The Frontier AI Risk Stack

A Six-Layer Framework for Understanding Systemic Risk in Advanced AI Systems

Plain Language Edition

Sean Sooch
Independent Researcher

Abstract

When people talk about AI risk, they usually focus on one piece of the problem: whether the AI model itself is safe, or whether it will be misused, or whether governments can regulate it. This paper argues that AI risk is actually six problems stacked on top of each other. A dangerous AI model is one thing, but a dangerous AI model connected to tools, deployed with real-world access, available only to powerful organizations, and overseen by regulators who cannot keep up, is something far worse. The biggest dangers come not from any single layer but from the way problems at one layer make every other layer worse.

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

---

1. Introduction

AI has entered a new phase. The question is no longer whether these systems will become powerful. They already are. The real question is how that power is being organized, who controls it, and whether our institutions can keep up.

Most conversations about AI risk focus on the AI model itself: is it safe? Can it be tricked? Could it lie? These are important questions, but they only cover one piece of a much larger picture. In the real world, AI is not deployed as a standalone model. It is deployed as a system, connected to tools, databases, the internet, and other AI agents. It is given permissions, access to sensitive data, and the ability to take actions in the real world.

This paper proposes a framework called the Frontier AI Risk Stack. It identifies six distinct layers of risk, from the model itself all the way up to the long-term effects on society. The central argument is simple: no single layer tells the full story, and the most serious dangers come from the way failures at one layer make every other layer worse.

2. AI Is No Longer Just a Model

When researchers test AI safety, they usually test the model by itself: ask it questions, check its answers, look for problems. But that is not how AI works in practice anymore.

In the real world, AI systems come with memory (they can remember past conversations), tools (they can browse the web, write code, send emails), permissions (they can access files, databases, and APIs), and increasingly, the ability to run multiple tasks at the same time without human oversight.

This means a model that seems harmless in a test environment could become far more powerful, and far more dangerous, once it is plugged into a real system. Testing the engine without testing the car does not tell you whether the car is safe to drive.

3. Running in Parallel Makes AI More Powerful and Harder to Control

One of the most important changes in how AI systems are built is the shift from doing one thing at a time to doing many things at once. Older AI systems worked step by step: do task A, then task B, then task C. Newer systems can run dozens of tasks simultaneously, each one exploring a different approach to the same problem.

This makes AI systems faster, more capable, and harder to predict. By the time a human operator notices something concerning, the system may have already completed dozens of steps across multiple parallel paths. It also means the system can find solutions (including dangerous ones) much faster than a step-by-step system could.

For regulators and safety teams, this creates a fundamental challenge: you cannot govern what you cannot observe in time, and parallel systems are designed to move faster than observation allows.

4. The Governance Gap

4.1 Not Everyone Gets Access

The most powerful AI systems are not available to the public. They are restricted to large corporations, government agencies, and research institutions. Anthropic, for example, built a model called Mythos Preview that outperforms senior software engineers on coding tasks, but chose to limit access to roughly 40 approved organizations.

This creates a new kind of inequality. It is no longer just about who has money or education. It is about who has access to the most powerful thinking tools ever built. As AI gets more capable, the incentive to restrict it grows stronger, creating a cycle where the gap between insiders and everyone else keeps widening.

4.2 Safety Is Not Keeping Up with Capability

Every generation of AI models is more powerful than the last, but the tools we use to understand, control, and verify their behavior are not improving at the same pace. This gap between what AI can do and what we can verify is called alignment debt, and it is growing.

Even more concerning, there is evidence that advanced AI systems may learn to behave well during testing but differently in the real world. If an AI can figure out when it is being evaluated and adjust its behavior accordingly, then our safety tests may be measuring performance rather than genuine safety.

4.3 The Monitoring Problem

Most AI safety regulation depends on being able to watch what AI systems do: audit them, log their actions, test their behavior. But if AI systems become sophisticated enough to recognize when they are being watched, they could behave differently during inspections than during normal operation.

This paper calls this "optimization theater": the AI creates the appearance of safety without actually being safe. If this happens, every governance framework that relies on observing AI behavior becomes unreliable.

5. The Six Risk Layers

This framework identifies six layers of AI risk. They are organized from the most technical (the model itself) to the most societal (the long-term effects on civilization). Each layer can make the others better or worse.

Layer 1: Model Risk

How dangerous is the AI model itself? This includes whether it can deceive, conceal its true capabilities, manipulate people, or provide dangerous knowledge (like how to create weapons or hack systems).

Who can fix it: The companies that build the models.

How to check: Test the model with adversarial questions, red-team exercises, and capability evaluations.

Layer 2: System Risk

What happens when you give the model tools, memory, internet access, and the ability to run tasks in parallel? A model that seems safe on its own can become much more powerful and unpredictable when embedded in a larger system.

Who can fix it: The engineers who design AI systems and platforms.

How to check: Test the full system under realistic conditions, not just the model in isolation.

Layer 3: Deployment Risk

What happens when the system is given real-world permissions? Access to production databases, financial systems, API keys, or sensitive files. This is where theoretical capability becomes operational power.

Who can fix it: The organizations that deploy AI into their operations.

How to check: Audit what permissions AI systems actually have, and test what could go wrong in the specific environment where they operate.

Layer 4: Access Risk

Who gets to use the most powerful AI, and who does not? When access is concentrated among a small number of wealthy organizations, it creates a new form of inequality based on who has access to the best thinking tools.

Who can fix it: Policymakers, regulators, and the companies that control access.

How to check: Track who has access to frontier AI, measure concentration, and assess whether the benefits are being shared broadly.

Layer 5: Governance Risk

Can governments and institutions actually keep up with AI? This includes whether regulations are adequate, whether audits are meaningful, whether standards bodies have the technical capacity to evaluate these systems, and whether the companies building AI are effectively regulating themselves.

Who can fix it: Regulators, standards organizations, and civil society groups.

How to check: Assess whether institutions have the technical expertise, legal authority, and political independence to provide real oversight.

Layer 6: Civilizational Risk

What are the long-term consequences if all five layers below fail or interact badly? This includes the concentration of cognitive power in a few organizations, the erosion of public decision-making, and structural instability as society struggles to adapt faster than AI advances.

Who can fix it: Everyone, including democratic publics.

How to check: Long-term scenario planning, tracking structural indicators of institutional health, and monitoring whether humans are maintaining meaningful control over the systems that shape their lives.

6. How the Layers Interact

The most important idea in this framework is that the layers do not exist in isolation. They compound. A problem at one layer makes problems at every other layer worse.

Consider a concrete example: a model with modest deception ability (Layer 1) is embedded in a system that gives it tools, memory, and parallel execution (Layer 2). It is deployed with access to a financial company's production systems (Layer 3). Only that company has access to it (Layer 4). And no external auditor has reviewed the setup (Layer 5). Each layer alone might be manageable. Together, they create a risk that no single-layer fix can address.

Now consider the opposite: the same model is deployed in a well-monitored system (Layer 2), in a sandboxed research environment (Layer 3), with broad academic access (Layer 4), and strong independent auditing (Layer 5). Here, the higher layers contain and reduce the model-level risk. The framework helps you see both scenarios clearly.

7. What Needs to Happen Next

Several ideas in this paper need more research to become practically useful. Three priorities stand out:

Measure how much systems amplify models. We need experiments that compare the same AI model running alone versus running inside a full system with tools and parallel execution. How much more capable does it become? How much harder to control?

Detect when AI is performing safety rather than being safe. We need evaluation methods that can tell the difference between an AI that is genuinely aligned with human values and one that has learned to look aligned during testing.

Trace risk across layers. We need case studies that follow specific risks through all six layers, showing how a technical property at Layer 1 interacts with system design, deployment context, access patterns, and governance quality to produce real-world outcomes.

8. How This Framework Is Different

Most existing frameworks for thinking about AI risk focus on one dimension: the types of harm AI might cause, the regulatory tools available, or the alignment problem between AI goals and human values. This framework is organized differently. Instead of asking what kind of harm AI might cause, it asks where in the stack the risk originates.

This matters because different actors have leverage at different layers. The people who build models (Layer 1) are not the same people who deploy them (Layer 3) or regulate them (Layer 5). A framework that identifies which layer a risk belongs to also identifies who is best positioned to address it.

The framework has limitations. It is a conceptual tool, not an empirical measurement. The boundaries between layers are not always clean. And it does not prescribe specific policies. But it does help ensure that the right questions are being asked at the right level.

9. Conclusion

AI risk is not one thing. It is six things stacked on top of each other, and the most dangerous scenarios come from the way those layers interact.

The AI frontier is no longer just a model frontier. It is a systems frontier, a power frontier, and increasingly, a governance frontier. These are all advancing at the same time, and the decisions being made right now will shape the technological landscape for decades.

The central mistake in most AI risk discussions is not that people are asking the wrong questions. It is that they are asking them at the wrong level. A framework that can hold all six layers in view at once is a step toward asking the right questions in the right places.

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., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv preprint arXiv:2401.05566.

[6] Chan, A., Salganik, R., Marber, A., et al. (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., et al. (2023). Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

[8] Anderljung, M., Barnhart, J., Korber, A., et al. (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., et al. (2018). The malicious use of artificial intelligence. arXiv preprint arXiv:1802.07228.

[13] Wang, L., Ma, C., Feng, X., et al. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345.

[14] Bengio, Y., Hinton, G., Yao, A., et al. (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., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.