TECHNOLOGY
POLICY
INSTITUTE

Building the Analytical Infrastructure for Governing Frontier AI Development

Scott Wallsten

March 2026

Building the Analytical Infrastructure for Governing Frontier AI Development

Scott Wallsten[*]

Technology Policy Institute

March 2026

[*] President and Senior Fellow, Technology Policy Institute. swallsten@techpolicyinstitute.org
This paper reflects my views only, and not necessarily those of TPI, its staff, or its boards. I thank Tim Bresnahan, Tyler Cowen, Joshua Gans, Avi Goldfarb, Matt Perault, Mary Shirley, and Catherine Tucker for comments. All errors are my own. Comments welcome.

Executive Summary

Artificial intelligence presents genuinely novel governance challenges. Its combination of rapid capability growth, omni-use applications, and nearly free proliferation is inconsistent with templates designed for other technologies that may have dangerous applications. Recent attempts to govern AI development have had mixed results, with many proposals vetoed, revoked, weakened, or still searching for enforcement mechanisms. The deeper problem is not political will. It is that the analytical foundations for thinking rigorously about AI governance do not yet exist. We lack even a shared vocabulary. The phrase "AI safety" means at least four different things across different communities, and as a result, summits, reports, and legislative proposals talk past each other.

This paper proposes building the missing analytical infrastructure through a dedicated, independent research institution with a long time horizon. The RAND Corporation's role in nuclear strategy is a useful analogy, not because AI is comparable to nuclear weapons, but because AI, like nuclear technology in the late 1940s, presents strategic questions that existing intellectual frameworks cannot answer. Over roughly fifteen years, RAND and affiliated researchers produced the conceptual vocabulary (deterrence, crisis stability, second-strike capability) that made nuclear governance possible. AI governance has no comparable vocabulary and no institution dedicated to building one.

The proposed institution would be lean, comprising roughly fourteen resident scholars and an annual budget of approximately $12 million, closer in spirit to the Institute for Advanced Study, though much smaller, than to the "CERN for AI" proposals that envision billions in shared infrastructure. It would not need access to proprietary models or training data. Its questions are about actors, incentives, and institutions, not about what any particular frontier AI model can do. Its outputs would include assessments of whether safety frameworks actually constrain behavior, formal models of AI competition dynamics, structured scenario analyses for regulators, and post-incident analysis producing shared facts.

The paper examines why no existing institution fills this role. Government-housed bodies are politically fragile (the U.S. AI Safety Institute was dismantled and repurposed in less than 2 years). University centers are vulnerable to institutional politics (Oxford closed the Future of Humanity Institute in 2024). Industry-funded bodies gravitate toward consensus. Each reason for failure informs the proposed design, which involves freestanding governance, an irrevocable endowment severing funder from research, geographic separation from political capitals, and scholar-driven rather than program-directed organization.

The paper tries to stress-test the idea. A "Why This Might Be Wrong" section considers whether governance risks increasing concentration, whether the object of control is too entangled with general-purpose infrastructure, whether existing institutions could simply be scaled, whether a distributed regime complex is preferable, and whether AI might be the kind of problem that resists institutional solutions entirely.

What the objections cannot answer is the cost of inaction. The longer the wait, the more governance builds on foundations no one has tested, and the harder those foundations become to replace.

Introduction

The central problem for AI governance, more than any particular policy issue, is that we do not have the intellectual frameworks to think clearly about it. Alongside enormous potential benefits, artificial intelligence brings legitimately frightening possibilities like autonomous lethal weapons, lower barriers to synthesizing dangerous compounds, and disinformation deployed at scale. As Geoffrey Hinton (2024) put it, “Anybody who says it's all going to be fine, it's crazy. Anybody who says they're inevitably going to take over, they're crazy too. We really don't know.”[1]

What makes AI unique, and not simply the latest in a long series of dual-use technologies, is a set of properties that interact in ways particularly difficult for existing governance templates.[2] First, the capability frontier is advancing faster than any institutional process can characterize, let alone regulate. Second, an AI model or set of model weights has a nearly limitless range of uses. It can be a medical research tool, a logistics optimizer, a weapons design assistant, or a disinformation engine depending entirely on how it is prompted. Dual-use usually means technology with both civilian and military applications. AI is omni-use. A user can move from beneficial to harmful deployment with a conversation, not a redesign. Third, proliferation, once model weights are released, is limited mostly by inference compute, which is orders of magnitude cheaper than frontier training.

AI as a quickly developing general purpose technology that can be adopted at low cost is a key reason why it has so much positive potential. But those same properties also mean that the standard policy responses to dangerous technologies all break down simultaneously. You cannot test exhaustively because capabilities are emergent and context-dependent. You cannot restrict access because the knowledge diffuses faster than any control regime can contain it.

This paper focuses on the development trajectory and cross-border strategic dynamics, not on sector-specific rules for particular applications, which address a different set of questions. Those, also, are difficult but we have the tools and often existing laws and institutions to think about them.

The speed of AI development makes the investment in analytical foundations more urgent, not less. Governance frameworks built on inadequate understanding will either fail to address actual

[1] The full quote: “The question is what’s going to happen when we’ve created beings that are more intelligent than us and we don’t know what’s going to happen? We’ve never been in that situation before. Anybody who says it’s all going to be fine, it’s crazy. Anybody who says they’re inevitably going to take over, they’re crazy too. We really don’t know. But because we really don’t know, it will make a lot of sense to do a lot of basic research now on whether we can stay in control of things that we create that are more intelligent than us.” “Nobel Prize Interview: Geoffrey Hinton,” Nobel Foundation, 2024. Available at https://www.nobelprize.org/prizes/physics/2024/hinton/1925103-interview-transcript/

[2] Several other domains share some of these features. Cyberattacks can be replicated at low marginal cost and can be difficult to observe, biotechnology has dual-use and verification problems, and finance shares fast-moving systemic risk. But in frontier AI, several factors intersect in a general-purpose tool that can be repurposed quickly while frontier development remains concentrated enough to create politically contested chokepoints.

risks or impose costs that outweigh the harms they prevent.[3] When the technology outruns specific rules, the only durable investment is in the capacity to evaluate and adapt, not in any particular regulation.

There is no shortage of policy proposals. The AI governance conversation has spent several years trying to adapt arms control, product safety regulation, climate agreements, and regulatory harmonization to a technology whose structural properties do not fit any of them. More summits, more declarations, and more legislation aimed at AI development have not produced especially meaningful results. The 2026 International AI Safety Report, authored by over 100 experts and backed by more than 30 countries, noted that most risk management frameworks remain voluntary, that policymakers have limited visibility into how risks are actually identified and managed in practice, and that information-sharing between developers, deployers, and infrastructure providers remains fragmented.[4] These efforts do not need better coordination. The issue is that the analytical foundations they would coordinate around have not yet been built. Without such a framework, proposals oscillate between ambitious 'international agency' ideas and incremental toolkits of standards, audits, and incident reporting.

Even the most basic vocabulary reveals the absence of shared frameworks. The phrase 'AI safety' alone refers to at least four distinct concepts across different communities. It can mean technical alignment research aimed at ensuring AI systems pursue intended goals, evaluating dangerous capabilities and red-teaming frontier models, robustness and reliability engineering for deployed systems, and designing governance institutions for frontier AI development.[5] When an alignment researcher, a Pentagon strategist, a corporate ethics team, and a frontier AI regulator each say 'AI safety,' they are describing different problems requiring different expertise and different solutions. Even the term "governance" is fraught, since it implies we already know what to govern and are just debating how. For AI development, we don't yet know what, if anything, to govern.

This paper argues that the right response is not to keep forcing familiar governance templates onto a technology they were not built for, nor to treat the absence of workable models as reason to ignore the risks. It is to invest in building the intellectual infrastructure necessary for thinking through these issues rigorously.

That investment requires a dedicated institution. Not another government advisory committee or industry self-regulation body, but an independent, well-funded research program with the mandate and patience to develop foundational frameworks over a decade or more. Something analogous to the role the RAND Corporation played in nuclear strategy, not because AI is a

[3] The speed of development also increases the costs of getting governance wrong. In a slow-moving field, a misguided regulation is costly but correctable. When the technology advances this quickly, premature rules risk channeling development onto unproductive paths that become self-reinforcing as investment, infrastructure, and talent organize around the distorted incentives.

[4] Yoshua Bengio et al., *International AI Safety Report 2026*, version 1 (arXiv, 2026), https://doi.org/10.48550/ARXIV.2602.21012.

[5] For examples of distinct usages, see Bengio et al., International AI Safety Report 2026 (using the term primarily for dangerous capability evaluation and risk management); OpenAI, 'How We Think About Safety and Alignment,' 2025 (distinguishing human misuse, misaligned AI, and societal disruption as separate categories under a single 'safety' umbrella). Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019).

weapon like nuclear bombs, but because AI today, like nuclear technology in the 1940s and 1950s, presents some novel strategic questions where existing intellectual frameworks are inadequate and the productive response is to build new tools for thinking.

Why Governance Is Difficult

AI governance faces two key obstacles.

First, AI development is distributed and proliferating. The knowledge base is open, and barriers to deploying and adapting high-capability models are falling, even as frontier training remains capital-intensive. Meta has released frontier model weights publicly. Models like DeepSeek R1 and Kimi K2 came from Chinese startups that drew attention for their relatively low reported costs. Some governments promote locally developed models, including France's Mistral and the UAE's Falcon. To be clear, openness and competition themselves generate enormous benefits by accelerating research, lowering barriers for developing countries, and enabling applications that proprietary models cannot support.[6] But proliferation is permanent and irrevocable, which makes development-level governance structurally harder than regimes designed for scarce physical goods.

Second, many countries consider AI development essential to their economic and strategic futures. President Trump called AI dominance "a national security imperative," declaring that the United States must "achieve and maintain unquestioned and unchallenged global technological dominance." Chinese President Xi Jinping described AI as "the next epoch-making technological transformation" and directed a "whole-of-nation" approach. French President Macron framed France's AI investment as "our fight for sovereignty, for strategic autonomy." To some countries, like Ukraine, AI development is literally existential as it is a key to defending against Russia.

The perceived importance of AI to countries' futures influences and creates competing governance visions. China's "Global AI Governance Action Plan" frames the challenge through inclusivity, sovereignty, and capacity-building, and recent proposals for a Chinese-led global AI cooperation organization represent an explicit alternative to the Western-anchored summit and safety institute process.[7] When major powers disagree not just about the terms of cooperation but about the purpose and structure of governance itself, the absence of shared analytical frameworks becomes a strategic liability rather than a bureaucratic gap.

The result is not surprising. Recent attempts to govern AI development have generally not produced convergence on verifiable commitments.

California's SB 1047 would have imposed pre-training safety protocols and developer liability. The Biden administration's Executive Order 14110 required reporting for models above certain compute thresholds. The AI Safety Summit series, from Bletchley Park through Seoul to Paris,

[6] GovAI, "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods" (2023).

[7] *Global AI Governance Action Plan* (Ministry of Foreign Affairs, People's Republic of China, 2025), https://www.fmprc.gov.cn/mfa_eng/xw/zyxw/202507/t20250729_11679232.html.

attempted to build an international architecture for frontier AI governance.[8] Each was vetoed, revoked, or significantly weakened, though the broader process generated some durable outputs, notably the International AI Safety Report series and expanded voluntary industry safety frameworks.

Global efforts have followed a similar trajectory. The EU has proposed delaying parts of its AI Act compliance deadlines and has withdrawn the AI Liability Directive.[9] International declarations proliferate without enforcement mechanisms.[10] The US and UK refused to sign the Paris AI Summit communiqué, although for opposite reasons. The UK said it was insufficient on governance and security, the US said it went too far toward regulation.[11] The India AI Impact Summit 2026 continued the drift toward broad economic themes rather than development-level constraints. The trajectory from Bletchley to Paris shows initial governance ambitions narrowing as competitive dynamics reasserted themselves.

When the strategic stakes are this high for this many actors, tools designed to restrict access or slow development will be ineffective from the start or face relentless pressure to erode. Stopgaps like export controls are worth maintaining, but they cannot substitute for the analytical frameworks needed to design durable governance.

The Analytical Gap

The literature on AI governance has matured rapidly in recent years. A leading synthesis by Anderljung et al. (2023) proposes focusing on highly capable foundation models and using requirements like evaluations, incident reporting, and risk management for organizations training frontier systems.[12] Brundage, et al (2020) outline mechanisms for supporting verifiable claims about development practices, aimed at making external scrutiny possible without relying purely on trust.[13]

Other work defines "technical AI governance" as the set of technical tools needed to make governance enforceable. These include measurement, audits, provenance tracking, compute monitoring, and controlled access to model artifacts.[14] Egan and Heim (2023) propose treating compute providers as chokepoints through "Know Your Customer" schemes.[15] Maas and Trager

[8] The Bletchley Declaration by Countries Attending the AI Safety Summit, Nov. 1-2, 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.

[9] EU AI Act, Regulation 2024/1689, provisions on general-purpose AI models with systemic risk.

[10] One partial exception is the Council of Europe's Framework Convention on Artificial Intelligence, opened for signature in 2024, which is the first legally binding international AI treaty and is open to non-member states. It has not yet entered into force and its enforcement mechanisms remain untested.

[11] Paris AI Action Summit Declaration, Feb. 2025. The US and UK refused to sign, for opposing reasons.

[12] Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety" (arXiv:2307.03718, 2023).

[13] Miles Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," version 2, preprint, arXiv, 2020, https://doi.org/10.48550/ARXIV.2004.07213.

[14] Anka Reuel et al., "Open Problems in Technical AI Governance," *Transactions on Machine Learning Research*, 2025.

[15] Janet Egan and Lennart Heim, "Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers," arXiv:2310.13625, preprint, arXiv, October 20, 2023, https://doi.org/10.48550/arXiv.2310.13625.

et al. (2023) propose an International AI Organization drawing analogies to aviation and maritime governance bodies.[16]

These are serious proposals. But international AI governance has difficulty moving from agenda-setting to enforceable coordination because of verification asymmetries, divergent state interests, and the central role of private developers.[17] Voluntary commitments generate uneven compliance absent strong incentives and monitoring. Anthropic illustrated this problem in February 2026, when it revised its Responsible Scaling Policy to remove the categorical commitment not to train models above capability thresholds without proven safeguards.[18] The revised policy conditions a development delay on Anthropic simultaneously being the frontrunner in AI capabilities and judging catastrophic risk to be significant, and explicitly cites competitive pressure as a reason the original commitment was untenable. The Future of Life Institute's AI Safety Index, which since 2024 has scored frontier AI companies across multiple governance dimensions, found in its 2025 assessment that the highest overall grade among seven leading companies was a C+, and that no company scored above D in existential safety planning.[19]

When the company that pioneered voluntary safety commitments retreats from them under market pressure, and independent evaluation finds no company meeting adequate standards, the evidentiary base for relying on self-governance as a primary mechanism is thin.

Callander (2025) formalizes a related dynamic, arguing that regulating a new industry is a race between policymakers learning enough to act and the industry accumulating enough political power to shape regulation in its own favor. The longer governance is deferred to allow learning, the narrower the window for effective intervention becomes.[20]

Early systematic evaluation of whether companies followed through on the 2023 White House voluntary commitments shows variation across firms, with compliance correlating to timing of signature rather than uniform good faith.[21] The OECD’s assessment of the Hiroshima Process reporting framework, the first international voluntary reporting tool, found that larger technology

[16] Robert F. Trager et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach* (Centre for the Governance of AI, 2023).

[17] See, e.g., Markus Anderljung, *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, 2023; Matthijs Maas and Jose Jaime Villalobos, *International AI Institutions: A Literature Review of Models, Examples, and Proposals*, AI Foundations Report no. 1 (Institute for Law & AI, 2023), https://www.law-ai.org/international-ai-institutions; Bengio et al., *International AI Safety Report 2026.*

[18] *Anthropic’s Responsible Scaling Policy: Version 3.0* (Anthropic, 2026), https://www-cdn.anthropic.com/e670587677525f28df69b59e5fb4c22cc5461a17.pdf; Billy Perrigo, “Anthropic Drops Flagship Safety Pledge,” *Time*, February 2026.

[19] Future of Life Institute, *AI Safety Index* (2025), https://futureoflife.org/wp-content/uploads/2026/01/FLI-AI-Safety-Index-Report-Summer-2025-Rev-Jan-2026.pdf.

[20] Steven Callander, “Regulating AI: The Race Between Policymakers and the Rise of a New Interest Group,” in *The Political Economy of Artificial Intelligence*, ed. Ajay Agrawal et al. (NBER, 2025), https://www.nber.org/books-and-chapters/political-economy-artificial-intelligence/regulating-ai-race-between-policymakers-and-rise-new-interest-group.

[21] Jennifer Wang et al., “Do AI Companies Make Good on Voluntary Commitments to the White House?,” arXiv:2508.08345, preprint, arXiv, September 24, 2025, https://doi.org/10.48550/arXiv.2508.08345.

firms have more sophisticated risk management practices, but technical provenance tools like watermarking and cryptographic signatures remain limited beyond a few major firms.[22]

Three gaps in the current literature matter most for the argument that a sustained research program is needed.

First, rigorous evaluation of governance interventions is thin. We have early scoring of voluntary commitments, but few empirical studies evaluate the causal effects of summits, codes of conduct, or safety institute networks on actual development choices and risk outcomes.[23]

Independent researchers have begun filling pieces of this gap. Coggins, et al (2025) performed a rigorous affordance analysis of OpenAI's Preparedness Framework and found that it structurally permits harm, including CEO authority to override safety determinations.[24] Schuett's (2024) systematic comparison found that frontier AI developers lack corporate governance mechanisms standard in nuclear and aviation industries.[25] METR's (2025) cataloging of frontier safety frameworks documents significant variation in what risks they cover and what actions they trigger.[26] These are valuable contributions, but they are ad hoc efforts by small teams, not the output of sustained institutional capacity. The question is whether governance evaluation at this level of rigor can be maintained and scaled.

Second, verification and monitoring is a fast-growing but still immature domain. Baker’s (2023) analysis of nuclear arms control verification methods argues that verification challenges for AI agreements could be manageable, especially if agreements focus on verifiable aspects like hardware inputs and large-scale training activity.[27] The Oxford Martin AI Governance Initiative’s 2025 report catalogs verification approaches including audits and inspection concepts tailored to model weights and compute infrastructure.[28] But these remain proposals rather than validated technical systems, and fewer political economy analyses exist of what states and firms would actually accept under conditions of strategic rivalry.

Third, strategic interaction modeling for AI competition remains under-specified. The canonical formal model of racing dynamics in advanced AI, Armstrong’s (2016) “Racing to the Precipice,” frames competition among development teams as a game where speed and safety trade off.[29] Newer work by Emery, Park, and Trager (2022) models how sharing or withholding safety-

[22] Karine Perset and Sara Fialho Esposito, *How Are AI Developers Managing Risks? Insights from Responses to the Reporting Framework of the Hiroshima AI Process Code of Conduct*, OECD Artificial Intelligence Papers, No. 45 (OECD, 2025).

[23] Wang et al., “Do AI Companies Make Good on Voluntary Commitments to the White House?”

[24] Sam Coggins et al., “The 2025 OpenAI Preparedness Framework Does Not Guarantee Any AI Risk Mitigation Practices: A Proof-of-Concept for Affordance Analyses of AI Safety Policies,” arXiv:2509.24394, preprint, arXiv, October 13, 2025, https://doi.org/10.48550/arXiv.2509.24394.

[25] Jonas Schuett, “Frontier AI Developers Need an Internal Audit Function,” *arXiv:2305.17038*, 2024, https://arxiv.org/abs/2305.17038.

[26] “Common Elements of Frontier AI Safety Policies (December 2025 Update),” METR, December 2025, https://metr.org/blog/2025-12-09-common-elements-of-frontier-ai-safety-policies/.

[27] Mauricio Baker, “Nuclear Arms Control Verification and Lessons for AI Treaties,” arXiv:2304.04123, preprint, arXiv, April 8, 2023, https://doi.org/10.48550/arXiv.2304.04123.

[28] Ben Harack et al., *Verification for International AI Governance* (Oxford Martin AI Governance Initiative, 2025).

[29] Stuart Armstrong, *Racing to the Precipice: A Model of Artificial Intelligence Development*, 31 (2016).

relevant information can change race dynamics.[30] RAND (2024) argues that uncertainty and low ability to observe AI capabilities can intensify security dilemma dynamics.[31] Lindsay (2025) reinforces this from an international-relations perspective, arguing that military automation does not translate from technical inputs to strategic advantage in the way that 'AI dominance' narratives assume, because political objectives, organizational competence, and the dynamics of bargaining and escalation intervene.[32] But existing models lack empirical anchoring and do not deliver a shared "strategic grammar" for policymakers comparable to what deterrence theory provided in the nuclear era.

This last gap is the most consequential. Relatively little formally integrates strategic stability concepts with development-focused instruments like compute governance, licensing, and model evaluation regimes. The analytical infrastructure for thinking rigorously about AI governance at the development level simply does not yet exist.

The International AI Safety Report process, now in its second edition, represents the most substantial product of the Bletchley-to-Paris summit series. It is a genuine achievement, with over 100 experts from "over 30 countries and international organisations," producing a rigorous evidence synthesis.[33] But the report itself illustrates the gap this paper identifies. It synthesizes existing evidence but does not produce new analytical frameworks. Its central finding is an "evidence dilemma" in which capabilities outpace the evidence base for evaluating risks and mitigations. It documents that twelve companies published or updated Frontier AI Safety Frameworks in 2025 but cannot assess whether those frameworks actually reduce risk, because the empirical evaluation capacity does not exist. The report is what the international community can produce with its current institutional infrastructure. The question is whether that infrastructure is sufficient.

The report model is well suited to the important task of synthesizing evidence across a large international expert community. It is structurally unsuited to what this paper argues is missing. Building strategic interaction models, designing and testing verification regimes, and empirically evaluating whether governance interventions actually work are sustained research enterprises, not periodic assessments. They require small teams working iteratively on specific problems over years, not large expert panels reaching consensus on the state of evidence. The report diagnoses the gaps. Filling them requires a different kind of institution.

The establishment of the UN Independent International Scientific Panel on AI in August 2025, modeled on the IPCC, further institutionalizes the evidence base.[34] Just as the IPCC did not

[30] Nicholas Emery et al., *Information Hazards in AI Races*, June 1, 2022, https://www.governance.ai/research-paper/information-hazards-in-races-for-advanced-artificial-intelligence.
[31] *Artificial General Intelligence and Strategic Competition*, RAND Perspectives nos. PEA4155-1 (RAND Corporation, 2024).
[32] Jon R. Lindsay, "The More Dismal Science: Perspectives from International Relations on Military Automation," in *The Political Economy of Artificial Intelligence* (2025), https://www.nber.org/books-and-chapters/political-economy-artificial-intelligence/more-dismal-science-perspectives-international-relations-military-automation.
[33] Bengio et al., *International AI Safety Report 2026*.
[34] United Nations, "Independent International Scientific Panel on AI," United Nations, 2025, https://www.un.org/independent-international-scientific-panel-ai/en; Elizabeth Gibney, "UN Creates New Scientific AI Advisory Panel: What Will It Do?," *Nature*, February 26, 2026, https://doi.org/10.1038/d41586-026-00542-8.

produce original climate science but instead synthesized the evidence, the ISP-AI, too, is designed to assess and synthesize evidence, not to generate original analytical frameworks. Without foundational strategic research on AI governance, there is nothing to synthesize. The proposed institution and the ISP-AI are complements, not competitors. One generates the research and the other assesses it.

A Precedent for Building New Analytical Tools

The need to develop new ways of thinking about strategy at the dawn of the nuclear era demonstrates both what we can learn from analogies and how they could lead us astray. As Maas (2023) documented, nuclear, biotech, cyber, and aviation comparisons can mislead if they carry governance intuitions that do not match AI's technical and political structure.[35] The nuclear analogy is loaded, and using it carelessly does more harm than good. Invoking nuclear strategy implies to many that AI poses extinction-level risk. That is not the argument here. The analogy is to the process by which a novel strategic problem was met with sustained institutional investment in new analytical tools, not to the magnitude or character of the threat itself.

When the United States first confronted nuclear weapons in the late 1940s, it had no framework for thinking about the strategic implications. A technology that promised unlimited energy alongside a weapon that could destroy a city in a single strike, and that would soon be held by an adversary, presented a genuinely novel problem. Early postwar analysts argued that the very purpose of armed forces shifted from achieving victory to preventing catastrophe.[36] The initial policy responses were either naive (the Baruch Plan's proposal to hand nuclear materials to an international authority) or reckless (advocacy of preventive nuclear war while the US held a monopoly).[37]

The response was an institutional investment in building new intellectual tools. The RAND Corporation, spun off as an independent nonprofit in 1948, became a focal point for strategic analysis blending technical modeling, operational research, and policy-facing theoretical innovation. Von Neumann pushed game-theoretic thinking into defense strategy. RAND scientists formalized the prisoner's dilemma. Wohlstetter's (1959) basing studies operationalized second-strike stability.[38] Brodie (1945) framed deterrence as a historically novel posture.[39] Kahn (1965) popularized escalation ladders.[40] Schelling's (1960) *The Strategy of Conflict* treated force as a means of influence through signaling and coercive diplomacy.[41]

Over roughly fifteen years, this effort produced a conceptual vocabulary that made nuclear governance possible. This vocabulary included deterrence, second-strike capability, crisis stability, mutually assured destruction, and escalation ladders. These were not obvious ideas. They were the product of sustained, interdisciplinary work by people who understood that the

[35] Matthijs Maas, *AI Is Like… A Literature Review of AI Metaphors and Why They Matter for Policy*, AI Foundations Report no. 2 (Institute for Law & AI, 2023).
[36] Bernard Brodie, *The Atomic Bomb and American Security* (RAND Corporation, 1945).
[37] Fred Kaplan, *The Wizards of Armageddon* (Stanford University Press, 1983).
[38] Albert Wohlstetter, "The Delicate Balance of Terror," *Foreign Affairs* 37, no. 2 (1959).
[39] Brodie, *The Atomic Bomb and American Security*.
[40] Herman Kahn, *On Escalation* (Praeger, 1965).
[41] Thomas Schelling, *The Strategy of Conflict* (Harvard University Press, 1960).

existing intellectual toolkit was inadequate. The NPT, signed in 1968, was built on this foundation.[42]

But the true story is not purely the romantic one about math and social science riding to the rescue. Any honest accounting of RAND must confront its failures. Abella (2008) documented that RAND's rationalist framework had a fundamental blind spot by assuming human behavior was predictable and rational, which led to certain catastrophic miscalculations, most notably in Vietnam.[43] RAND wielded enormous influence without democratic oversight, and the accountability gap persisted throughout its most productive period. Schlosser's (2013) account of nuclear weapons management reveals a further gap between the elegant strategic theories and the messy, accident-prone reality of actual operations, where safety was achieved more through engineering improvisation and near-miss learning than through the top-down frameworks RAND produced.[44] Samaan's (2012) institutional analysis shows how RAND's model degraded over time through federal funding dynamics, the revolving door, and loosening ties to academia.[45] Finally, of course, we will never know the counterfactual: whether these concepts and ensuing equilibrium would have emerged even without RAND.

The intellectual products also had structural limitations that their creators understood imperfectly. Schelling's insights depended on a strategic landscape with particular features, including two principal actors, roughly observable capabilities, physical irreversibility, and identifiable escalation dynamics. The mathematical tools of the era reinforced this orientation. Game theory in the 1950s was largely limited to two-player interactions. The strategic concepts that emerged from RAND were brilliant, but they were built for a bilateral world because that was the reality and the limit of what the theory could handle.[46]

AI is fundamentally different. It involves not two superpowers but dozens of state actors, thousands of private companies, open-source communities, and individual researchers. Capabilities are difficult to observe. AI "attacks" are more ambiguous, harder to attribute, and in some cases more reversible than nuclear strikes. There is no clear escalation ladder, no enrichment threshold, and no separable "peaceful" track. Scharre (2023) noted that banning AI the way you could ban chemical weapons is impossible.[47] And the dual-use problem is far more entangled since the same model that accelerates drug discovery can also generate instructions for synthesizing dangerous compounds.

None of this means Schelling's insights are irrelevant. Mixed-motive dynamics clearly exist. The US and China are competing intensely but both have an interest in avoiding catastrophic AI failures. Some norms may emerge without formal agreements. The question is whether these concepts can be adapted to a multilateral, low-observability, software-defined strategic landscape, or whether entirely new concepts are needed. Today's analytical toolkit, including

[42] Lawrence Freedman, The Evolution of Nuclear Strategy, 4th ed. (Palgrave Macmillan, 2019).

[43] Alex Abella, *Soldiers of Reason: The RAND Corporation and the Rise of the American Empire* (Harcourt, 2008).

[44] Eric Schlosser, *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety* (Penguin, 2013).

[45] Jean-Loup Samaan, *The RAND Corporation (1989-2009): The Reconfiguration of Strategic Studies in the United States* (Palgrave Macmillan, 2012).

[46] Thomas Schelling, *Arms and Influence* (Yale University Press, 1966).

[47] Or, as he more colorfully put it, "'no AI' lacks the clarity of 'no gas.'" Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence* (W.W. Norton, 2023).

computational game theory, agent-based modeling, mechanism design for large populations, and network theory, offers tools for analyzing multilateral strategic interactions that did not exist in the 1950s. The problem of AI governance may be harder than nuclear strategy in important respects, but the analytical toolkit available is vastly more capable.

Institutional Principles

The literature on institutional design for sustained technology governance research identifies several requirements.

First, scientists and policymakers must find the work credible. Haas's (1992) foundational work shows how networks of experts can shape international policy coordination by framing issues and providing authoritative knowledge that reduces uncertainty for decision-makers.[48] Guston's (2001) concept of "boundary organizations" argues that institutions mediating between science and policy must be accountable to both domains and produce outputs credible to technical experts and legible to policymakers.[49] The proposed institution's outputs must be usable by both communities, and capture by either political principals or commercial stakeholders would destroy its credibility.

Second, it should strive for procedural independence while recognizing that true independence is impossible. McCubbins, Noll, and Weingast's (1987) work on administrative procedures as instruments of political control shows that agencies and quasi-public bodies are rarely truly independent, even from the beginning. Political principals shape them through oversight and structural design,[50] and this influence is strongest at the moment of creation, when founding actors embed their preferences into governance structures that persist long after they step back. For the proposed institution, this means the founders' choices about board composition, appointment mechanisms, and research mandate will shape the institution's trajectory more than any subsequent declaration of independence. Transparency requirements, conflict-of-interest constraints, and mechanisms for external challenge can shift oversight to broader audiences and reduce the risk that any single principal quietly steers research outputs. But these protections must be designed with full awareness that the founders themselves are the first principals whose influence the structure needs to constrain.

Third, the institution must be resilient. A research program that requires a decade or more to produce foundational frameworks must be designed to survive changes in political leadership, funding cycles, and shifts in public attention. The history of technology governance institutions suggests this is the hardest design requirement to meet because the forces that erode institutional independence operate on shorter time horizons than the research itself.

The institution would need to bring together economists, computer scientists, game theorists, international relations scholars, military strategists, and AI safety researchers. It would need the

[48] Peter Haas, "Introduction: Epistemic Communities and International Policy Coordination," *International Organization* 46, no. 1 (1992).
[49] David Guston, "Boundary Organizations in Environmental Policy and Science," *Science, Technology, & Human Values* 26, no. 4 (2001).
[50] Matthew D. McCubbins et al., "Administrative Procedures as Instruments of Political Control," *Journal of Law, Economics, and Organization* 3, no. 2 (1987): 243–77.

freedom to pursue questions without predetermined answers, the institutional patience to sustain inquiry over a decade or more, and connections to government and industry without being captured by either.

The Problems of Existing and Proposed Governance Institutions

The potentially enormous economic and societal effects of AI have brought some equally enormous proposals, particularly the growing number of "CERN for AI" proposals that have appeared since 2023. Kohler's (2025) taxonomy cataloged fourteen such proposals in four clusters, ranging from large-scale shared compute infrastructure to international safety testing facilities.[51] The most detailed, from the Centre for Future Generations, envisions an institution with a €35 billion budget and 3,000 staff.[52] The CERN framing implies that AI governance requires large shared physical infrastructure, which invites political battles over location, funding shares, and operational control.[53]

Appendix B surveys existing institutions that address pieces of this challenge, assessing their funding structures, independence mechanisms, and mandates. Organizations like Georgetown's CSET, Oxford's GovAI, and parts of RAND itself are doing valuable work on pieces of this problem. But they are fragmented efforts across institutions with different missions and funded on short grant cycles, meaning they lack the institutional mandate for the kind of sustained, integrated research program the problem requires.

No existing body combines sustained foundational research, technical capacity, structural independence, stable long-term funding, and the analytical ambition to build strategic frameworks comparable to what RAND built for nuclear strategy. The inventory also reveals why the gap persists, through several distinct failure modes that any new institution would need to be designed to avoid.

Governments can and do sustain serious analytical institutions for decades. But government agencies face two general pitfalls. New ones are politically fragile, as the recent fate of national AI safety institutes illustrates. The U.S. AI Safety Institute, housed at NIST and chronically underfunded at $10 million,[54] was dismantled and repurposed within eighteen months of its creation. In June 2025, it was renamed the Center for AI Standards and Innovation, completely transforming its mission from safety to deregulation, with the Commerce Secretary publicly referring to its prior mission as "censorship." The UK AI Safety Institute suffered a similar fate; despite a £100 million budget and top-tier technical talent, it was renamed the AI Security

[51] Kevin Kohler, *CERN for AI: One Analogy, Many Visions* (Simon Institute for Long-Term Governance, 2025).

[52] *Building CERN for AI: An Institutional Blueprint* (Center for Future Generations, 2025), https://cfg.eu/building-cern-for-ai/.

[53] Anna-Lena Rüland, "'We Need a CERN for AI': Organized Scientific Interests and Agenda-Setting in European Science, Technology, and Innovation Policy," *Minerva*, ahead of print, March 17, 2025, https://doi.org/10.1007/s11024-024-09568-6.

[54] Chuck Schumer, *Majority Leader Schumer Announces First-of-Its-Kind Funding to Establish a U.S. Artificial Intelligence Safety Institute* (2024), https://www.democrats.senate.gov/newsroom/press-releases/majority-leader-schumer-announces-first-of-its-kind-funding-to-establish-a-us-artificial-intelligence-safety-institute; Sharon Goldman, "Biden Appoints AI Safety Institute Leaders as NIST Funding Concerns Linger," *Venture Beat*, February 7, 2024, https://venturebeat.com/ai/biden-appoints-ai-safety-institute-leaders-as-nist-funding-concerns-linger.

Institute in February 2025, signaling a narrowed scope. The international AISI network is only as durable as its most fragile node, and two of its anchor nodes have already degraded.

Ones that survive long enough to become entrenched are nearly impossible to reform or shut down, develop their own institutional interests and constituencies, and lose independence as they become subject to the oversight dynamics of appropriations and appointments. Standing up such an organization before we even know how to best define "governance" is unlikely to yield conceptual advances.

University affiliation carries intellectual prestige but can be subject to changing university priorities. Oxford's Future of Humanity Institute was arguably the most influential AI governance research group in the world, with annual funding that grew from roughly £1 million to several million pounds, largely from philanthropic sources like Open Philanthropy and the Musk Foundation.[55] Yet the university closed it in April 2024. It died not because the work was poor or the funding disappeared, but because an interdisciplinary center with external funding and a public profile that exceeded its host department's did not fit the governance structures built for managing faculty and degree programs. CSET at Georgetown remains productive but carries the same structural vulnerability, compounded by extreme funding concentration in a single philanthropic source. The proposed institution should be freestanding, with its own board and governance, able to maintain university affiliations without being subject to university politics. RAND itself was structured this way from the start.

Industry-funded multi-stakeholder bodies gravitate toward consensus rather than rigorous independent analysis. The Partnership on AI, founded and substantially funded by the tech companies it nominally oversees, has produced useful public goods like the AI Incident Database. But it ultimately functions as a very well-funded conversation. It produces consensus guidance, not the kind of independent, rigorous work that makes its funders uncomfortable. That gravitational pull toward consensus is structural, not a failure of leadership. Any institution whose funding depends on the goodwill of the organizations it studies will face it.

Funding concentration is as dangerous as government dependence, even when the funder's intentions are good. CSET's reliance on Open Philanthropy, the Oversight Board's sole link to Meta, and the IAEA's sensitivity to voluntary contribution dynamics all illustrate the same point: a single funding source creates a single point of institutional failure. Some institutions have engineered partial solutions. The Meta Oversight Board achieved structural separation through an irrevocable $280 million trust. The Ada Lovelace Institute started with a grant of £5 million over five years. But the most resilient approach would combine diversified funding streams with structural protections on each, rather than relying on any single mechanism. Independence is an engineering problem, not a declaration. The institutions in the inventory that achieved genuine independence did so through specific legal and financial structures. The ones that lost it had independence as an aspiration rather than a design specification.

[55] Open Philanthropy, *Future of Humanity Institute — General Support* (2016), https://www.openphilanthropy.org/grants/future-of-humanity-institute-general-support/; University of Oxford, *£13.3m Boost for Oxford's Future of Humanity Institute* (2018), https://www.ox.ac.uk/news/2018-10-10-£133m-boost-oxfords-future-humanity-institute.

The subtlest way an organization can fail is when its mandate drifts toward tractable, but arguably less consequential, questions. The Facebook Oversight Board reviews individual content decisions because that is concrete and achievable. The OECD monitors because monitoring is what the OECD does. National AI Safety Institutes test models because testing is measurable. Every institution in the inventory did what it could measure and report on rather than what the problem may actually have required. The proposed institution would need to define its mandate by the analytical gap rather than by what is immediately feasible, and then resist the gravitational pull toward tractability from the start. This is probably the hardest institutional design problem the founders would face, and no clean solution exists. But naming the problem is a precondition for managing it.

External threats are not the only danger. Research institutions also fail from within, in ways that are harder to detect because they can look like success. Without a clear organizing problem, researchers left to their own judgment often cannot distinguish important questions from interesting ones. The orienting objective of nuclear strategy was to "minimize the probability of catastrophic war," which disciplined every research program RAND undertook. AI governance has no equivalent, and absent one, an institution of smart people will produce smart work on questions that may not matter.

The interdisciplinary mandate compounds this risk. Interdisciplinary research is inherently difficult to do well. Rhoten's (2004) study of six university-based interdisciplinary research centers found that most functioned as "loosely connected individuals searching for intersections, not cohesive groups tackling well-defined problems," despite adequate funding and genuine researcher motivation.[56] Interdisciplinary work fails when it yields economists, computer scientists, and international relations scholars each contributing a diluted version of their discipline and the synthesis is weaker than any individual contribution would have been.

Avoiding this requires recruiting at the top of every contributing discipline and maintaining standards that make second-rate interdisciplinary work unpublishable, not just unfashionable. That is easier to prescribe than to achieve.

The Proposal

Proposing a new institution is often a substitute for thinking clearly about a problem. The case here is different. The institution's purpose is not to govern but to determine whether, when, and how governance is warranted.

Its output would be analysis, not authority, conducted by a lean analytical research body, closer to the Institute for Advanced Study than to CERN. It does not need a particle accelerator or a shared compute cluster. It needs fourteen scholars and the institutional patience to let them work. It has to be durable enough to sustain a decade of foundational research, independent enough to produce uncomfortable findings, and structured so it can be wound down if it stops being useful.

Notably, the institution should not need access to proprietary models, compute clusters, or training data. Los Alamos built the bomb and RAND figured out how to live with it. Schelling

[56] Diana Rhoten, "Interdisciplinary Research: Trend or Transition," *Items and Issues* 5, nos. 1–2 (2004).

and Wohlstetter did not need enriched uranium to develop deterrence theory. They needed to understand the macro-parameters of the technology: destructive capacity, delivery speeds, and the economic incentives of the actors involved. Similarly, a new institution does not need to audit neural weights or inspect training runs. That is application-level safety work, appropriate for internal red teams, AI safety institutes, or engineering bodies. The goal is mapping the strategic landscape: the dynamics of competition, the conditions under which governance interventions help or hurt, and the verification mechanisms that are technically feasible and politically acceptable. The central questions are about actors, incentives, and institutions, not about what any particular model can do. Someone evaluating frontier model capabilities is doing important work, but they are answering a different question than whether a governance regime will hold when the actors subject to it find compliance commercially inconvenient.

Untethering macro-strategic research from physical model access avoids the trap of relying on corporate goodwill and bypasses intractable battles over intellectual property entirely. Depending on goodwill and access creates a dynamic that pulls multi-stakeholder bodies toward consensus rather than independent analysis.

The institutional economics literature on durable institutions finds that the most successful begin with a narrow but valuable function, produce shared methods and shared facts, and create increasing returns as participants build internal routines around the institution's outputs.[57] Once regulators, insurers, procurement offices, and international coordination forums embed the institution's frameworks into their own processes, switching costs rise and other organizations depend on rather than merely consult the institution.

But durability also creates vulnerability. Noll's (1989) account of regulation emphasizes that the same embeddedness that makes institutions persist can steer them toward visible, low-impact outputs that satisfy overseers rather than solve the underlying problem.[58] Shirley's (2005) synthesis of institutional reform research warns that formal blueprints often fail at implementation because incentives, monitoring capacity, and distributional conflict determine what an institution actually does, not what its charter says.[59] That path dependence is how a small research institution becomes durable, but only if durability is designed rather than assumed.

The Research Agenda

A sustained research program would need to address questions that the current literature identifies but cannot yet answer as well as to identify new questions we have not yet been able to pose.

In practice, the institution’s first task would be to establish a shared strategic grammar for AI development dynamics. Nuclear strategy had “deterrence,” “first strike,” “crisis stability.” AI governance has nothing equivalent. Policymakers, industry leaders, and researchers use the same

[57] Douglass C. North, *Institutions, Institutional Change and Economic Performance* (Cambridge University Press, 1990).
[58] Roger Noll, “Economic Perspectives on the Politics of Regulation,” in *Handbook of Industrial Organization*, with Richard Schmalensee and Robert Willig (North-Holland, 1989).
[59] Mary M. Shirley, *Institutions and Development*, ed. Claude Ménard and Mary M. Shirley (Springer, 2005), 611–38.

words to describe fundamentally different concerns, and the result is that every summit, every proposal, and every international declaration talks past itself. The Bletchley-to-Paris trajectory illustrates this directly. Countries did not converge on answers because they had not converged on questions. A credible research program would begin by defining the core problems with enough precision that disagreements become productive rather than circular. What are the specific mechanisms by which frontier AI development creates risks that market incentives and domestic regulation cannot address?[60] Under what conditions does international coordination improve outcomes versus simply adding transaction costs? Which risks are artifacts of the current moment versus structural features that will persist as the technology diffuses?

Getting those questions right is not a preliminary step. It is the central intellectual contribution. Recent work on AI and strategic stability suggests that AI's impact on competition, escalation, and stability is partly shaped by the analytical frameworks through which states understand the technology.[61] If so, framework-building is itself a strategic act, not merely an academic one. Everything else follows from clearly stating the right questions. How successful have different governance regimes been? The empirical literature reviewing the impacts of various governance regimes on technological development is almost nonexistent. This research would not look simply at the effects of AI rules, but on a number of governance regimes across a number of technologies and history.

Can partial verification support partial governance? You cannot inspect a training run from orbit. But compute monitoring through cloud providers, energy consumption signatures of large training runs, and chip tracking through the supply chain offer partial verification mechanisms. Baker's (2023) analysis argues that with preparation, verification challenges for AI agreements could be reduced to levels that nuclear arms control managed, especially for hardware inputs and large-scale training.[62] The Oxford verification report catalogs further approaches.[63]

Comprehensive verification was not achievable for nuclear weapons either. Nuclear safeguards did not rely on trust alone; verification made certain agreements feasible that would otherwise have been impossible. The question is whether partial verification can support partial governance for AI, and the answer is not yet known.

What does "strategic stability" mean in a multilateral, software-defined landscape? Nuclear crisis stability meant avoiding configurations where either side had incentive to strike first. The equivalent for AI is unclear when capabilities can be copied and deployed in hours, when offense and defense are blurred, and when actors include non-state organizations. Analysis by the Stockholm International Peace Research Institute on AI and strategic stability emphasizes practical confidence-building measures, and separate academic work examines how AI-enabled

[60] The range of relevant mechanisms extends beyond safety and labor displacement. Acemoglu, Kong, and Ozdaglar (2026) model how agentic AI can erode human learning incentives by substituting for the costly effort that jointly produces private signals and public knowledge, potentially undermining the long-run information ecosystem that governance institutions need. Daron Acemoglu et al., "AI, Human Cognition and Knowledge Collapse," NBER Working Paper 34910, 2026.

[61] See, e.g., Vincent Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (Stockholm International Peace Research Institute (SIPRI), 2020).

[62] Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties."

[63] Harack et al., *Verification for International AI Governance*.

capabilities can increase escalation risk through entanglement of nuclear and non-nuclear systems.[64]

When does competitive development itself produce safety? Competition and openness create risk, but they also generate safety properties that centralized governance might not. Redundancy across developers means no single point of failure. Open-source models enable independent security auditing. Rapid iteration can surface vulnerabilities faster than centralized review. A rigorous analysis must account for the possibility that some risks are better managed by the competitive process itself.

Under what conditions do governance interventions make things worse? Export controls can accelerate indigenous chip development. Safety requirements can concentrate development among large firms. Liability regimes may drive development offshore. International agreements can create false confidence in unverifiable commitments. The history of technology regulation is full of interventions whose costs exceeded their benefits. Any analytical framework must take seriously the possibility that the best response to some risks is adaptation, resilience, or simply accepting a level of risk as the cost of the technology's benefits.

Which risks are governable at the development level, and which require resilience? Some risks, particularly those posed by individuals or small groups using widely available models, may not be governable through any development-level framework. For those, the relevant question shifts to detection tools, defensive infrastructure, and response capacity. Distinguishing which risks fall into which category is a central task. As the literature notes, effective governance increasingly focuses on frontier development chokepoints, especially compute, evaluation, and model artifact controls, because these are among the few levers where verification might be feasible without requiring comprehensive control of the entire digital ecosystem.[65]

How should governance evaluation become empirical rather than aspirational? We have almost no rigorous evidence on whether specific AI governance interventions causally affect development choices and risk outcomes. The OECD's Hiroshima Process reporting is a start, but early evidence shows uneven compliance correlated with firm size and timing rather than genuine norm internalization.[66] Building the capacity to evaluate governance interventions with the same rigor applied to economic or health policy is itself a research program.

How will we know if the institution is succeeding? The institution should treat evaluation of its own usefulness as part of the research program, not as a separate accountability exercise. We do not yet know what successful AI governance research looks like. If the institution produces a

[64] Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk*; James M. Acton, "Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War," *International Security*, Summer 2018, 56–99.

[65] See Egan and Heim (2023) on compute providers as governance chokepoints; Reuel et al. (2024) on technical governance tools including evaluation and model artifact controls; Baker (2023) on verification feasibility for hardware inputs. Mueller (2025) argues that governance ambitions beyond these chokepoints imply control over the entire digital ecosystem. Janet Egan and Lennart Heim, *Compute Governance and AI Safety*, 2023; Reuel et al., "Open Problems in Technical AI Governance"; Mauricio Baker, *Verification Methods for AI Agreements*, 2023; Milton Mueller, "It's Just Distributed Computing: Rethinking AI Governance," *Telecommunications Policy* 49, no. 3 (2025), https://doi.org/10.1016/j.telpol.2025.102917.

[66] Perset and Fialho Esposito, *How Are AI Developers Managing Risks? Insights from Responses to the Reporting Framework of the Hiroshima AI Process Code of Conduct*.

formal model of AI competition dynamics and no policymaker uses it, that could mean the model was wrong, or that the problem did not require formal modeling, or that the translation from research to policy failed. These are different failures requiring different responses, and distinguishing among them is itself an analytical task. The institution should commit from the start to asking whether its work is changing how anyone thinks about a specific problem, and to publishing honest answers. If after five years the research has not demonstrably informed a single policy decision, regulatory design, or corporate governance choice, that is evidence the design is wrong and should be revised or abandoned. The institution must also resist the temptation to evaluate itself on its own terms. Any organization asked whether its work matters will find reasons to say yes. External review by people with no stake in the institution's survival or in receiving the next evaluation contract is the only credible check, and the founders should build that in structurally rather than leaving it to good intentions.

What It Would Produce

The research agenda above lists potential input questions. The institution should produce concrete outputs that downstream actors use. The list of possibilities is long, but consider the following:

- An annual assessment of whether frontier safety frameworks actually constrain behavior, applying rigorous methodology systematically across all published frameworks rather than leaving evaluation to ad hoc academic efforts.
- A library of formal models for AI competition dynamics, analogous to the arms race stability models RAND maintained during the Cold War, updated as competitive conditions change.
- Structured scenario analyses that regulators and legislators can use, separating what is known from what is uncertain from what is genuinely contested.
- Post-incident analysis when governance failures occur, producing shared facts rather than competing narratives.

Over time, the shared strategic grammar would let policymakers, researchers, and developers reason about the same problems using common concepts. None of these outputs requires access to proprietary model weights or training data. All of them require sustained institutional capacity that does not currently exist.

How It Would Be Structured

The founding director would need to be someone with a rare combination of genuine intellectual standing in at least one of the contributing disciplines, credibility with both the technical AI community and the policy world, and the institutional temperament to build something that outlasts their own tenure. This is not a job for a convener or a manager. It is a job for someone who can look at the work being produced and know whether it is good enough. RAND’s most productive period was shaped not by its organizational structure but by the quality of the people it attracted and the judgment of the leaders who recruited them.

The founding scholars matter as much as the founding director. The first five hires would define the institution’s intellectual identity for a decade or more. They should be drawn from at least three disciplines, should have demonstrated the ability to produce work that changes how other researchers think, and should be willing to commit at least three years of primary effort. The temptation will be to recruit established names for legitimacy. The better strategy is to recruit people doing the best work on the actual questions, regardless of seniority, and let the quality of early output build the institution's reputation.

How research reaches policymakers matters for impact as much as what the research says. The standard think tank model, where researchers produce technical work and a separate communications team "translates" it into simplified derivatives, almost invariably strips out the nuance and caveats that made the work valuable. A better model invests in editorial infrastructure rather than communications staff. A senior editor who works directly with researchers on the structure and clarity of their major outputs, not producing simplified summaries but editing the research itself, ensures that a working paper can be read directly by a policy advisor without passing through a layer of dilution. This also means hiring for analytical brilliance need not be constrained by communication skills. Some of the most valuable researchers will be poor communicators. Editorial infrastructure ensures their work reaches policymakers regardless.

An open question is whether the institution should begin by looking at the issue from one country's or group's perspective. As a starting point, we could assume that the frameworks it develops cannot serve only one side of a competition. AI governance is not exclusively a great-power problem. Countries across the Global South, from India to members of the African Union, have distinct governance interests that are not reducible to keeping pace in a technology race, and any durable framework will need to account for them. However, we should remain open-minded even on that. RAND produced strategic frameworks from within one side of a bilateral competition, and those frameworks proved useful to both sides because they clarified dynamics all parties faced. It is possible that rigorous analysis from any institutional base would produce insights with broad applicability. The question of institutional composition is itself a design problem the founders would need to resolve.

A Structural Precedent

The Santa Fe Institute offers a structural precedent. Founded in 1984 by Los Alamos scientists, including Murray Gell-Mann and George Cowan, who wanted to work on complex problems outside disciplinary boundaries, SFI has survived four decades as a freestanding, independently governed research institution. It maintains a small resident faculty supplemented by a large network of affiliated researchers whose primary appointments remain elsewhere, which solves part of the talent problem without requiring people to abandon existing positions. Its Applied Complexity Network brings corporate partners who pay for access to ideas without directing research, with explicit rules protecting the scientific agenda from funder influence.

SFI demonstrates that the institutional form this paper proposes—independent, interdisciplinary, privately funded, outside both government and universities—can be built and sustained. But SFI's own operating principles state that they have "in general avoided becoming involved in matters of policy. But if you are working on a program that involves sustainability or the environment or human welfare and you think we might have something you can use pick up the phone."[67] Its mission is fundamental science, and its culture actively resists being instrumentalized for governance purposes. The proposed institution would need SFI's structural independence with a fundamentally different orientation toward producing frameworks that policymakers can actually use. That combination, rigorous independence with policy-facing output, is what no existing institution has achieved.

SFI's model also suggests that internal organization matters as much as institutional form. SFI does not organize its resident faculty into managed research programs with directors and subordinates. Scholars are individually appointed, pursue their own research agendas, and collaborate when the intellectual fit is there. The most productive work emerges from proximity and shared questions, not from top-down direction. The people who produced the foundational insights in nuclear strategy were not working under program directors. Schelling, Wohlstetter, Brodie, and Kahn were peers pursuing their own questions within a loosely shared intellectual space. An institution designed to produce comparable insights for AI governance should be organized similarly: a community of independently appointed scholars working within a set of orienting questions, not a hierarchy of managed research teams.

Physical location probably also matters. The pattern among institutions designed for sustained, foundational thinking is that they locate where the environment supports concentration, not access. SFI is in Santa Fe. RAND was founded in Santa Monica. The Institute for Advanced Study is in Princeton. The Salk Institute is in La Jolla. Locating a research institution in Washington or Brussels would embed it in the ecosystem it is supposed to analyze from the outside. Every hire would carry networking expectations. Proximity to political principals is the dynamic that McCubbins, Noll, and Weingast's work warns about, and that Samaan documents as a factor in RAND's own degradation over time. Geography enforces separation from power in ways that conflict-of-interest policies alone cannot.

Why This Might Be Wrong

A serious literature argues that poorly designed AI governance can backfire, and any research program worth funding must take this literature seriously rather than assume governance is the answer. Scholars make several arguments.

First, regulation and governance always risk increasing concentration. Lancieri's (2024) work on AI regulation, competition, and regulatory capture shows that heavy development regulation can raise barriers to entry, favor incumbents, and encourage regulatory arbitrage across jurisdictions.[68] Korinek and Vipra (2025) add that foundation models generate concentration dynamics, and that concentration interacts with safety in complex ways by allowing internalization of some externalities while also creating systemic fragility and increasing the risk

[67] Cormac McCarthy, *Operating Principles* (Santa Fe Institute, 2015), https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/ckeditor/2018/04/24/cormac-sfi.jpg.

[68] Filippo Lancieri et al., "AI Regulation, Competition, Arbitrage, and Regulatory Capture," *Georgetown Law Faculty Publications and Other Works 2647*, November 2024, https://scholarship.law.georgetown.edu/facpub/2647/.

of regulatory capture.[69] Incumbent interests can leverage safety discourse to entrench themselves if regulation becomes aligned with their interests.

Second, some governance objectives may simply not be feasible. Mueller (2025) argues that many headline proposals to govern AI, especially through compute control, imply effective control over distributed computing as a whole.[70] "Governing AI" can collapse into governing the entire digital ecosystem, including data, networks, and software, which may be politically unacceptable or practically impossible across borders. This line of criticism suggests that certain governance ambitions fail not because coordination is hard in general, but because the object of control is too entangled with general-purpose digital infrastructure.

Export controls on AI chips illustrate the difficulty. The US has imposed escalating controls since 2022, treating the concentrated hardware supply chain as a chokepoint analogous to fissile material control.[71] In reality, enforcement has been limited. Operation Gatekeeper in December 2025 revealed a $160 million chip smuggling network. At least $1 billion in restricted chips have reached China since April 2025, at 50% black-market premiums.[72] Meanwhile, algorithmic efficiency improvements progressively weaken the relationship between compute access and AI capability.[73]

Export controls are not worthless. They raise costs and buy time, which may be the most we can expect from tools designed for a different era. But as scholars of dual-use governance have noted, they create uncertainty for allies and firms, provoke countermeasures, and their durability depends on political coalitions that shift with administrations.[74] Stopgaps are worth maintaining, but they are not substitutes for the analytical frameworks needed to design durable governance, any more than ad hoc nuclear postures in the early 1950s were substitutes for the deterrence theory that eventually stabilized competition.

The innovation costs are uncertain but plausible. Precautionary approaches can suppress beneficial experimentation and slow the discovery of safety improvements that come from real-world feedback. Over-constraining frontier development could reduce the variety of approaches, centralize control, and slow safety learning, yielding worse long-run outcomes even if short-run risks are reduced.

And open versus closed development creates a genuine dilemma that the literature does not resolve. Heavy governance may push development into less transparent channels, while weak governance may enable unsafe proliferation. One plausible backfire mechanism works in each

[69] Anton Korinek and Jai Vipra, "Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence," *Economic Policy* 40, no. 121 (2025).
[70] Mueller, "It's Just Distributed Computing: Rethinking AI Governance."
[71] Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items, 87 Fed. Reg. 62,186 (Oct. 13, 2022).
[72] Zijing Wu and Eleanor Olcott, "Nvidia AI Chips Worth $1bn Smuggled to China after Trump Export Controls," *Financial Times*, July 24, 2025, https://www.ft.com/content/6f806f6e-61c1-4b8d-9694-90d7328a7b54?syn-25a6b1a6=1.
[73] *Export Controls: Commerce Department Should Improve Semiconductor Controls*, GAO-24-106417 (Government Accountability Office, 2024).
[74] Lancieri et al., "AI Regulation, Competition, Arbitrage, and Regulatory Capture."

direction. Distinguishing when governance helps from when it hurts is itself a central research task, not a question with a known answer.

Third, perhaps the governance problem does not require new analytical infrastructure at all. Some argue that recombining existing institutional patterns is the better path to AI governance rather than building from scratch. Trager et al. (2023) propose an international organization modeled on aviation and maritime governance, where compliance is incentivized through jurisdictional certification and market access rather than novel strategic frameworks.[75] Baker's (2023) work on verification argues that AI development, despite its differences from nuclear weapons, relies on physical inputs and supply chains that create enforcement hooks comparable to those managed under existing arms control regimes.[76] Aviation-style safety engineering, now being operationalized through national AI safety institutes and their international network, represents a third adaptation strategy that is already producing early results in pre-deployment evaluation. If the adaptation camp is correct, the investment this paper proposes is not merely premature but misdirected. The right response would be to strengthen execution of known institutional designs rather than to build new analytical capacity. This paper's premise is that adaptation is insufficient because the underlying strategic dynamics of AI development are not yet well enough understood to know which institutional patterns to adapt and how. But that premise is contestable, and the adaptation argument deserves to be taken seriously as an alternative to what is proposed here.

Klein and Patrick (2024) argue against a single new institution and instead propose multiple bodies each addressing different aspects of the problem, coordinating through peer review and mutual recognition rather than centralized authority.[77] This proposal for a "regime complex" draws on how governance actually works for other complex international challenges. Ostrom's (1990) work on polycentric governance reinforces the regime complex intuition because successful governance of shared resources often emerges from overlapping, semi-autonomous institutions rather than centralized authority.[78] Her framework also supports this paper's emphasis on shared analytical vocabulary, since she argued that polycentric systems work not because of decentralization per se but because of common principles that let diverse institutions coordinate without hierarchy.

But a regime complex is incomplete, not wrong. Distributed governance bodies need shared analytical foundations to coordinate effectively. Without a common strategic grammar and a shared evidence base on what governance interventions actually accomplish, the regime complex produces fragmentation rather than complementarity. The proposed institution is not a replacement for the regime complex. It is the research infrastructure the regime complex requires in order to function.

A reasonable extension of this argument is that the research infrastructure itself should be plural rather than singular, as it described here. Multiple institutions pursuing overlapping questions

[75] Trager et al., *International Governance of Civilian AI: A Jurisdictional Certification Approach.*

[76] Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties."

[77] Klein and Patrick, *Envisioning a Global Regime Complex to Govern Artificial Intelligence* (Carnegie Endowment for International Peace, 2024).

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

from different disciplinary and geographic starting points could produce more robust frameworks than any single body. But the field of strategic AI governance research barely exists. The talent pool is thin, the orienting questions are not yet well-defined, and the intellectual critical mass needed to produce foundational work requires concentration before it can support dispersion. Nuclear strategy did not emerge from a dozen competing institutes. It emerged primarily from one, and the concepts then diffused. The argument here is for building the first such institution, not the only one.

Finally, the simplest reading of the evidence presented so far is that such an institution is unnecessary. If you cannot define the problem, maybe there is not one.

Every governance analogy examined in this paper required a definable threat. Nuclear weapons could destroy cities. Chemical weapons killed soldiers in trenches. Even climate change, for all its complexity, could be measured in atmospheric CO2 concentrations and degrees of warming. AI governance has no equivalent. The risks are real but diffuse, entangled with enormous benefits, and dependent on development trajectories that no one can predict. It is possible that what looks like an analytical gap is actually the absence of a well-defined problem, and that the right response is not to build institutions to fill the gap but to accept that governance will be incremental, adaptive, and ad hoc because the technology demands it.

One does not need to believe that AI poses no risks to reach that conclusion. It is that AI may be the kind of problem that resists the institutional approach this paper proposes. Nuclear weapons were a discrete technology, held by a small number of state actors, with a binary catastrophic outcome. That structure rewarded systematic analysis and sustained institutional investment. AI is a diffuse, general-purpose technology with millions of actors and a continuous range of outcomes. Governance for the first kind of problem looks like RAND and the NPT. Governance for the second kind might look more like the internet, where order emerged from standards bodies, market competition, norms, and patchwork regulation rather than from any strategic framework developed in advance. The internet analogy cuts both ways. Order did emerge without a strategic framework, but so did governance failures that democracies have spent two decades struggling to correct. The question is whether repeating that experience is an acceptable outcome.

Each of these objections identifies a real risk. None of them eliminates the need for the analytical work this paper proposes, and most of them presuppose it. This paper cannot resolve that uncertainty. But it can acknowledge the expected value calculation honestly. The cost of building the proposed institution and discovering it was unnecessary is roughly $12 million a year, comparable to the Santa Fe Institute's current operating budget, supporting a community of fourteen resident scholars supplemented by a large visiting program (Appendix B provides an illustrative steady-state budget). The cost of not building it and discovering, five or ten years from now, that we needed foundational analytical frameworks and did not have them is potentially much larger, and much harder to recover from. That asymmetry does not make the investment obviously correct. It makes it a bet worth considering seriously.

The Case for Investing

The goal would not be to build a case for or against governing AI development. It would be to determine rigorously which risks require intervention, which do not, and what tools might work for those that do. A serious analysis might conclude that for some categories of risk, structural barriers to governance are so high that resources are better spent on resilience, redundancy, or adaptation, just as early nuclear strategists realized that building survivable infrastructure was more stabilizing than attempting to ban the bomb. It might conclude that competitive AI development, for all its risks, produces safety properties that centralized governance would not. Those would be valuable findings, not failures of the research program.

This is also not a substitute for governing AI applications domestically. Regulating deepfakes, algorithmic discrimination, healthcare AI safety, and sector-specific liability in ways that yield expected net benefits is important and does not require the kind of analytical infrastructure described here. Much of what AI does can be governed through existing frameworks for pharmaceuticals, financial products, vehicles, and consumer safety, adapted to the specifics of AI-enabled applications. Frontier labs' internal safety work, including dangerous capability evaluations and responsible scaling commitments, represents useful early experimentation. The OECD's reporting framework and the network of national AI safety institutes are steps toward institutionalizing evaluation capacity.[79]

But application-level governance cannot address questions about the trajectory of AI development itself, which unfolds across borders and outside any single jurisdiction's reach. That gap is real, and it will not be filled by continuing to rely on approaches whose effectiveness remains undemonstrated.

The current policy environment in the United States is not particularly conducive to this effort. Cuts to federal science funding and the dismantling of AI safety infrastructure at NIST reduce institutional capacity. The need for rigorous analytical frameworks does not depend on any particular administration's priorities, and the work could be housed outside the government entirely. The original RAND was independent of the agencies it served. An AI equivalent would need to be even more so.

A funding path may exist. The companies developing frontier AI have the resources to endow such an institution outright. At current valuations, the leading frontier labs are collectively worth hundreds of billions of dollars. A $250 million endowment, sufficient to fund the institution in perpetuity at roughly $12 million annually, would represent a negligible fraction of that value. An irrevocable endowment trust with an independent board[80] would be a structural, rather than aspirational, solution to the independence problem. Unlike ongoing corporate funding, which creates perpetual leverage, an endowment severs the connection between funder and institution from day one. The founding companies would have no mechanism to direct research, withdraw support, or retaliate against uncomfortable findings. There is also a collective-action logic. Each company individually faces competitive pressure to weaken safety commitments, as recent

[79] "International Network of AI Safety Institutes Mission Statement," National Institute of Standards and Technology, 2024.

[80] "Independent" is not well-defined and, as evidenced by changes at the U.S. Federal Communications Commission and Federal Trade Commission, can be challenged. Because this would not be a government agency, that specific assault on independence is not the risk, but other influences can affect independence. For this organization, board rules protecting independence should include no donor seats, no donor veto, no donor clawbacks, staggered trustee appointments, and public conflict rules.

developments have illustrated. But all companies are worse off if the absence of credible independent analysis leads to regulatory overreaction after a preventable failure, or to public loss of confidence in the technology itself. An endowed institution that produces rigorous, independent evidence about what governance works and what does not is a public good that no single company can credibly produce on its own. The companies developing frontier AI are better positioned than any other actors to create it, and the endowment model ensures they cannot capture it once created.

The practical path would begin with a small founding commitment, likely from two or three frontier AI companies and one or two major foundations, sufficient to establish a board of trustees, hire a founding director, and appoint the first four or five resident scholars. The founding board's first task would be to recruit a director with enough intellectual credibility to attract top-tier researchers and enough institutional judgment to resist the gravitational pull toward tractable but unimportant work. The director's first task would be to commission the initial research portfolio around two or three of the orienting questions described above, selected for their combination of analytical tractability and policy relevance. The endowment campaign would run in parallel with the first two years of operation, using early research outputs to demonstrate value to prospective donors. This is how SFI was built. Cowan and Gell-Mann started with a summer workshop, proved the model produced valuable work, then raised the endowment. The sequence matters: demonstrate intellectual seriousness first, then capitalize the institution, not the reverse.

The investment has not been made. It should be.

Appendix A: Inventory of AI Governance Institutions

Compiled March 2026. Reference document for Wallsten, "Building the Analytical Infrastructure for Governing Frontier AI Development."

Institution	Type	Founded	Funding Model	Annual Budget	Independence Structure	Core Function	Relevance to Proposed Institution	Sources
UNIVERSITY-HOUSED RESEARCH CENTERS								
CSET (Georgetown)	University research center	2019	Philanthropic (Open Philanthropy primary funder, plus Hewlett Foundation, PIT-UN)	$100M+ total through 2025	Housed at Georgetown SFS. Research agenda set internally. Publishes donor list. Nonpartisan mandate.	Data-driven policy analysis on AI, security, and emerging tech. Talent tracking, export controls, compute analysis.	Closest existing model for data-driven analytical work. But focused on security/intelligence applications rather than governance framework-building.	CSET press release (June 2021); Open Philanthropy grants database.
FHI (Oxford)	University research center	2005	Philanthropic (Open Philanthropy, Elon Musk, ERC, Future of Life Institute, Leverhulme Trust)	~£1M/yr pre-2018; 2018 grant of up to £13.3M (partly contingent on hiring)	Housed at Oxford Philosophy Faculty. Formally part of university structure.	Long-term existential risk research, AI safety, governance theory.	Closed April 2024 after Faculty froze fundraising (2020) and declined to renew contracts (2023). FHI attributed closure to administrative friction; some accounts also cite director controversies.	Sandberg, FHI Final Report, EA Forum (April 2024); Wikipedia; Open Philanthropy grants database.
FOUNDATION-FUNDED INDEPENDENT INSTITUTES								
Ada Lovelace Institute (UK)	Independent research institute	2018	Nuffield Foundation (£5M over five years). Independent of government and industry.	~£5M over five years plus ongoing Nuffield support	Legally independent. Nuffield Foundation provides structural independence. Own board and leadership.	Data and AI ethics research, public deliberation, policy analysis. UK and EU focused.	Good model for independence structure and honest-broker positioning. Wrong mission profile for foundational framework-building.	Nuffield Foundation announcement (2018); Ada Lovelace Institute website.
GOVERNMENT-CREATED INSTITUTES								
UK AI Safety Institute (now AI Security Institute)	Government directorate (DSIT)	2023	UK government public funding	£100M (~$127M)	Part of DSIT. Led by Ian Hogarth (chair). Recruited from OpenAI, DeepMind, Oxford.	Pre-release frontier model evaluation. Safety testing. Open-source evaluation tools (Inspect). Has tested 16+ models pre-release.	Demonstrates government can build serious technical capacity. Also demonstrates vulnerability to political shifts: renamed to AI Security Institute Feb 2025, signaling narrowed scope.	UK DSIT (2024); Wikipedia; TIME (Jan 2025).
US AI Safety Institute (now CAISI)	Federal agency program (NIST)	2023	US federal appropriation	$10M (chronically underfunded)	Housed at NIST. Subject to congressional	Originally: AI safety evaluation and testing. Now: standards-	Strongest cautionary example of government institutional fragility. Took <2 years to completely	Wikipedia; NIST documentation; Lutnick statement (June 2025).

Institution	Type	Founded	Funding Model	Annual Budget	Independence Structure	Core Function	Relevance to Proposed Institution	Sources
					appropriations and executive direction.	setting, risk identification.	reverse institutional purpose. Renamed June 2025.	
International AISI Network	Intergovernmental network	2024	Individual member government funding	Varies by node	Agreed at Seoul Summit May 2024. Members: UK, US, Japan, France, Germany, Italy, Singapore, South Korea, Australia, Canada, EU.	Coordinate safety testing approaches across national institutes. Share evaluation methods.	Illustrates both the appeal and fragility of government-led coordination. Only as durable as its most fragile node.	Seoul AI Summit communique (May 2024); NIST mission statement (2024).
INDUSTRY-FUNDED MULTI-STAKEHOLDER BODIES								
Partnership on AI	Industry-founded nonprofit	2016	Founded by Amazon, Meta, Google/DeepMind, Microsoft, IBM. Now 126+ partner organizations.	Not publicly disclosed	Independent 501(c)(3). Own board and CEO (Rebecca Finlay). But founded and substantially funded by companies it convenes.	Best-practice frameworks, synthetic media standards, AI Incident Database, convenings.	Demonstrates that industry multi-stakeholder bodies gravitate toward consensus rather than rigorous independent analysis.	Wikipedia; PAI website.
Meta Oversight Board	Corporate-funded independent body	2020	Irrevocable Delaware trust ($280M: $130M initial + $150M in 2022). Meta cannot claw back funds.	$280M trust, multi-year operations funded	Irrevocable trust structure. Own leadership, fixed terms, dedicated staff. Board members selected by trustees, not Meta.	Review individual content moderation decisions on Facebook/Instagram. Issue binding decisions and policy recommendations.	Best existing model for funding independence (irrevocable trust). Cautionary tale for mandate scope: limited to individual content decisions.	Oversight Board (July 2022); Oversight Board FAQ; Lawfare (Nov 2023).
INTERNATIONAL GOVERNANCE BODIES								
OECD AI Policy Observatory	Intergovernmental body	2019	OECD member state contributions	Part of OECD budget	Operates within OECD institutional structure. Consensus-based among member states.	AI Principles (2019). Monitoring and reporting. Hiroshima Process AI reporting. Policy coordination.	Good monitoring and coordination model. Not designed for foundational analytical research.	OECD AI Principles (2019); OECD Hiroshima Process framework.
IPCC (climate analogy)	Intergovernmental panel	1988	UNEP and WMO sponsorship. Government contributions.	~$6–8M/yr (leverages unpaid expert time)	Independent scientific assessments. Structured review process. Government approval of summaries only.	Assess climate science for policymakers. Synthesize research. Periodic assessment reports.	Model for procedural independence and scientific legitimacy (35+ years). But assessment-synthesis model too slow for AI and does not generate original research.	IPCC website; IPCC Trust Fund financial statements.

Appendix B: Illustrative Steady-State Budget

This appendix provides a bottom-up estimate of annual operating costs for the proposed institution at steady state (Year 5+). The budget is organized around the research environment and institutional design requirements described in the main text. All figures are in 2026 dollars and represent annual operating costs, not startup or capital expenditure.

Research

Resident Scholars

10 senior scholars + 4 postdoctoral researchers | $3.0M annually

Ten senior scholars, each individually appointed, form the permanent intellectual core. These are people at the level of a tenured professor or senior research fellow at a top institution, hired for the quality and relevance of their thinking, not to fill a slot in an organizational chart. Compensation of $200,000 to $300,000 fully loaded is competitive with top policy research institutions but below what industry or some university endowed chairs offer. Non-monetary benefits include independence, interesting colleagues, no teaching or committee obligations, and work on the most consequential policy question of the era.

Four postdoctoral researchers provide junior capacity. They work with senior scholars on specific projects but also pursue their own research. The postdoc program produces research output and identifies the next generation of scholars for eventual senior appointment, either at this institution or elsewhere.

Visiting Scholars Program

$1.5M annually | 20–30 visitors per year | No permanent FTEs

With fourteen permanent researchers, the institution cannot cover every methodological angle or substantive domain on its own. The visiting program provides the breadth that a small permanent community cannot.

Visitors receive stipends of $50,000 to $75,000 for residencies of three to six months, plus travel and housing support. A game theorist comes for a semester to work on a specific model of multilateral AI competition. An arms control specialist visits for four months to collaborate on verification design. An economist spends a summer building an empirical evaluation of export control effects. The permanent scholars provide continuity and institutional memory. The visitors provide range.

The visiting model also addresses the talent recruitment challenge. A tenured professor does not need to give up tenure to spend a semester at the institution. A RAND analyst does not need to leave RAND to contribute. The institution gets the talent without requiring the career sacrifice that deters senior people from joining unproven organizations. Visitors also carry the institution's frameworks back to their home departments and agencies, extending intellectual influence far beyond the permanent staff.

Computing and Data

$300K annually | No permanent FTEs

The budget does not include dedicated computational infrastructure or a data engineering team. The research community's computational and data needs cannot be specified in advance because the research itself does not yet exist. Standing infrastructure purchased for imagined future requirements is a reliable way to waste money. It is always easier to add computing capacity than to find productive uses for computing capacity purchased prematurely.

A shared pool of $300,000 covers cloud computing, full university-level online library access, dataset licensing, and occasional contractor support as specific projects require them. Most of the empirical work the paper envisions, evaluating governance interventions, tracking compliance with voluntary commitments, analyzing competitive dynamics, requires good data and statistical tools, not large-scale computing. If the institution's work eventually evolves toward computationally intensive methods, that investment can be made when the need is demonstrated rather than anticipated.

Editorial Infrastructure and Convenings

4 FTEs | $1.6M annually

This section reflects a deliberate choice to invest in editorial capacity and substantive convenings rather than a communications or policy engagement apparatus.

Getting Research to Decision-Makers

The institution needs one senior person whose knows where AI governance decisions are being made and when. When a scholar produces an evaluation of export controls and Senate Commerce is preparing a markup three weeks later, this person ensures the scholar knows and that the paper reaches the relevant staff. When the OECD verification working group meets in April, this person ensures the relevant work is circulated beforehand.

The job is knowing what decisions are imminent and routing the institution's work to the people making them at the right moment. The founding director and external affairs lead conduct briefings when scholars prefer not to or cannot.

This person has no team, no budget for relationship cultivation, and no incentive to accumulate meetings. Their value is measured by whether the institution's work arrives on the right desk at the right moment, not by how many conferences they attend.

Editorial Infrastructure

The editorial team consists of a senior structural editor and a junior editor. The senior editor is someone who has edited for a top policy journal, a publication like The Economist or Foreign Affairs, or an equivalent venue where technical rigor and accessibility coexist. This person works directly with scholars on their major outputs, not copy-editing but structural editing, such as identifying where a paper assumes knowledge the target audience does not have, where an argument's structure obscures its conclusion, where jargon substitutes for explanation. Repeated over several papers, this process teaches scholars to internalize the skill.

The freelance budget ($100,000) supports a stable of three or four specialist editors who understand specific policy communities. A paper on verification mechanisms aimed at arms control policymakers may need an editor who knows that world. A paper on compute governance aimed at trade officials needs someone who understands trade policy language. This capacity is not necessary full-time because the need is intermittent and domain-specific.

This structure reflects two beliefs. First, translation should happen through editing the research itself, not through producing simplified derivative content. A well-edited working paper that a policy advisor can engage with directly is worth more than a simplified derivative that strips out the caveats and qualifications that made the analysis valuable. Second, hiring for analytical brilliance should not be constrained by communication skills. Some of the institution's most valuable scholars will be poor communicators. The editorial infrastructure exists so that their work reaches policymakers regardless. But scholars must be willing to work with editors to make their research accessible to policy audiences. The institution cannot employ people who refuse to let their work be adapted for the people it is meant to reach.

Convenings

The convening budget ($800,000 for four to six major workshops plus two flagship events annually) is higher than academic conference costs because the participants are different. Getting deputy ministers, intelligence officials, and corporate chief scientists into the same room costs more than assembling professors. But the paper's argument about building shared vocabulary depends on these communities actually engaging with the frameworks the institution produces. Convenings are not a perk. They are a core output.

Leadership and Administration

8 FTEs | $2.5M annually

The founding director's role in this model is closer to a university president or SFI's president than to a think tank CEO. The job is to define the institutional vision, hire brilliant people, curate the community, and represent the institution externally. It is not to manage research programs or direct scholarly output. Compensation of $500,000 reflects the need to attract someone who commands respect in both academic and policy communities and who has the vision to build a new institution from scratch.

The VP of Research manages the operational environment that enables research, including the visiting program logistics, workshop schedule, research budgets, and practical coordination that a community of independent scholars requires. This person is not directing research but ensuring the scholars have what they need to do their work. The VP of Operations handles finance, HR, facilities, and compliance.

The institution requires only a lean administrative staff. A controller plus two staff handle finance and HR. Two IT and general support staff maintain the facility and systems. The $300,000 board governance budget (external review panels, audit fees, board operations) implements the external challenge mechanisms the paper's institutional design section describes. The compliance infrastructure reflects the argument that independence requires specific engineering, not just good intentions.

Facilities and Indirect Costs

$2.8M annually

Location

The budget assumes a primary campus outside a major metropolitan area. The location choice is itself a design decision with implications for institutional independence.

A location that signals "we are here to think, not to lobby or hobnob" provides a structural independence mechanism that organizational charts cannot replicate. Geography enforces separation from power in ways that conflict-of-interest policies alone cannot.

The facility budget of $800,000 assumes approximately 10,000 square feet for roughly twenty-five people at a non-metropolitan campus, including common spaces designed to produce the informal interactions that drive interdisciplinary work. When scholars need to spend extended time near multilateral processes or other institutions for specific projects, that comes from the travel budget, not from standing positions elsewhere.

The $800,000 contingency and strategic reserve exists because the SFI data shows operating deficits in both 2022 and 2023, with declining net assets despite forty years of institutional reputation. The funding concentration discussion in the main text argues that “independence is an engineering problem, not a declaration.” Part of that engineering is maintaining reserves sufficient to sustain operations when a major donor is late, a pledge falls through, or a funding cycle turns unfavorable.

Budget Summary

Line Item	Annual Cost	FTEs	Notes
I. RESEARCH			
Resident scholars (senior)	$2,500,000	10	Individually appointed; $200–300K fully loaded
Postdoctoral researchers	$500,000	4	Junior researchers; $100–150K fully loaded
Visiting scholars program	$1,500,000	—	20–30 visitors/year, 3–6 month residencies
Cloud computing & data acquisition	$300,000	—	Shared pool; cloud, datasets, occasional contractors
Subtotal: Research	**$4,800,000**		
II. EDITORIAL & CONVENINGS			
External affairs (1 senior position)	$250,000	1	Tracks policy calendars, routes research to decision-makers
Convenings & workshops	$800,000	—	4–6 workshops + 2 flagship convenings annually
Editorial team (senior + junior editor)	$270,000	2	Structural editing of research outputs
Freelance editorial & specialist editors	$100,000	—	Domain-specific editing for targeted policy audiences
Website & event logistics	$150,000	1	Working paper series, web, event support
Subtotal: Editorial & Convenings	**$1,570,000**		
III. LEADERSHIP & ADMINISTRATION			
Founding Director	$500,000	1	Sets vision, curates community, represents institution
VP Research	$350,000	1	Manages operations of the research environment
VP Operations	$300,000	1	Finance, HR, facilities, compliance
Finance, HR, legal staff	$600,000	3	Controller + 2 staff
IT & general admin	$400,000	2	Support staff
Board governance & external review	$300,000	—	Audit, external review panel, board operations
Subtotal: Leadership & Admin	**$2,450,000**		
IV. FACILITIES & INDIRECT COSTS			
Facility lease & operations	$800,000	—	Primary campus (non-metro); ~10,000 sq ft
Travel (non-convening)	$500,000	—	Research travel, conferences, field visits
Benefits loading	$700,000	—	~25% on ~$2.8M in compensation not otherwise fully loaded

Contingency & strategic reserve	$800,000	—	Funding volatility buffer
Subtotal: Facilities & Indirect	$2,800,000		
TOTAL ANNUAL OPERATING BUDGET	$11,620,000		

Staffing Summary

Category	FTEs	
Resident scholars (senior)	10	
Postdoctoral researchers	4	
Editorial, external affairs & logistics	4	
Leadership	3	
Administration & support	5	
Total permanent FTEs	**26**	

Plus 20–30 visiting scholars annually (not counted as permanent FTEs).

Comparative Context

Institution	Annual Budget	Notes
Santa Fe Institute (2022)	$11.4M	Lean resident faculty + large visiting network; recurring operating deficits (Form 990, ProPublica Nonprofit Explorer)
Santa Fe Institute (2023)	$9.2M	Revenue declined year-over-year; funding volatility despite 40-year track record
CSET, Georgetown	$8–10M est.	Single research focus (security), concentrated philanthropic funding (est. from $100M+ total, 2019-2025; CSET press release, June 2021)
Institute for Fiscal Studies (UK)	£12M	Single-country fiscal policy; narrower mandate (https://ifs.org.uk/about/finance)
Peterson Institute	$16M	International economics; closer in scope (PIIE Annual Report 2024)
Proposed institution	**~$12M**	14 resident scholars + large visiting program + editorial + convenings

At approximately $12 million, the proposed institution operates at roughly the scale of the Santa Fe Institute, which is the closest structural precedent. The difference is the policy-facing capacity (editorial infrastructure, convenings, external affairs) that SFI deliberately avoids. The proposed institution’s budget is smaller than institutions like the Peterson Institute or IFS because it avoids high spending on infrastructure, communications staff, or administrative overhead.

Phasing

The steady-state budget of approximately $12 million assumes a mature institution with its full complement of resident scholars and visiting program at scale. The institution would not start at this level.

Year 1 ($3–5M). Founding director, VP Research, and a small administrative team. Three to four resident scholars recruited. Initial visiting scholar cohort (perhaps 8–10 visitors). Senior editor hired. First workshop convened.

Years 2–3 ($6–9M). Six to eight resident scholars in place. Postdoc program launched. Visiting program approaching full scale. Initial empirical projects underway. Full convening program. First major publications.

Years 4–5 ($10–12M). Full complement of scholars. Steady-state operations. By this point the community should have produced the foundational frameworks the paper describes: a shared strategic grammar, initial governance evaluations, and a risk taxonomy that disciplines the institution's ongoing work.

Funding Model

At approximately $12 million annually, the funding challenge is serious but not unprecedented. This is roughly the scale at which institutions like SFI, CSET, and the Institute for Fiscal Studies appear to operate, based on recent public figures. It is a bit less than what a $250 million endowment would yield at a 5 percent draw rate. The funding concentration discussion in the main text argues that “a single funding source creates a single point of institutional failure.” The most resilient model would combine several streams.

Philanthropic general support from multiple foundations and individuals would provide the base.

Government research grants for specific research (verification, governance evaluation) could support a portion of the scholarly community's work without compromising the institution's independence on strategic questions.

Applied network or membership revenue, modeled on SFI's Applied Complexity Network, would bring corporate and government partners who pay for access to research and convenings without directing the scholarly agenda.

Partial endowment income would reduce dependence on annual fundraising. Even a $50 million endowment at 5 percent would cover $2.5 million annually, roughly 20 percent of operating costs.

The SFI experience is instructive on the difficulty even at this scale. SFI's founders envisioned a $300 million endowment. Forty years later, the institute operates on $9 to $11 million annually with recurring deficits and funding volatility.

Key Assumptions

Salary benchmarks. Senior scholars at $200,000 to $300,000 fully loaded; postdoctoral researchers at $100,000 to $150,000. These rates are competitive with top policy research institutions but below industry compensation for comparable technical talent. The visiting scholar program and the independence of the appointment are the primary non-monetary attractions.

No dedicated computational infrastructure. The scholarly community's computational needs are met through cloud computing purchased as needed. Most of the empirical work the paper envisions requires good data and statistical tools, not large-scale computation. If the institution's work evolves toward computationally intensive methods, that investment can be made when the need is demonstrated.

No endowment income at steady state. The budget is constructed assuming the institution operates entirely on annual revenue. Any endowment income would reduce the annual fundraising burden proportionally.

Benefits loading. Applied at approximately 25 percent on roughly $2.8 million in compensation lines not otherwise described as fully loaded. This covers health insurance, retirement contributions, and payroll taxes.

Flat organizational structure. No program directors or departmental hierarchy. The founding director and VP Research curate the community and manage the research environment. Scholars are individually appointed and pursue their own research agendas within the institution's mandate. Intellectual coordination emerges from proximity, shared questions, and the convening and visiting programs, not from management.

Facility costs. Assume a non-metropolitan primary campus. A DC location would increase facility costs substantially and, more importantly, would compromise the institutional independence the main text argues is essential.