REGULATING AI DEVELOPMENT:
CHALLENGES, CONSTRAINTS, AND A REALIST AGENDA

Steven E. Koonin

Edward Teller Senior Fellow, Hoover Institution

Stanford University

June 2026

Abstract. Calls for comprehensive regulation of artificial intelligence have intensified as the technology's capabilities have grown. Such calls conflate two distinct regulatory objects: the domestic application of AI systems, which is already underway and broadly feasible, and the development of increasingly capable AI systems at the global frontier, the subject of this paper, where the structural constraints attenuate traditional regulatory approaches. This paper examines three structural features of AI development that define the constraint envelope within which any workable governance policy must operate: a verification problem that makes enforcement fundamentally different from previous dual-use technology regimes; a self-interest problem that will shape international compliance in predictable ways; and a beneficial applications problem that creates an unfavorable political economy for sustained restraint. Understanding these constraints does not counsel despair, but it does counsel realism — a shift to shaping AI development rather than preventing it, to strengthening resilience, and less ambitiously but more honestly, ensuring that whoever sits at the frontier does so with some transparency and accountability. The result is a menu of policy recommendations that are achievable under real-world constraints.

I. The Regulatory Impulse

The arrival of large language models capable of sophisticated reasoning, code generation, and autonomous action has produced a predictable institutional response. Governments, international bodies, and a significant portion of the AI research community have called for regulatory frameworks that would subject the development and deployment of advanced AI systems to oversight, mandatory testing, and in some proposals, prerelease approval requirements. The European Union enacted the AI Act in 2024. At the state level, California's Transparency in Frontier AI Act and New York's RAISE Act, both of which impose transparency and safety obligations on frontier model developers, have taken effect in 2026; some 370 AI-related measures were introduced across state legislatures this year alone.

The federal picture is more contested. The Biden administration issued an executive order in 2023 directing federal agencies to develop AI governance standards; it was rescinded by the Trump administration in January 2025 and replaced by a "light touch" posture prioritizing innovation and U.S. competitiveness. The administration has since issued a December 2025 executive order seeking to preempt state-level AI regulation, a March 2026 legislative framework urging Congress to establish a uniform national standard, and most recently a June 2026 executive order creating a voluntary framework for frontier labs to share models with federal evaluators — explicitly prohibiting any interpretation of the order as authorizing mandatory licensing or preclearance. That last order was itself a scaled-back version of a more ambitious draft that the president withdrew, saying he did not want to do anything that would get in the way of American AI leadership.

The administration's trajectory — rescinding safety requirements, preempting state regulation, resisting even voluntary oversight mechanisms — reflects a coherent if contested philosophy. Yet it has not fully escaped its own interventionist impulses: senior economic advisors have argued that frontier models should be released only after safety is demonstrated, and the June order's voluntary evaluation framework embeds at least a residual precautionary logic. The Pope has also weighed in, illustrating how far the regulatory impulse has spread beyond technical and policy circles.

The regulatory impulse is not irrational. Capable AI systems, even absent the still elusive “general intelligence,” pose genuine challenges. The concentration of frontier model development in a small number of private companies, the opacity of large neural networks even to their builders, the speed

at which new capabilities emerge, and the breadth of potential applications — including national security, healthcare, and critical infrastructure — all create legitimate grounds for public concern. The question is not whether the concern is warranted. It is whether the proposed responses are actually well matched to the structure of the problem.

Before examining these structural challenges, note an essential distinction that much of the current regulatory debate obscures. There are in fact two quite different objects of potential regulation. The first is the domestic use and deployment of AI systems — in hiring, lending, medical diagnosis, law enforcement, content moderation, autonomous vehicles, and the countless other applications now entering commercial and governmental practice. Regulation here is not only feasible but already underway: the EU AI Act, various US state-level initiatives, and sector-specific agency guidance all represent genuine, potentially workable governance of how AI is applied within jurisdictions. This paper does not contest that project.

The second and distinct object of regulation is the development of increasingly capable AI systems — the frontier research and large-scale training that produces the next generation of foundation models. It is here that the structural challenges analyzed below apply with full force. Frontier AI development is a global activity pursued by dozens of nations and hundreds of institutions; the knowledge is published; the compute is commercially available; and no jurisdiction can unilaterally constrain what capabilities exist. Conflating these two objects — treating "regulate AI" as a single coherent project — produces policy confusion: it imports the apparent feasibility of use regulation to lend credibility to development regulation, where the constraint envelope is entirely different.

This paper argues that three structural features of AI development — call them the verification problem, the self-interest problem, and the beneficial applications problem — define hard constraints on what any governance regime can achieve. These features do not make AI governance impossible, but they do make certain approaches not merely futile but counterproductive, and others more promising. Understanding the constraint envelope is the necessary first step toward formulating policy that could actually work.

[Author's note: This paper stands alongside two other Hoover working papers addressing the governance of advanced AI, each prepared independently. Niall Ferguson, drawing on the history of nuclear brinkmanship and arms control, argues that the unconstrained AI races between leading

companies and between the superpowers pose dangers grave enough to warrant a rapid transition to a US-China détente built on AI arms control. Philip Zelikow, Eric Schmidt, and colleagues — while acknowledging that "past arms control precedents are not a very good template for this problem" — propose an ambitious coalition defense enterprise, organized around threat assessment, countermeasures, and public-private partnership on a historic scale. That three independent analyses converge on the gravity and urgency of the challenge, while diverging on what can realistically be done about it, is itself instructive: the diagnosis is clear, but the prescription is where the hard questions live. The present paper locates those hard questions in three structural constraints — verification, self-interest, and beneficial applications — and argues that these have been underweighted in much of the current debate.]

II. The Verification Challenge: No IAEA for AI

The most instructive comparison for AI regulation is not the FDA, which approves discrete products with identifiable chemical compositions and testable mechanisms of action, but international nuclear nonproliferation — and the comparison is instructive precisely because the analogy breaks down where it matters most.

Nuclear weapons development has proven relatively, if imperfectly, containable over the past eight decades. The key to that partial success is physical. Creating nuclear weapons requires large industrial facilities — enrichment plants, reactors, reprocessing and testing infrastructure — whose signatures are detectable by satellite, seismic monitoring, and atmospheric sampling. It requires specialized materials: highly enriched uranium or plutonium, whose precursors must be mined, processed, and handled through supply chains that can be monitored and interdicted. The centrifuge cascades that produce weapons-grade uranium cannot be hidden in a laptop, nor can one download weapons-grade fissile material.

These physical chokepoints are what make the International Atomic Energy Agency possible. The IAEA exists because there is something for inspectors to inspect: facilities to visit, materials to inventory, signatures to detect. The verification regime is imperfect — Iraq, North Korea, and Libya all evaded it for periods — but it is not nothing. Physical reality gives inspectors leverage that is simply unavailable in the software domain.

AI development has none of these properties. The compute infrastructure that enables frontier model training is inherently dual-use: a data center optimized for advertising, weather modeling, or financial analytics differs from one training a large language model only in degree, not in kind. The chips involved are commercially available. The models themselves are software: weight files that can be copied in seconds, transmitted across the globe without physical movement, and executed without any observable external signature. A national AI program has no detectable seismic or atmospheric signature nor any unique observable footprint; a 70-billion-parameter model trained on a rented cloud cluster is indistinguishable from a weather-forecasting workload.

More fundamentally, the knowledge required to build capable AI systems is published openly. Transformer architecture, reinforcement learning from human feedback, the scaling laws that predict model capability are all in the open literature, widely understood, and being actively

pursued by researchers in almost every country and a myriad of institutions. Open-source models of considerable capability are freely available to anyone with an internet connection. A determined actor does not need a classified industrial complex; a credit card and a cloud account will suffice. A bedroom (or a bunker) can be a laboratory in a way it has never been for any previous category of dual-use technology with significant destructive potential.

The resulting verification challenge is not merely technically difficult but sits at the far end of a detectability spectrum that illuminates why AI governance is uniquely hard. Nuclear weapons development occupies one end of that spectrum: high detectability, strong governance traction, a functioning international inspection regime. Biological weapons occupy the middle. Like AI, they do not require massive industrial infrastructure — a sophisticated laboratory, the right expertise, and access to specific biological materials or synthesized DNA sequences can suffice. But bioweapons retain partial detectability that AI entirely lacks. It has long been understood that unique pathogen sequences can in principle be flagged at commercial DNA synthesis facilities, and nascent international screening efforts exploit this chokepoint.[1] Large-scale efficacy trials require human or animal subjects on a detectable scale. A domestic vaccination campaign to protect one's own population before agent release — historically cited as an intelligence indicator of offensive bioweapons intent — would leave observable traces. These signatures are imperfect and exploitable, but they are real: they give inspectors and intelligence services something to look for. AI development provides nothing equivalent. There is no sequence to flag, no trial to observe, no vaccination campaign to detect. A state or non-state actor training a dangerous AI system is, from the outside, indistinguishable from one training a system for commerce, science, or entertainment. AI sits at the far end of the detectability spectrum, where governance traction is weakest and the gap between declared norms and actual behavior is hardest to close.

There is a further complication that the spectrum framing makes visible. AI is not merely presenting its own governance challenge in parallel with biological weapons — it is actively eroding whatever verification traction bioweapons governance has achieved. AI-assisted protein structure prediction, automated laboratory systems, and large-scale biology foundation models are simultaneously reducing the expertise barrier for dangerous pathogen development, compressing

[1] While leading AI executives have signed an open letter calling for stricter regulation of DNA sequences, the letter remains silent on whether biosecurity regulation should be applied to AI firms.

the timeline from concept to dangerous agent, and enabling the design of novel sequences that existing synthesis screening databases might not recognize. The partial chokepoint that made biological weapons governance partially tractable is being narrowed by the same technology whose own governance is the subject of this paper. This interaction effect deserves attention in its own right; for present purposes it reinforces the urgency of getting AI governance right before the window closes further.

The verification challenge does not make domestic oversight pointless. It does define what domestic oversight can and cannot accomplish: it can create transparency and accountability for actors within a jurisdiction, but it cannot constrain global AI development. Policy that conflates these two objectives will achieve neither.

III. The Self-Interest Challenge: Countries Do What Serves Them

The history of international regulatory agreements is largely a history of nations signing commitments and defecting when compliance costs become visible. This is not cynicism but political science: states are the primary actors in the international system, their behavior is shaped by domestic interests and competitive pressures, and agreements that impose significant asymmetric costs tend to erode.

The analogy with greenhouse gas mitigation is sobering. Decades of international negotiation have produced a succession of agreements — Kyoto, Copenhagen, Paris — whose ambitions to restrain human emissions of greenhouse gases have consistently exceeded their results by enormous margins. Nations that signed binding commitments subsequently withdrew, quietly revised their targets, or reported compliance while emissions continued to rise. The gap between declared norms and actual behavior has been persistent and substantial. This is not because the negotiators were insincere. It is because the costs of decarbonization are concentrated, immediate, and fall on identifiable domestic constituencies, while the supposed climate benefits are vague, diffuse, long-term, and shared with free riders.

Regulation of AI development would face an even more challenging calculus of self-interest. The advantages of AI capability are not merely economic but also military and geopolitical. A nation at the frontier of AI development has advantages in intelligence collection, autonomous systems, cyber operations, strategic planning, and logistics that translate directly into national power. The

incentive to defect from any agreement constraining AI development is therefore not merely economic — as with emissions — but strategic in the security sense. Nations that accept binding constraints while adversaries proceed unconstrained are not only forgoing economic opportunity; they are accepting strategic disadvantage. And even if a treaty were signed, the incentives to defect would only grow as AI becomes more militarily decisive.

This dynamic has played out in every previous domain where dual-use technology intersected with national security. Nuclear nonproliferation succeeded in slowing but not preventing proliferation and succeeded only where the United States and its allies were willing to apply sustained pressure against specific states. Chemical and biological weapons conventions have been violated repeatedly by signatories. Export control regimes have been circumvented through third-country routing and indigenous development. In each case, the pattern is the same: nations comply when compliance is cheap and defect when it is not.

AI regulation would impose costs that are neither cheap nor symmetric. For democratic societies with open research environments and rule-of-law institutions, compliance would be genuine and observable. For authoritarian states with opaque research programs and different values, compliance would be voluntary and unverifiable (*i.e.*, "optional"). An international AI regulatory regime would therefore not constrain the actors whose behavior is of greatest concern, but it would constrain their competitors. Governance approaches that ignore this asymmetry will not merely fail but will produce outcomes contrary to their stated aims.

IV. The Beneficial Applications Challenge: The Political Economy of Restraint

The nuclear case offered a relatively favorable political economy for restraint. The primary application of nuclear weapons is mass destruction. Whatever the deterrence logic that sustained the Cold War balance, the humanitarian case for limiting the spread of weapons of mass destruction was compelling, widely shared, and politically sustainable across administrations and lawmakers. The costs of nuclear restraint — forgoing a weapons program — were real but abstract for most states that might have pursued one.

AI is different in kind. The beneficial applications of capable AI systems are not hypothetical or long-term. They are arriving now, across virtually every domain of human activity. AI-assisted

drug discovery is accelerating the identification of candidates for diseases that have resisted treatment for decades. AI systems are improving diagnostic accuracy in radiology, pathology, and genomics. The same technology is accelerating materials science, weather forecasting, software development, and mathematical research. For education, it offers the prospect of individualized instruction at scale. For economic productivity, it may be the most significant general-purpose technology since electrification.

This breadth of benefits creates a political economy of restraint that is fundamentally different from the nuclear case. Asking nations to forgo AI development is not asking them to forgo weapons of mass destruction. It is asking them to forgo better cancer diagnostics, faster drug development, higher agricultural yields, improved weather forecasts, and economic productivity gains that compound over time. The constituencies opposing restraint are not defense contractors and weapons scientists. They are oncologists, farmers, software developers, educators, and — eventually — every citizen who might benefit from an AI-assisted medical system.

This means that even if an international regulatory agreement could be negotiated, the domestic politics of sustained compliance would be treacherous. The constituencies that benefit from AI development — in medicine, agriculture, education, and industry — are too broad and too vocal for any administration to sustain development constraints indefinitely, regardless of what international agreements nominally require. The political economy of restraint is not merely unfavorable. It is probably unsustainable. Governance approaches must reckon with this reality rather than assume it away.

The political economy of restraint is further complicated by a domestic resistance movement that might seem, at first glance, to cut the other way. Concerns about labor displacement, algorithmic bias, synthetic media, and AI-enabled surveillance have generated genuine political energy in the United States, and one might argue that this resistance creates space for sustained constraints on AI development. But the inference does not follow. The domestic resistance movement is directed overwhelmingly at deployment and use — AI's energy needs and how it is applied in hiring, content moderation, policing, and media — not at whether the United States should remain at the frontier of development. The concerns animating domestic resistance — about bias, accountability, and the values embedded in powerful systems — are precisely the concerns that make U.S. frontier leadership important. A world in which the development frontier is ceded to actors immune to

domestic civil society pressure, litigation, and democratic oversight is a world in which those concerns cannot be addressed at all. The domestic critics of AI deployment are, in this sense, making the case for the policy recommendation in Section VI: that it matters enormously who sits at the frontier, and under what conditions of transparency and accountability they operate.

V. What the Regulatory Consensus Gets Wrong

The most straightforward regulatory scheme, requiring formalized "safety" tests of an AI model as a condition for its release, is unlikely to be effective given the breadth of risks, the flexibility of model employment, and the rapidity of the technology's development. A more sophisticated regulatory scheme, exemplified by a recent proposal for bank-examiner-style supervision of frontier AI labs, concedes the weaknesses of existing schemes and offers a genuinely thoughtful alternative. The banking analogy captures something real: frontier AI labs are private actors making fast-moving, technically complex decisions with significant public consequences, and the regulatory models designed for discrete products with testable properties do not fit.

But even this sophisticated proposal embeds an assumption that it does not examine: that the actors whose behavior matters most can be brought within the supervisory framework. Bank supervision works because the banking system is jurisdictionally captive. Banks need charters, deposit insurance, and access to the central bank. These dependencies give supervisors genuine leverage. But an AI lab that found domestic supervision intolerable could relocate, distribute, or open-source. And no domestic supervisor has any jurisdiction over the thousands of actors worldwide who are developing capable systems using published methods and commercially available compute.

The error at the center of the emerging regulatory consensus is the assumption that AI governance is amenable to approaches that have worked for previous dual-use technologies: controlling materials, monitoring facilities, certifying compliant actors, establishing norms that gradually acquire the force of customary international law. These approaches work when there are physical chokepoints, when the technology is difficult to replicate independently, and when the costs of the technology's misuse fall primarily on those who possess it. AI shares none of these properties.

Recognizing this is not defeatism, but rather a necessary precondition for asking the right question to design governance that actually works: given that capable AI systems will continue to be

developed by multiple actors under varying conditions of transparency and accountability, what governance approaches can actually shape outcomes at the margin? That reframing — from prevention to shaping, from comprehensive control to targeted resilience — points toward a different and more tractable policy agenda.

VI. A Realist Policy Agenda

The three structural challenges identified above define a constraint envelope. Policy that ignores the envelope will fail. Policy designed within it can succeed, not by preventing the development of capable AI, but by shaping who develops it, under what conditions of transparency, and with what resilience against misuse. The following recommendations follow directly from the analysis.

Accept the constraints explicitly. Any serious governance framework must begin by acknowledging what it can and cannot achieve. Approaches premised on comprehensive control — preventing capable AI from being developed by actors outside the framework — will waste resources, impose costs on compliant actors, and produce the illusion of safety without substance. Honest framing isn't an aesthetic preference. It is the precondition for frameworks that last.

Require domestic transparency without conflating it with global control. Rather than prerelease approval, which implies the ability to stop something, require frontier labs operating in the United States to maintain detailed internal documentation of capabilities, red-team findings, safety evaluations, and incident reports, accessible to a designated government body under appropriate confidentiality protections.[2] The bank examiner concept has genuine merit here: continuous engagement, confidential disclosure, and the ability to require mitigations before problems become public crises. The critical adjustment is to be clear about what this accomplishes. It creates visibility and accountability within domestic jurisdiction. It does not constrain global AI development, and policy should not pretend otherwise.

Invest in defensive capabilities rather than preemptive restraint. If capable AI cannot be prevented from existing, the policy priority shifts from prevention to detection and resilience. This means hardening critical infrastructure against AI-enabled attack; building the capacity to detect, analyze, and counter adversary AI systems; and securing the model weights and supply chains on

[2] Details will matter here. Documentation requirements that sound straightforward in legislation often become ineffective bureaucratic labyrinths in practice, in the worst case ossifying into compliance theater.

which that capacity depends. This is the analog of civil defense and deterrence rather than nonproliferation — a lesser aspiration, more achievable in practice.

Maintain US frontier leadership as a policy objective in itself. This is the most counterintuitive recommendation given the current political mood, but it follows directly from the analysis of the self-interest problem. If capable AI is coming regardless of what any single jurisdiction does, it matters enormously who is at the frontier. US and allied labs operating under democratic norms, with genuine safety research cultures and some transparency obligations, are preferable to a world in which the frontier is ceded to actors with different values and no accountability. Regulation that slows Western development without equivalently slowing others is not a safety policy. It is strategic abdication dressed in the language of precaution.

Pursue selective international engagement on narrow, verifiable issues. Not a comprehensive treaty — the self-interest and verification challenges make that futile — but targeted agreements on specific applications where verification is more tractable and the case for restraint is compelling enough to sustain political will across administrations. Autonomous lethal weapons systems with no meaningful human control, AI-assisted design of biological or chemical weapons, and AI-enabled attacks on nuclear command and control infrastructure are candidates: the potential harms are severe, the beneficial applications are limited, and the case for restraint is one that democracies and some authoritarian states might find it in their interest to support. The arms control analogy here is not the Nuclear Non-Proliferation Treaty but the more targeted prohibitions of the Chemical Weapons Convention — imperfect and partially violated, but not nothing.

The gene synthesis screening model is instructive here as a template. International coordination to require screening of DNA synthesis orders against databases of dangerous pathogen sequences is precisely the kind of narrow, technically grounded, partially verifiable agreement that can attract genuine compliance: the beneficial applications of dangerous sequence synthesis are limited, the potential harms are severe, and the chokepoint is real enough to give the agreement some teeth. Analogous agreements for AI might focus on the highest-risk, most narrowly defined applications such as AI systems designed to autonomously develop or optimize biological or chemical agents, or to enable attacks on nuclear command and control infrastructure. In those cases, the harmful application is sufficiently distinct from beneficial uses to make targeted prohibition at least partially meaningful.

Prioritize interpretability research as a public good. The core problem that regulators face — that frontier AI systems do things their builders do not fully understand — is likely not a permanent feature of the technology. Interpretability research aims to make neural network behavior legible: to identify what a model has learned, how it represents information, and what it will do in conditions outside its training distribution. Progress here would change the underlying verification landscape, making future oversight more tractable than current oversight. It would also change the safety calculus for the labs themselves. This research is undersupplied by the market because the benefits are diffuse and the findings are public goods. It deserves priority public funding since the governance problem will not get easier until the underlying science does.

Taken together, these recommendations constitute a coherent realist policy agenda: domestic transparency without the pretense of global control; resilience rather than prevention; frontier leadership as a strategic objective; selective international engagement on the narrowest tractable issues; and sustained investment in research that could expand the constraint envelope over time. None of this is as satisfying as the comprehensive frameworks that current regulatory proposals envision. But satisfaction is not the test. Effectiveness is.

VII. Conclusion

The argument advanced here is not that AI poses no risks, or that governance is impossible, or that the current trajectory of AI development is without concern. It is a more specific claim: that the principal proposed response to those risks — comprehensive regulation designed to constrain the development of capable AI systems globally — faces three structural challenges that cannot be wished away by sophisticated regulatory design.

The verification challenge is real: there is no physical signature to detect, no material to interdict, no facility to inspect at the international level, and the bedroom (or a bunker) is a plausible laboratory. The self-interest challenge is real and historically persistent: nations comply with constraints that are cheap and defect from those that are costly, and AI capability is too directly connected to economic and strategic power for restraint to be durable. And the beneficial applications challenge will not dissolve under political pressure: the constituencies for AI development span every domain of human flourishing, and the political economy of sustained restraint is probably unsustainable.

These are not arguments for passivity. They are arguments for realism — for governance approaches designed within the constraint envelope rather than premised on transcending it. The right question is not how to prevent AI from improving, but how to ensure that those who develop it operate with transparency and accountability, that societies are resilient against its misuse, and that the frontier is held by actors whose values align with the interests of open, democratic societies. The answer to that harder question will be the foundation for governance that can endure as the technology evolves.

About the Author Steven E. Koonin is the Edward Teller Senior Fellow at the Hoover Institution, Stanford University. He was previously University Professor and Director of the Center for Urban Science and Progress at New York University, Chief Scientist for BP plc, Provost of the California Institute of Technology, and Under Secretary for Science in the U.S. Department of Energy.
