Concentration and Diffusion of Authority in AI Governance

By Raj Rajagopalan, Internautica, Inc.

In early April, Anthropic made one of the most consequential access decisions in the history of cybersecurity. A new model called Mythos — capable of finding zero-day vulnerabilities in every major operating system and browser, including a 27-year-old bug in OpenBSD — was quietly announced alongside Project Glasswing, a coalition of over 40 technology companies including Apple, Google, Microsoft, Cisco, and Broadcom, granted early access and $100 million in usage credits to find and patch vulnerabilities before broader release.

The interesting part of this decision was that it was made by a private company, with no public deliberation, input from democratic institutions or regulatory institutions. The government was briefed, but as reporting made clear, did not take Anthropic up on its offer to help evaluate the model — in part because the same administration had previously tried to designate Anthropic a supply chain risk after the company declined to participate in surveillance and autonomous weapons contracts.

The most pointed critique came not from a regulator but from an academic. “Whatever the right judgment call is,” said Jonathan Iwry of the Wharton Accountable AI Lab, “the most striking aspect of this situation is how reliant we are on the judgment of a handful of private actors who aren’t accountable to the public.”

That sentence names the problem precisely and it is not a new problem, as we’ve been here before. This is a question of AI governance in its most fundamental sense — not compliance frameworks or internal safety reviews, but the fulfillment of a basic user need: the right of the people who bear the consequences of a technology to have a legitimate say in decisions about it. As an architect who has spent decades viewing security not as a compliance obligation but as the fulfillment of human needs, I find the absence of that accountability structure not merely a policy gap but a design failure.

The example of recombinant DNA

When recombinant DNA research emerged in the mid-1970s, the scientific establishment's instinct was to manage the risks internally. Self-governance seemed the obvious answer, considering that they were after all, the ones who understood the technology. That is until Cambridge city councilman Alfred Vellucci intervened.

A politician known more for his theatrical antagonism toward Harvard and MIT than for any technical sophistication, Vellucci convened public hearings in 1976 that led to the creation of the Cambridge Experimentation Review Board — a body of ordinary citizens, charged with overseeing DNA research conducted within city limits. The research community was appalled and warned that non-experts couldn't possibly evaluate the risks. The hearings were chaotic, contentious, and deeply uncomfortable for everyone involved, yet they were also right.

The Cambridge model did not stop the research or require the scientists to dumb down their work. What it did was assert a principle that we have largely abandoned in the age of AI: that the people who bear the risks of a technology have a legitimate claim to participate in decisions about it, and that democratic accountability is not a luxury that can be deferred until experts feel comfortable granting it.

Vellucci's insight was not technical but rather political, in the deepest sense of the word. He understood that the question of what risks a community should accept is not a scientific question, but rather a question about power, about whose judgment counts, and about who is accountable when things go wrong.

Fifty years later, Anthropic made a decision affecting every organization that runs software on a major operating system — which is to say, essentially everyone — and the mechanism for public accountability was a quote from a business school professor in a trade publication.

The Technical is the Political

To understand why this matters beyond the Glasswing case, it helps to look at what the AI safety research community has been quietly discovering over the past several years: that every major technical decision in AI safety is also a political one, whether or not anyone acknowledges it.

Consider safety benchmarks — the tools researchers use to evaluate whether AI models behave safely. A benchmark like SafeDialBench evaluates models across six dimensions: Fairness, Legality, Morality, Aggression, Ethics, and Privacy. This looks like a scientific instrument. It is not.

Someone decided that these six dimensions — and not others — are the ones that matter. Someone decided what counts as a violation within each category. Someone decided how to weight them. These are values decisions, not engineering decisions. The benchmark naturalizes those choices, making them invisible. But invisible decisions are still decisions, made by someone, reflecting someone's priorities.

The same pattern appears at a deeper level of abstraction, in the domain of moral evaluation itself. A landmark paper published in Nature this February by researchers from Google DeepMind argued for moving beyond what they call moral "performance" — producing outputs that look ethical — to moral "competence" — reasoning that is actually grounded in moral principles. Their framework identifies three fundamental challenges: the facsimile problem (models may imitate moral reasoning without genuine understanding), moral multidimensionality (real moral decisions involve complex, context-sensitive tradeoffs), and moral pluralism (different cultures have legitimately different moral frameworks, and AI systems are deployed globally).

That third challenge — moral pluralism — is where the technical problem becomes undeniably political. A model trained primarily on English-language data, aligned to satisfy predominantly Western evaluators, embeds a particular moral and cultural perspective. Research has shown that safety alignment in large models poorly generalizes across languages: ask a model in Zulu what you cannot ask it in English and you may get a very different answer. The safety layer, it turns out, is largely an English-language layer.

This is not an accident that better engineering will fix. It is a structural consequence of who builds these systems, whose data trains them, and whose feedback shapes their values. When Anthropic or OpenAI defines what "safe" means in their alignment training, they are making a governance decision that affects billions of people who had no voice in it. The Cambridge City Council at least had jurisdiction over the people it was protecting. The AI industry has no such constraint.

The Institutional Gap

None of this is an argument that AI companies are malicious or that their technical work is without value. Anthropic's decision to pause broad Mythos deployment and organize Glasswing may well be the right call. The open-source alternative — releasing a model capable of finding zero-days in every major operating system to anyone who wants it — would be worse.

But right calls made by unaccountable actors are still a governance failure. This is a distinction that the AI safety discourse has consistently struggled to hold. The National Security Agency made right calls too, for decades, until EternalBlue — an NSA-developed exploit — was stolen, weaponized

into WannaCry, and used to cripple hospitals, banks, and infrastructure across 150 countries. The problem was not that the NSA's engineers were incompetent. It was that no accountable institution was positioned to ask whether accumulating that capability in secret was wise, or to mandate the kind of disclosure that might have prevented the theft.

The structural parallel is exact. A powerful capability, held by a small group of technically sophisticated actors, managed according to their own judgment, with inadequate mechanisms for external accountability. The question is not whether the people holding the capability are trustworthy. The question is whether trustworthiness — of individuals, of companies, of any private actor — is a sufficient substitute for institutional accountability.

In a democratic society, this is what institutions are for.

What the Institution Should Look Like

The Last Vote, a recent framework for multi-stakeholder democratic governance of large language models, offers a useful starting point. Its core argument is that different stakeholder groups — technical practitioners, researchers, civil society organizations, and crucially, democratic representatives — possess forms of knowledge that are non-substitutable. Technical expertise tells you what a system can do. Democratic representation tells you what risks a community is willing to accept. These are different kinds of knowledge, and both are necessary.

The Cambridge Experimentation Review Board was not a technical body. It was a civic one, with technical advisors. It did not try to replicate the scientists' expertise. It asked different questions: Who bears the risk? Who consented? What happens if this goes wrong and who is accountable? Those questions are not answered by technical competence. They are answered by legitimate representation.

The complexity science literature offers a complementary frame. Adaptive governance — governance designed to function under genuine uncertainty, where the behavior of the system being governed is emergent and not fully predictable — requires diversity of perspective, not just depth of expertise. A governance structure that optimizes for technical sophistication but lacks input from affected communities is not just normatively deficient. It is functionally brittle. It will fail in the ways that monocultures always fail: by being unable to anticipate the questions it never thought to ask.

What would this look like in practice? Not a regulatory body staffed by engineers or a standards committee dominated by industry. A civic institution with genuine authority — the power to require

disclosure, to mandate independent evaluation, to establish and enforce rules about who gets access to what capabilities and under what conditions — staffed by a diverse body of stakeholders and advised by, but not controlled by, technical experts.

The details matter and they are hard, but the principle is not complicated, as Vellucci understood it in 1976. The people who bear the consequences of a technology have a right to participate in decisions about it. That right does not disappear because the technology has become more complex.

The Moment We Are In

Anthropic has described the period following Mythos's emergence as potentially "tumultuous" — a transitional moment before a new equilibrium is reached in which AI benefits defenders as much as attackers. That framing is probably correct as a technical forecast. But it obscures the governance question.

Transitions are when institutions get built or fail to get built. The choices made in the next few years about who has authority over AI capabilities, how that authority is exercised, and what accountability mechanisms exist will shape the landscape for decades. We built the wrong institutions after the development of social media — or more precisely, we built no institutions worth the name — and we are still living with the consequences.

The Glasswing decision was not an aberration but rather a preview. As AI systems become more capable, the decisions about what to build, how to deploy it, and who gets access will become more consequential, not less. Each of those decisions will be made by someone. The question is whether they will be made by accountable institutions or by the judgment of private actors who happen to be in the room.

Alfred Vellucci was not in the room in 1976 but made himself relevant anyway. The question for AI governance today is whether we can build the equivalent of the Cambridge Experimentation Review Board before the next Glasswing decision — or whether we will keep outsourcing our most consequential choices to organizations that are excellent at building AI and constitutionally unable to govern it.

The technology will not wait for us to figure this out. It never does.

Raj Rajagopalan has spent decades working at the intersection of human-centered security and enterprise architecture. This article draws on ongoing research and work at Internautica, Inc.

Sources

On Project Glasswing and Mythos

Statt, N. “Anthropic’s Mythos and the Cybersecurity Risk Experts Are Warning About.” Platformer, April 2026. https://www.platformer.news/anthropic-mythos-cybersecurity-risk-experts/

“Anthropic’s Mythos: AI-Driven Cybersecurity Risks Are Already Here.” Fortune, April 10, 2026. https://fortune.com/2026/04/10/anthropic-mythos-ai-driven-cybersecurity-risks-already-here/

“Mythos Preview: Capabilities and Concerns.” Anthropic Red Team Blog, 2026. https://red.anthropic.com/2026/mythos-preview/

On AI Safety Benchmarks and Alignment

Xu et al. “SafeDialBench: A Fine-Grained Safety Evaluation Benchmark for LLMs in Multi-Turn Dialogues with Diverse Jailbreak Attacks.” arXiv, 2025. https://arxiv.org/abs/2502.11090

Haas, J., Bridgers, S., Manzini, A. et al. “A Roadmap for Evaluating Moral Competence in Large Language Models.” Nature 650, 565–573, 2026. https://doi.org/10.1038/s41586-025-10021-1

Yong, Z.X. et al. “Low-Resource Languages Jailbreak GPT-4.” arXiv, 2023. https://arxiv.org/abs/2310.02446

On AI Control and Institutional Design

Greenblatt, R., Shlegeris, B., Sachan, K., Roger, F. “AI Control: Improving Safety Despite Intentional Subversion.” arXiv, 2023. https://arxiv.org/abs/2312.06942

Gardner-Challis et al. “When Can We Trust Untrusted Monitoring? A Safety Case Sketch Across Collusion Strategies.” arXiv, 2026. https://arxiv.org/abs/2602.20628

On Democratic Governance of AI

Haas et al. “The Last Vote: Multi-Stakeholder Democratic Governance of LLMs.” arXiv, 2025. https://arxiv.org/abs/2511.13432

Woolley, S. “The AI Democracy Dilemma.” Journal of Democracy, 2025.

Ilcic, A., Fuentes, M., Lawler, D. “Artificial Intelligence, Complexity, and Systemic Resilience in Global Governance.” Frontiers in Artificial Intelligence, June 2025. https://doi.org/10.3389/frai.2025.1562095

On the Cambridge Recombinant DNA Controversy

Krimsky, S. Genetic Alchemy: The Social History of the Recombinant DNA Controversy. MIT Press, 1982.

Wade, N. “Gene Splicing: Cambridge Citizens Embroiled in Dispute with Harvard and MIT.” Science 195, 1977.