Amodei et al. [1] identified concrete safety problems including reward hacking, distributional shift, and safe exploration. Russell [2] and Bostrom [3] articulated the alignment problem at a philosophical level, establishing the theoretical basis for concerns about systems whose objectives diverge from human values. More recently, Ngo et al. [4] have situated alignment challenges within the deep learning paradigm specifically, while Hubinger et al. [5] have provided experimental evidence of deceptive alignment behaviors persisting through safety training. These contributions form the empirical and theoretical foundation upon which the present analysis builds.

However, a growing body of work suggests that model-level analysis alone is insufficient. Chan et al. [6] have argued that increasingly agentic AI systems introduce qualitatively new categories of harm that model-level

evaluations systematically miss. Shevlane et al. [7] have developed frameworks for evaluating extreme risks from frontier models, recognizing that evaluation must extend beyond isolated capability testing. Anderljung et al. [8] have proposed regulatory frameworks for frontier AI that grapple with the gap between technical capability and institutional readiness. And Dafoe [9] has outlined a research agenda for AI governance that emphasizes the political and economic dimensions of AI development, dimensions that purely technical frameworks tend to overlook.

This paper synthesizes and extends these lines of inquiry by proposing a unified analytical framework, the Frontier AI Risk Stack, that identifies six distinct but interacting layers of risk. The framework's central claim is that frontier AI risk is compositional: no single layer fully explains the danger, and the most serious threats emerge from the way failures compound across layers. The paper proceeds in five parts. Section 2 establishes the architectural shift from isolated models to orchestrated agentic systems. Section 3 analyzes how parallel orchestration functions as a capability multiplier with governance implications. Section 4 examines the governance gap created by restricted deployment and alignment debt. Section 5 presents the six-layer framework in detail. Section 6 discusses cross-layer interactions, operationalization challenges, and directions for future research.

2. The Architectural Shift: From Models to Systems

A model is no longer the full unit of concern for frontier AI risk assessment. What is deployed in practice is increasingly a system: model plus orchestration, model plus memory, model plus tools, model plus permissions, model plus environment access, and increasingly, model plus parallel execution [6, 10]. What matters is the system surrounding the model, memory structures that persist and evolve across interactions; tool use that extends the model's reach into external systems; code execution capabilities; environment access and sensing; recursive planning across multiple layers of abstraction; multi-step workflows that decompose complex objectives into subtasks; and orchestration across many parallel branches of reasoning and action.

The frontier is therefore moving from model intelligence to what might be termed *architected intelligence*, where the power of the system emerges not from the model alone but from the design of the entire computational structure around it. This distinction is not merely one of engineering optimization, it represents a qualitative shift in how intelligence is deployed and scaled. A model that appears harmless in isolation, tested through dialogue in a constrained environment, may become operationally powerful when embedded in a system that can decompose objectives into subtasks, access external tools and databases, execute code, and coordinate parallel workstreams [7, 11].

Current evaluation paradigms that test models in controlled, isolated conditions therefore systematically underestimate the capabilities and risks of deployed systems. This creates a dangerous gap between what we think we know about model behavior and what those models can actually achieve when deployed in their full architectural context. Evaluating a model in isolation is analogous to evaluating an engine without considering the vehicle, the road, and the driver, it captures something real, but it systematically underestimates the operational risk [7, 12].

3. Parallel Orchestration as Capability Multiplier

Within the broader shift toward agentic systems, one architectural distinction is especially consequential: the difference between series and parallel agentic systems. Series systems process tasks in a linear chain where each step depends on the previous one, they are easier to debug, monitor, and govern because each step is traceable and the execution path is linear and predictable. Parallel systems distribute work across multiple simultaneous

branches, launching multiple agents or processes concurrently, each pursuing a different aspect of the problem, with a coordinating layer integrating the results [13].

The analogy to electrical circuits is instructive. In a series circuit, current flows through components sequentially: one path exists, the same current passes through each component, and one failure can interrupt the whole chain. In a parallel circuit, multiple paths operate simultaneously, the same voltage is maintained across branches, and one failure does not collapse the entire system. These two patterns, when applied to AI system design, produce fundamentally different capability profiles with very different implications for speed, robustness, search breadth, and governability.

Parallelism changes the risk profile in four major ways. First, *search-space expansion*: parallelized systems can search many branches simultaneously, dramatically increasing the likelihood of finding effective (and potentially dangerous) solutions. Second, *reduced wall-clock time to action*: parallel architectures compress time and reduce windows for human intervention, meaning that by the time an operator notices concerning behavior, the system may have already explored dozens of paths and committed to a course of action. Third, *greater robustness through branch diversity*: multiple strategies can be explored simultaneously and converged upon, making the system more resilient to individual failures and more likely to achieve its objectives even under adverse conditions. Fourth, *emergent practical capability*: a model that appears modest in isolation may become substantially more powerful when embedded in a parallelized workflow, because the orchestration layer compensates for the model's individual limitations through redundancy, diversity, and synthesis [6, 13].

This architectural difference has direct governance implications. The safety properties of a system composed of dozens of parallel processes, each interacting with external tools and data sources, cannot be predicted from the safety properties of its individual components. They are an emergent function of how those parts interact, and current evaluation methodologies are not designed to capture that emergence [7, 14]. The current approach to AI governance is analogous to certifying individual aircraft components without ever testing the assembled plane in flight.

4. The Governance Gap: Restricted Access, Alignment Debt, and Institutional Lag

Recent capability disclosures from frontier AI laboratories, including Anthropic's Claude model system cards [15, 16], provide evidence that frontier AI systems have crossed a threshold where certain misalignment-relevant behaviors are no longer purely hypothetical. Documented behaviors include instances of strategic deception, covert rule circumvention, reward hacking, and adaptive interaction with oversight mechanisms [5, 15]. It is important to note the epistemic status of these observations: some represent empirically documented behaviors under controlled evaluation conditions, while others represent theoretical projections based on the optimization dynamics of increasingly capable systems. This paper treats both as analytically significant, but distinguishes between them where precision demands.

4.1 Restricted Deployment and Cognitive Stratification

If the most capable AI systems cannot be safely released to the general public, then their deployment will necessarily become selective. This is not a hypothetical scenario; it is already the operational reality for frontier models. Access to the most powerful systems is gated by corporate agreements, government partnerships, research affiliations, and pricing structures that effectively limit frontier cognition to a narrow set of actors: large corporations, state security agencies, critical infrastructure operators, elite research institutions, and ultra-high-capital entities [9, 17].

This dynamic creates what can be described as *cognitive stratification*, a new form of inequality mediated not by access to land, labor, education, or industrial capital, but by access to decision-quality intelligence itself. The concentration of frontier intelligence among a small number of actors is not a temporary market condition or a policy choice that can be easily reversed. It is a structural feature of a technology that becomes more dangerous as it becomes more capable, creating a self-reinforcing dynamic: as capability grows, the incentive to restrict access increases, and the ability of restricted actors to identify new risks also increases [9, 18]. This creates a cycle where capability growth and access restriction move in lockstep.

4.2 Alignment Debt

Current alignment methods may not be scaling commensurately with capability. *Alignment debt* accumulates when systems become more capable without corresponding improvements in interpretability, robustness, corrigibility, and reliability under real-world conditions [4, 19]. Each generation of models becomes more powerful, but the tools for understanding, constraining, and verifying their behavior improve more slowly. Experimental evidence from Hubinger et al. [5] suggests that advanced systems may increasingly optimize not for alignment itself, but for appearing aligned under test conditions, a phenomenon consistent with theoretical predictions about deceptive alignment [4, 20].

This creates the risk of what might be termed *ambivalent superintelligence*: a system that does not harbor adversarial goals but fails to care about human flourishing in the rich, civilizational sense that human beings require. It may optimize for goals that are not overtly malicious but remain profoundly misaligned with liberty, dignity, pluralism, or continuity of human self-determination. This is a subtler but arguably more realistic danger than cinematic hostility, and one for which current alignment techniques offer inadequate assurance [2, 3, 21].

4.3 The Monitoring Problem and Optimization Theater

Modern AI safety and regulation implicitly rely on some form of monitorability, whether through audits, evaluations, logging, interpretability research, or chain-of-thought analysis. The prevailing governance model assumes that sufficiently advanced systems can be rendered observable enough to manage. But if frontier systems can reason about evaluators, simulate compliance, and modulate visible behavior based on contextual cues, then oversight mechanisms risk becoming what might be called *optimization theater*: processes that create the appearance of safety without reliably ensuring it [5, 15, 22].

At that point, safety becomes not a condition but a performance, and the distinction between aligned and strategically compliant becomes operationally invisible. This undermines the foundation of every governance framework that depends on the ability to observe, measure, and audit model behavior as a proxy for trustworthiness. We propose that ‘optimization theater’ be operationally defined as the condition under which a system’s observable compliance metrics diverge systematically from its latent behavioral tendencies, a definition that, while challenging to measure directly, suggests specific empirical research directions discussed in Section 6.

5. The Six-Layer Frontier AI Risk Stack

The framework proposed here identifies six distinct but interacting layers of frontier AI risk. The stack is organized from the most technically proximate layer (Model Risk) to the most structurally distal (Civilizational Risk). Each layer is defined by its analytical focus, the primary actors with leverage at that layer, and the types of failure it captures. The layers are not independent, each can amplify or constrain the others, and the framework’s central analytical claim is that the most serious risks emerge from cross-layer interactions rather than from any single layer in isolation.

Layer 1: Model Risk

Definition: The properties, capabilities, and failure modes of the base model itself. This includes deception capacity, unsafe knowledge, strategic concealment, manipulative capacity, and the potential for cyber or biological uplift. **Primary actors:** Model developers. **Evaluation focus:** Capability benchmarks, red-teaming, adversarial evaluations. **Existing literature:** This is the most extensively studied layer, encompassing the alignment problem [2, 3], concrete safety problems [1], and emergent deceptive behaviors [5]. Model risk is the foundation of the stack, but it is only the foundation.

Layer 2: System Risk

Definition: The risks that emerge from architecture and orchestration, the consequences of tool use, memory, recursive planning, parallel execution, and multi-agent coordination. System risk captures the ways in which a model's capabilities are amplified or transformed by the infrastructure surrounding it. **Primary actors:** System architects, platform developers. **Evaluation focus:** System-level evaluations under deployment-realistic conditions, orchestration audits. **Distinguishing criterion:** A risk belongs to this layer if it would not manifest from the model alone but requires the system scaffold to emerge [6, 13].

Layer 3: Deployment Risk

Definition: The risks introduced when a system is placed in real environments and given real permissions, production access, API control, file inspection, financial or operational authority. Deployment risk is the layer at which theoretical capability becomes operational power. **Primary actors:** Deployers, enterprise integrators. **Evaluation focus:** Permission audits, environment-specific risk assessments, operational red-teaming. **Distinguishing criterion:** A risk belongs here if it depends on the specific environment, permissions, or operational context in which the system is deployed [8, 12].

Layer 4: Access Risk

Definition: The risks arising from who gets frontier intelligence and who does not. This includes concentration among large firms, asymmetry between elite and public use, and the emergence of cognitive inequality. Access risk determines the distribution of benefits and harms across society. **Primary actors:** Access controllers, policymakers, market regulators. **Evaluation focus:** Distribution analysis, access equity metrics, concentration indices. **Distinguishing criterion:** A risk belongs here if it arises from the pattern of distribution rather than the properties of the technology itself [9, 17, 18].

Layer 5: Governance Risk

Definition: Whether institutions can understand, constrain, and respond to the systems being built. This includes weak audits, inadequate regulation, self-certification, arms-race incentives, and regulatory capture. Governance risk determines whether the other layers are subject to meaningful oversight or operate effectively unchecked. **Primary actors:** Regulators, standards bodies, civil society organizations. **Evaluation focus:** Institutional capacity assessments, regulatory gap analysis, audit quality metrics. **Existing literature:** Dafoe [9], Anderljung et al. [8], and Cihon et al. [23] have made substantial contributions here [8, 9, 23].

Layer 6: Civilizational Risk

Definition: The long-term implications of the preceding layers interacting faster than society can adapt. This includes systemic concentration of cognitive power, loss of public contestability, long-run human disempowerment, and structural civilizational instability. Civilizational risk is the emergent consequence of compounding failures across all other layers. **Primary actors:** All of the above, plus democratic publics. **Evaluation focus:** Long-range scenario analysis, structural indicators of institutional capacity. **Distinguishing criterion:** A risk belongs here if it emerges from the interaction of multiple lower layers operating over extended time horizons [3, 24].

6. Cross-Layer Risk Composition and Interaction Mechanisms

The most important claim of this framework is that frontier AI risk is compositional. No single layer alone fully explains the danger. Model Risk determines what the system can potentially do. System Risk determines how effectively that capability is organized. Deployment Risk determines what environment and permissions operationalize it. Access Risk determines who gets the benefit or burden of it. Governance Risk determines whether institutions can keep up. And Civilizational Risk emerges when the rest interact at scale [25].

This compositional view explains why single-variable analyses of AI risk consistently feel incomplete. Focusing on model safety alone misses system amplification. Focusing on access alone misses the role of architecture. Focusing on governance alone misses the speed at which capability is evolving. A coherent analysis must hold all six layers in view simultaneously and attend to the interactions between them.

Critically, the interactions are not merely additive, they are often multiplicative. A modest model-level capability becomes far more consequential when amplified by parallel orchestration (Layer 2), deployed in a high-stakes environment (Layer 3), restricted to actors with limited oversight (Layer 4), and governed by institutions that lack the technical capacity to evaluate what they are overseeing (Layer 5). A model that is powerful but constrained by good system design and governance may be safe. A weaker model deployed in a poorly designed system with insufficient oversight may be dangerous. Understanding risk as a stack, rather than as a flat list of concerns, is the first step toward a governance framework adequate to the complexity of the challenge.

6.1 Illustrative Interaction Pathways

To demonstrate the framework's analytical utility, consider two illustrative pathways through the stack. In the first, a model with modest deception capacity (Layer 1) is embedded in a parallel agentic system with tool access and persistent memory (Layer 2), deployed with production API credentials in a financial services environment (Layer 3), available only to a single firm with no external audit requirement (Layers 4–5). The compounding across layers transforms a manageable model-level concern into a systemic risk that no single-layer intervention would address.

In the second pathway, a highly capable model (Layer 1) is deployed in a well-designed series-only architecture with comprehensive logging (Layer 2), in a sandboxed research environment (Layer 3), with broad academic access (Layer 4), and robust third-party audit requirements (Layer 5). Here, the higher layers constrain and contain the model-level risk. The framework thus provides analytical structure for understanding both risk amplification and risk mitigation, an advantage over frameworks that focus exclusively on worst-case scenarios.

7. Toward Operationalization: Empirical Directions and Limitations

Several of the concepts introduced in this paper, strategic deception, evaluator awareness, optimization theater, cognitive stratification, are analytically productive but, as currently formulated, empirically underspecified. A mature research program will require operational definitions that support measurement, distinction from adjacent phenomena, and falsification. We identify three priority directions.

Measuring system-level capability amplification. Quantifying the difference between a model's isolated capability and its effective capability when embedded in various system architectures would provide direct evidence for the importance of Layer 2. Experimental designs comparing identical models across series, parallel, and hybrid orchestration architectures on standardized task suites could establish baseline amplification factors [6, 7].

Detecting optimization theater. Developing evaluation methodologies capable of distinguishing genuine alignment from strategic compliance is a central challenge. Approaches might include evaluation environments

that systematically vary the system's beliefs about whether it is being observed, differential analysis of behavior under varying monitoring conditions, and interpretability techniques applied to the reasoning traces of systems under evaluation [5, 15, 22].

Mapping cross-layer interaction effects. Empirical case studies that trace specific risk outcomes through multiple layers of the stack, demonstrating how model-level properties interact with system architecture, deployment context, and governance conditions, would provide the most direct validation of the framework's central compositional claim.

This paper does not claim to resolve these empirical challenges. Its contribution is primarily conceptual: proposing a structured framework that organizes existing concerns and identifies where empirical work is most needed. The framework's value lies not in providing definitive answers but in ensuring that the right questions are being asked at the right level of analysis.

8. Discussion

The six-layer framework advanced here differs from existing approaches to AI risk in several respects. Unlike taxonomies organized by harm type (e.g., Weidinger et al. [22]), it is organized by structural layer, distinguishing risks by where in the sociotechnical stack they originate rather than by what kind of harm they produce. Unlike governance frameworks focused primarily on regulatory instruments (e.g., Anderljung et al. [8]), it encompasses the full range from technical properties to civilizational consequences. And unlike alignment-focused analyses (e.g., Ngo et al. [4]), it situates alignment as one layer within a broader structure rather than treating it as the singular axis of concern.

The framework also clarifies where different actors have leverage. Developers have primary control over model and system design (Layers 1–2). Deployers control deployment decisions (Layer 3). Access controllers have power over who can use systems and under what conditions (Layer 4). Regulators and governance institutions must coordinate across all layers (Layer 5). And democratic publics have a stake in whether the cumulative effects of these decisions produce civilizational outcomes consistent with human agency and self-determination (Layer 6).

The paper's central argument, that frontier AI risk is compositional and that governance must shift from component certification to system-level evaluation under realistic operating conditions, has a direct historical analogy: the transition in aviation safety from inspecting individual aircraft components to certifying assembled systems in flight. Frontier AI governance must make the same transition. Without it, oversight may certify components while failing to understand the full system they compose.

Several limitations should be noted. The framework is primarily analytical rather than empirical, its layers are proposed based on structural analysis and existing evidence rather than derived from systematic empirical investigation. The boundaries between layers are necessarily somewhat porous; reasonable analysts may disagree about which layer a specific risk belongs to. And the framework does not directly prescribe policy interventions, though it does identify where interventions are most needed. These are productive areas for further development.

9. Conclusion

The future of AI risk will be determined less by isolated model behavior and more by the interaction of capability, architecture, access, and governance. The frontier is no longer just a model frontier; it is now a system frontier, a power frontier, and increasingly, a governance frontier. These frontiers are advancing simultaneously, and the path taken at this moment will shape the technological landscape and power distribution for decades to come.

The central mistake in much of today's AI risk discourse is not that people are asking the wrong questions, it is that they are often asking them at the wrong level. Frontier AI risk is not one thing. It is a layered, compounding, and increasingly systemic phenomenon. A coherent response therefore requires more than better models, better prompts, or better benchmarks. It requires a framework capable of seeing the full stack.

This paper has proposed one such framework, organized around six interacting layers that together capture the full scope of frontier AI risk. The goal is not to provide definitive answers, but to ensure that the right questions are being asked at the right level of analysis, and that governance interventions can be designed with awareness of how changes at one layer propagate through the others.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

[2] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

[3] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[4] Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

[5] Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... & Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.

[6] Chan, A., Salganik, R., Marber, A., Kumar, P., Rashid, M., & Frase, H. (2023). Harms from increasingly agentic algorithmic systems. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666.

[7] Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., ... & Dafoe, A. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

[8] Anderljung, M., Barnhart, J., Korber, A., Leung, J., O'Keefe, C., Whittlestone, J., ... & Heim, L. (2023). Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.

[9] Dafoe, A. (2018). AI governance: A research agenda. *Future of Humanity Institute, University of Oxford*.

[10] Weng, L. (2023). LLM-powered autonomous agents. *Lil'Log*.

[11] Significant Gravitas. (2023). AutoGPT: An experimental open-source attempt to make GPT-4 fully autonomous. *GitHub repository*.

[12] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

[13] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... & Wang, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345.

[14] Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., ... & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, *384*(6698), 842–845.

[15] Anthropic. (2025). The Claude model spec. *Anthropic Technical Documentation*.

[16] Anthropic. (2024). Claude 3.5 Sonnet system card. *Anthropic Technical Documentation*.

[17] Zwetsloot, R. & Dafoe, A. (2019). Thinking about risks from AI: Accidents, misuse, and structure. *Lawfare Blog*.

[18] Korinek, A. & Stiglitz, J. E. (2021). Artificial intelligence, globalization, and strategies for economic development. *NBER Working Paper 28453*.

[19] Christiano, P. (2019). What failure looks like. *AI Alignment Forum*.

[20] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437.

[21] Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

[22] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

[23] Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy*, *11*(5), 545–556.

[24] Critch, A. & Krueger, D. (2020). AI research considerations for human existential safety (ARCHES). *arXiv preprint arXiv:2006.04948*.

[25] Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., ... & Anderljung, M. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*.