

The Efficiency Shock in AI Governance: Open-Weight Models, Regulatory Fragmentation, and the Emerging Limits of Chokepoint Control

Alexander Lee

Hong Kong Journal of Social Sciences Press

Abstract

This article argues that the prevailing “Chokepoint Paradigm” in AI governance—predicated on controlling compute, capital, and centralized infrastructure—has been destabilized by an “Efficiency Shock”: the decoupling of state-of-the-art capabilities from prohibitively high resource thresholds. Situating this structural break within historical institutionalism, we term the governance inflection window the “DeepSeek Moment,” exemplified by open-weight models achieving frontier-level reasoning at a fraction of prior training costs. We identify three consequences: the emergence of “Open-Weight Sovereignty,” where capability diffusion outpaces export controls; a “Liability Vacuum” in frameworks such as the EU AI Act, whose provider-centric architecture cannot capture distributed open-weight ecosystems; and a geopolitical shift from containment to “Fragmented Adaptation,” characterized by three divergent governance blocs. Three testable propositions are advanced regarding export control decay, provider-concept failure, and sovereignty-driven model adoption. We conclude that governance must pivot from supply-side restriction to impact-based monitoring and adaptive liability.

Keywords Efficiency Shock, open-weight models, Chokepoint Paradigm, regulatory fragmentation, AI sovereignty, DeepSeek Moment, algorithmic governance, EU AI Act

1 Introduction: The Chokepoint Paradigm Under Stress

1.1 The Prevailing Governance Logic

Since the emergence of generative artificial intelligence as a global strategic priority, the prevailing logic of governance has been predicated on scarcity. The dominant “Chokepoint Paradigm”—evident in U.S. export controls on advanced semiconductors (Bureau of Industry and Security, 2022), the compliance thresholds of the European Union’s AI Act (European Parliament, 2024), and the centralized safety commitments of the Bletchley Declaration (UK Government, 2023)—assumes that “frontier” AI models are, and will remain, industrial mega-projects. This paradigm holds that because state-of-the-art (SOTA) capabilities require tens of thousands of GPUs, billions of dollars in capital, and massive energy infrastructure, the entities capable of producing them will remain few, visible, and regulatable. Under this logic, governance is a matter of regulating the “provider”—the central node in a linear supply chain (Bradford, 2020; Engler, 2022).

This assumption is not unique to AI. It mirrors established governance logics in nuclear non-proliferation (where fissile material production is the chokepoint), pharmaceutical regulation (where centralized manufacturing enables quality control), and financial regulation (where licensed intermediaries

serve as compliance nodes). Across these domains, the presumption is that strategic capabilities require strategic resources, and that controlling the resource controls the capability.

1.2 Literature Gap: The Unexamined Assumption of Scale

Work on algorithmic governance has documented how the rise of large platforms and automated decision-making systems reconfigures state authority and accountability (Yeung, 2018; Katzenbach & Ulbricht, 2019; Danaher et al., 2017). In international relations, scholars have framed AI competition through the lens of semiconductor chokepoints and the extraterritorial power of US-based cloud providers (Farrell & Newman, 2019; Allen, 2022; Ding & Dafoe, 2023). Platform governance scholarship has analyzed the “Brussels Effect,” whereby EU regulation sets de facto global standards through market power (Bradford, 2020).

However, this body of work shares an implicit and largely unexamined assumption: that frontier AI capabilities are *structurally* tethered to massive scale. This assumption of a monotonic relationship between compute investment and capability level underpins the entire edifice of chokepoint-based governance. If capabilities can be produced at dramatically lower cost—if the “moat” of capital is shallow rather than deep—then the regulatory architecture built on scarcity assumptions confronts a fundamental design flaw.

1.3 The Efficiency Shock and Research Questions

We argue that this hypothesis was falsified by an “Efficiency Shock”: the sudden decoupling of SOTA-level reasoning capabilities from prohibitively high resource barriers. The release of open-weight models that rival closed-source frontier systems—exemplified most prominently by the DeepSeek R1 series (DeepSeek-AI, 2025)—demonstrates that algorithmic innovation can substitute for raw compute scale, compressing the cost of intelligence by orders of magnitude.

This article terms this structural rupture the “DeepSeek Moment.” We use this term to denote the governance inflection window (late 2025–early 2026) in which efficiency breakthroughs and open-weight diffusion became *politically salient*, rather than the publication date of a single technical report. The DeepSeek Moment is not merely a commercial or technical milestone, but a critical juncture (Capoccia & Kelemen, 2007) in institutional history where the “moat” of capital—the primary barrier to entry and the primary lever of control—evaporated for a significant class of strategic capabilities. It signals the erosion of the “containment era” in AI policy and the beginning of an era of “Open-Weight Sovereignty,” characterized by the uncontrolled diffusion of powerful reasoning agents across borders and regulatory jurisdictions.

This article addresses three research questions: (1) How does the emergence of highly capable open-weight models alter the assumptions underpinning provider-centric AI regulation? (2) What forms of legal and institutional “liability vacuum” arise when powerful models become downloadable artifacts rather than mediated services? (3) How might states with different geopolitical positions adapt to this shift, and what patterns of “fragmented adaptation” can be anticipated?

1.4 Contributions and Structure

This article makes three contributions. *Theoretically*, we propose the concept of “Efficiency Shock” as a general governance failure mode—a class of events where capability thresholds drop below regulatory activation thresholds, rendering existing institutional architectures obsolete. We embed this within historical institutionalism’s theory of critical junctures. *Institutionally*, we reveal how the EU AI Act’s liability architecture, built around the legal fiction of a unitary “Provider,” suffers from what we term *Regulatory Dysgraphia*—a structural inability to “read” and capture the technical reality of open-weight model distribution. *Geopolitically*, we propose a three-bloc typology of “Fragmented Adaptation” (Containment

Bloc, Sovereign Bloc, Open Commons) that maps how states will differentially respond to uncontrolled capability diffusion.

The article proceeds in five parts. Section 2 defines the “DeepSeek Moment” as a structural break and analyzes the mechanism of the Efficiency Shock. Section 3 generalizes the Efficiency Shock as a governance failure mode, drawing on historical analogies from cryptography, nuclear technology, and drone proliferation. Section 4 examines open-weight models as a distinct regulatory object, demonstrating how they create a “Liability Vacuum” in frameworks like the EU AI Act, and proposes a design space for post-provider liability. Section 5 analyzes the geopolitical consequences, proposing the “Fragmented Adaptation” typology with observable implications. Section 6 concludes with a research agenda, testable propositions, and policy design principles.

2 The DeepSeek Moment: A Structural Break

2.1 From the Scaling Hypothesis to the Efficiency Shock

In historical institutionalism, a “critical juncture” occurs when a stable institutional pathway is disrupted by an exogenous shock that renders existing rules obsolete (Capoccia & Kelemen, 2007; Mahoney, 2000). The “DeepSeek Moment” represents such a juncture for AI governance. Before late 2025, the trajectory of AI development appeared to confirm the “Scaling Hypothesis” (Kaplan et al., 2020): that better performance strictly required more compute, more data, and more money. This trajectory naturally favored a concentrated market structure, which in turn supported a “chokepoint” regulatory strategy: if you control the chips (hardware) and the largest data centers (infrastructure), you control the AI.

The release of DeepSeek R1—a model that achieved reasoning benchmarks at strategically relevant capability levels comparable to leading closed-source models (e.g., OpenAI’s o1 series) at a fraction of the training cost and inference latency—shattered this linearity (DeepSeek-AI, 2025). Specifically, DeepSeek R1 employed a Mixture-of-Experts (MoE) architecture, multi-stage reinforcement learning from human feedback (RLHF), and knowledge distillation techniques that allowed it to match or approach the performance of o1-mini on benchmarks including AIME 2024, MATH-500, and LiveCodeBench, while reportedly incurring training costs orders of magnitude below those of comparable closed-source systems.

We emphasize that this comparison is not intended as a definitive performance ranking, but as an illustration of the declining resource threshold required to reach strategically relevant capability levels. We focus on *strategically relevant thresholds*, not absolute frontier rankings. The profound governance implication lies in the *trend line*: the resource threshold required to achieve “dangerous” or “strategic” utility is falling faster than regulatory capacity can adapt.

2.2 The Mechanism of the Efficiency Shock

The “Efficiency Shock” operates through a specific causal mechanism: lowering the barrier to entry for *sovereign hosting*. Previously, deploying a frontier model required reliance on API services provided by US-based tech giants (Microsoft, Google, OpenAI, Anthropic). This dependency created a de facto governance layer: these companies could refuse service, filter outputs, and comply with US sanctions. The closed-source API model was, in effect, a regulatory chokepoint by design.

Open-weight models like R1, however, can be downloaded from public repositories (e.g., HuggingFace), fine-tuned on domain-specific datasets, and hosted on local infrastructure—including consumer-grade GPUs or modest national supercomputers. Once the weights are public, the “switch” that central providers could flip to disable the model no longer exists. The Efficiency Shock thus transforms not just access, but *control*. It converts high-level AI from a “service” (rented, centralized, governable) into an “artifact” (owned, decentralized, ungovernable).

Moreover, DeepSeek is not an isolated data point. It is part of a broader trend of open-weight model proliferation. Meta’s Llama series, Mistral AI’s models (France), and the Technology Innovation Institute’s Falcon series (UAE) all demonstrate that high-capability open-weight models are emerging from multiple geopolitical origins. The DeepSeek Moment is merely the most dramatic illustration of a structural shift that was already underway.

2.3 The Irreversibility Thesis

Critically, this shock is functionally irreversible under conditions of open diffusion. Knowledge of algorithmic optimizations (e.g., Mixture-of-Experts architectures, distillation pipelines, synthetic data generation) diffuses instantly through technical papers, open-source repositories, and informal developer networks. Even if future frontier models re-establish a “scale advantage” at the absolute peak of performance, the DeepSeek Moment has established that “tier-one-minus” capabilities—relevant to high-risk domains such as autonomous cyber operations, biological sequence analysis, and mass-scale persuasion—are now permanently accessible to any actor with a modicum of hardware.

This creates what we term the “Moving Frontier Problem” for governance: even if the absolute SOTA continues to require massive investment (Bommasani et al., 2021), the “*dangerous enough*” tier is now structurally democratic. Governance based on controlling the apex is irrelevant if the threshold of strategic utility has already been crossed by widely available models. DeepSeek R1 functions here as a *focal case* rather than a unique driver; the analytical claims hold insofar as similar efficiency shocks are observable across multiple open-weight lineages (Llama, Mistral, Falcon, Qwen).

3 Efficiency Shocks as a General Governance Failure Mode

3.1 Defining the Mechanism

We propose that the “Efficiency Shock” is not unique to AI but represents a *general class of governance failure* that occurs whenever a technological breakthrough reduces the resource requirements for a capability below the threshold at which existing regulatory institutions are designed to activate.

Formally: **When capability cost < regulatory activation threshold → institutional failure.**

The governance architectures of the 20th and early 21st centuries were implicitly designed around a scarcity assumption: that dangerous capabilities require rare, expensive, and therefore controllable inputs. Nuclear weapons require enriched uranium; biological weapons require BSL-4 laboratories; precision-guided munitions require military-grade electronics. Regulation is organized around controlling these inputs—the “chokepoints.” An Efficiency Shock occurs when innovation renders the controlled input unnecessary or substitutable. The capability persists, but the chokepoint vanishes.

3.2 Historical Analogies

Cryptography and the PGP Wars (1990s). The U.S. government classified strong encryption as a “munition” under the International Traffic in Arms Regulations (ITAR), assuming that only state-level actors could produce cryptographic systems of military grade. Phil Zimmermann’s release of Pretty Good Privacy (PGP) in 1991 demonstrated that strong encryption could be implemented on personal computers by a single developer. The chokepoint (specialized hardware and classified knowledge) was bypassed by mathematical efficiency. The result: decades of failed regulatory attempts (the “Crypto Wars”), followed by an eventual capitulation to universal encryption (Levy, 2001).

Nuclear Technology and Centrifuge Miniaturization. The Manhattan Project assumed that nuclear capability required industrial-scale gaseous diffusion plants—facilities so large they were visible from

satellites. The development of high-speed gas centrifuge cascades (and the A.Q. Khan proliferation network) demonstrated that enrichment could be achieved with far smaller, more concealable facilities. The chokepoint (industrial scale) was eroded by engineering efficiency (Albright & Hinderstein, 2005).

Commercial Drones and the Democratization of Lethality. Military-grade unmanned aerial vehicles were once the exclusive domain of advanced militaries. The commoditization of GPS modules, brushless motors, and lithium-polymer batteries reduced the cost of a capable drone from millions to hundreds of dollars. The result: non-state actors now deploy drone swarms, a capability that existing arms control regimes were not designed to address (Scharre, 2018).

3.3 Generalizable Propositions

From these cases, we extract three propositions that apply to the AI Efficiency Shock:

Proposition 1 (Threshold Collapse): Once the resource threshold for a strategic capability falls below the level accessible to non-state or small-state actors, the capability becomes structurally uncontrollable through supply-side regulation.

Proposition 2 (Regulatory Lag): Institutions designed around chokepoint control cannot self-reform fast enough to address diffusion-driven threats, because the institutional logic (controlling inputs) is incompatible with the new reality (ubiquitous outputs).

Proposition 3 (Irreversibility): Efficiency gains, once demonstrated, cannot be “unlearned.” The knowledge of how to achieve high capability at low cost diffuses through open technical discourse and becomes part of the permanent technological commons.

These propositions are not intended as unfalsifiable axioms but as testable claims that can be evaluated against empirical evidence from the AI domain and beyond.

3.4 Boundary Conditions and Scope

The Efficiency Shock framework is not, however, universal. We acknowledge three boundary conditions under which chokepoint-based governance may retain partial efficacy:

Data as a residual chokepoint. While algorithmic efficiency can substitute for raw compute, the quality and specificity of training data remain significant barriers in certain domains. Models trained on publicly available web data may achieve general reasoning capabilities, but specialized applications (e.g., classified intelligence analysis, proprietary drug design) may still depend on proprietary datasets that are not subject to open diffusion. To the extent that data—not compute—becomes the binding constraint, chokepoint governance may persist in narrow verticals.

Energy and infrastructure constraints. Even efficient models require electricity and hardware for inference at scale. States that can control energy infrastructure or impose high costs on computational hosting retain some leverage over *deployment density*, if not over the existence of the capability itself. This is particularly relevant for latency-sensitive or high-throughput applications (e.g., real-time surveillance, autonomous weapons). Moreover, the Efficiency Shock primarily concerns *training* costs and capability thresholds. Large-scale *inference*—deploying a model for millions of concurrent queries—still requires substantial compute infrastructure, providing a residual chokepoint for high-throughput malicious applications such as personalized disinformation at scale.

Regulatory adaptation speed. Our Proposition 2 (Regulatory Lag) assumes that institutions reform slowly relative to technology. This is an empirically common but not inevitable condition. In cases where regulatory bodies adopt agile, iterative rulemaking processes—such as the UK’s “pro-innovation” regulatory framework or Singapore’s model AI governance framework—the lag may be compressed, though not eliminated.

These boundary conditions do not invalidate the Efficiency Shock framework; rather, they specify the conditions under which its predictions are strongest (general-purpose capabilities, open diffusion channels, slow-moving regulatory institutions) and weakest (domain-specific capabilities requiring proprietary data, infrastructure-constrained deployment, adaptive governance regimes).

Box 1: Toward an Efficiency Shock Index (ESI)—A Measurement Blueprint

To operationalize the concept of “Efficiency Shock” for future empirical research, we propose a preliminary measurement framework based on three proxy variables:

Variable 1: Cost-to-Capability Ratio (CCR). Defined as the ratio of estimated training cost (in USD or GPU-hours) to a standardized capability benchmark score (e.g., composite of MMLU, HumanEval, MATH-500, AIME). A declining CCR over time indicates an active Efficiency Shock. Formally: $CCR_t = \text{TrainingCost}_t / \text{CapabilityBenchmark}_t$. An Efficiency Shock is detected when: $\Delta CCR / \Delta t < -k$, where k is a threshold calibrated to historical baselines.

Variable 2: Hardware Accessibility Score (HAS). Defined as the minimum hardware tier required to run a model at a given capability level for inference. Categories: (a) consumer-grade GPU (e.g., RTX 4090), (b) workstation-grade (e.g., A6000), (c) data center single-node (e.g., 8×H100), (d) multi-node cluster. A shift from category (d) to (a) for a given capability level constitutes an accessibility shock.

Variable 3: Diffusion Velocity (DV). Defined as the rate at which derivative models (fine-tuned, quantized, merged, distilled variants) appear on public repositories following the release of a base model. Formally: $DV = \Delta \text{DerivativeModels} / \Delta t$ (measured in models per week post-release). High DV indicates rapid capability diffusion beyond the control of the original provider.

A composite Efficiency Shock Index (ESI) could be constructed as a weighted function of these three variables. We do not claim to provide a definitive metric here, but rather a *measurement blueprint* that future empirical research can calibrate, validate, and refine. The ESI is intended as a comparative heuristic rather than a definitive index; its value lies in making efficiency shocks *commensurable* across domains and time periods.

4 Open-Weight Models as a “Slippery” Regulatory Object

4.1 From Services to Artifacts: An Ontological Shift

Before analyzing specific regulatory failures, it is necessary to identify the deeper ontological shift that the Efficiency Shock produces. The governance of AI has, to date, been organized around the assumption that AI is a *service*—something accessed via an API, mediated by a corporate intermediary, and subject to terms of use, content filters, and usage policies. This “AI-as-a-Service” model maps neatly onto existing paradigms of platform regulation (Gillespie, 2018).

Open-weight models represent a fundamentally different category: *AI-as-an-Artifact*. An artifact, once distributed, exists independently of its creator. It can be copied, modified, recombined, and deployed on infrastructure entirely opaque to the original developer. This distinction has profound regulatory implications, as Table 1 illustrates.

This table illustrates a fundamental category error in current regulation: applying platform governance tools to an artifact distribution problem. The correct analogy for open-weight governance is not “regulating Facebook” but “regulating PGP encryption”—a comparison that, historically, did not end well for regulators.

Table 1: AI-as-a-Service vs. AI-as-an-Artifact: Regulatory Implications

Dimension	AI-as-a-Service (Closed API)	AI-as-an-Artifact (Open Weights)
Control Chain	Provider retains kill switch	Control severed at download
Recallability	Provider can patch, filter, or disable	Impossible to recall distributed weights
Liability Locus	Provider is clear “cheapest cost avoider”	Liability dispersed across fine-tuners, hosts, users
Regulatory Model	Platform regulation (TOS, content moderation)	Closer to arms/crypto regulation (diffusion control)
User Visibility	Provider logs all queries	Usage entirely opaque
Modification	Users cannot alter model	Users can fine-tune, merge, quantize freely

4.2 Regulatory Dysgraphia and the EU AI Act

Current AI regulations suffer from what we term *Regulatory Dysgraphia*: the structural inability of legal texts, drafted under the assumptions of the platform economy, to “read” and capture the technical reality of the system they seek to govern. Regulatory dysgraphia, as we use the term, extends existing concerns about law’s difficulty in “reading” algorithmic systems (Brownsword & Goodwin, 2012; Burrell, 2016; Pasquale, 2015) to a context in which the core regulatory object—the model weights—can be endlessly forked, recombined, and redeployed.

The European Union’s AI Act (European Parliament, 2024)—the world’s most comprehensive attempt at algorithmic constitutionalism—constructs its entire liability regime around the legal fiction of the “Provider” (Article 3(3)). The Provider is conceptualized as a centralized control node: an entity that develops an AI system and places it on the market or puts it into service. According to the European Commission’s official FAQ on General-Purpose AI models, the “provider of a GPAI model” is defined as the natural or legal person who develops or commissions the development of such a model (European Commission, 2024). This definition assumes a stable, linear chain of custody: the Provider trains the model, the Deployer implements it, and the User interacts with it. This “waterfall” governance model is inherited directly from product safety regulations (e.g., the Machinery Directive), where physical goods have clear manufacturers and immutable properties.

It is important to distinguish between “open-source” and “open-weight” in this context. DeepSeek R1 is an *open-weight* model—the trained parameters are publicly downloadable—but it is not fully open-source in the conventional software sense: the training data recipes, RLHF reward model specifications, and infrastructure configurations are not disclosed. This “semi-transparent” state exacerbates the regulatory challenge: downstream fine-tuners lack the information needed to conduct meaningful conformity assessments, while the original provider can disclaim responsibility for modifications made by actors with incomplete visibility into the base model’s properties.

Box 2: Why Open-Weight Models Break the EU AI Act’s Liability Architecture

Legal Assumption 1: The Provider is identifiable and accountable. The EU AI Act presumes that a single entity—the Provider—bears primary responsibility for compliance (conformity assessments, risk management, quality management systems under Article 17). In the open-weight paradigm, when a lab releases weights under a permissive license (MIT, Apache 2.0), the “original provider” initiates the chain. But the moment weights are downloaded and fine-tuned by a developer in São Paulo on local legal data, the control tether is cut. Is this downstream actor a “Deployer” (minimal obligations) or a new “Provider” (full compliance burden)? The Act provides no clear answer for this gray zone.

Legal Assumption 2: Systemic risk correlates with compute scale. The EU AI Act’s GPAI (General-Purpose AI) provisions classify models as posing “systemic risk” when trained with cumulative compute exceeding 10^{25} FLOPs (European Parliament, 2024, Article 51). The Code of Practice guidelines further elaborate the obligations triggered by this threshold (AI Act Code of Practice, 2024). This threshold was designed to capture only the largest frontier models (GPT-4 class and above). The Efficiency Shock renders this threshold structurally obsolete: a model trained *below* the FLOPs threshold—through algorithmic efficiency, distillation, or synthetic data—may nonetheless exhibit systemic-risk-level capabilities. The legal net is calibrated to catch whales (high-compute models), but the ocean is filling with sharks (high-efficiency models) that swim through the mesh.

Legal Assumption 3: Models can be recalled or patched. Product safety regulation assumes that defective products can be recalled. Open-weight models, once distributed via torrent or direct download, cannot be recalled. A “safety failure” in a neural network is a distributed property of the weight matrix, not a discrete bug in readable code. “Patching” requires re-training or fine-tuning, which creates a *different* model version. The state cannot “turn off” weights that exist on thousands of hard drives worldwide.

4.3 The Liability Vacuum and Distributed Agency

This structural decoupling creates a “Liability Vacuum.” In tort law, liability typically attaches to the entity best positioned to prevent the harm—the “cheapest cost avoider” (Calabresi, 1970). In the closed-source API model (e.g., GPT-4 via OpenAI), the provider is clearly that entity: it controls access, filters outputs, logs usage, and can disable the service. In the open-weight paradigm, no single actor occupies this position. The original developer has released the weights; the fine-tuner has modified them for a specific purpose; the infrastructure host provides compute; the end-user generates the harmful output. Responsibility is distributed across a network of partial contributors, none of whom holds full accountability.

This is not a novel problem in law—the concept of “distributed agency” has been analyzed in environmental liability and supply chain regulation (Locke, 2013). However, open-weight AI models exacerbate it to an unprecedented degree because the “product” is infinitely replicable at zero marginal cost, modification requires no permission from the original creator, and the modified downstream artifact bears no trace of its provenance unless deliberately watermarked (which current open-weight releases are not).

4.4 Design Space for Post-Provider Liability

Given the structural inadequacy of provider-centric liability, we map a design space of four alternative or complementary liability architectures for open-weight ecosystems. These are presented as ideal types, not policy prescriptions; their relative viability is an empirical question for future research.

Model A: Strict Usage Liability. Liability attaches exclusively to the end-user who deploys the model for harmful purposes. The model creator and distributor bear no liability. This mirrors the legal treatment of dual-use tools (e.g., knives, encryption software) and maximizes innovation incentives. *Weakness:* enforcement depends on identifying and prosecuting end-users, which is difficult when usage is anonymous and cross-jurisdictional.

Model B: Intermediary/Host Liability. Liability is shared by infrastructure providers (cloud hosts, model repositories, fine-tuning platforms) who facilitate access to open-weight models. This mirrors the intermediary liability framework of the Digital Services Act (DSA) for online platforms. *Weakness:* it faces the same scaling challenge as content moderation—millions of model variants cannot be individually screened—and risks driving distribution to unregulated channels (torrents, IPFS).

Model C: Provenance and Audit Duties. Model creators are required to embed machine-readable

provenance metadata (training data summary, capability profile, intended use boundaries) in the model weights or accompanying documentation. Downstream users who strip or ignore provenance data face heightened liability. This is analogous to “nutrition labeling” for AI. *Weakness*: provenance can be stripped or forged; enforcement requires global consensus on metadata standards.

Model D: High-Risk Domain Licensing. For specific high-risk application domains (biological research, autonomous cyber operations, critical infrastructure control), a licensing regime governs who may deploy open-weight models. This is analogous to controlled substance licensing in pharmaceuticals. *Weakness*: it shifts the regulatory burden to the deployment context, requiring domain-specific regulators (bio-safety, nuclear, financial) to develop AI-specific competencies.

No single model will suffice. A realistic post-chokepoint liability regime will likely combine elements of Models A–D, calibrated to the risk profile of the application domain. The key insight is that all four models share a common departure from the status quo: *none assumes a unitary provider as the primary locus of control*.

5 Geopolitical Consequences: From Containment to Fragmented Adaptation

5.1 The Erosion of the “Hardware Wall”

Since the imposition of comprehensive export controls on October 7, 2022, U.S. strategy has been predicated on the “Chokepoint Paradigm”: the belief that by restricting access to the most advanced logic chips (e.g., NVIDIA H100/B200) and semiconductor manufacturing equipment, the U.S. could freeze China’s AI capabilities at a standard generationally behind the frontier (Allen, 2022; Miller, 2022; Federal Register, 2022). The syllogism was clear: *SOTA AI requires SOTA Hardware; SOTA Hardware is controlled by the US/Allies; therefore, China cannot achieve SOTA AI*.

The DeepSeek Moment demonstrates the leakage in this logic. It proves that *software efficiency*—innovations in mixture-of-experts (MoE) architectures, higher-quality synthetic data curation, and distillation techniques—can act as a substitute for raw hardware horsepower. By achieving reasoning benchmarks at strategically relevant levels comparable to OpenAI’s o1 series using legacy or restricted hardware clusters, DeepSeek R1 has empirically weakened the “Hardware Wall” thesis. While export controls impose a “tax” on training speed and cost, they have failed to prevent the emergence of frontier-class capabilities. Theoretical containment has collided with practical diffusion—a dynamic consistent with Proposition 1 (Threshold Collapse) articulated in Section 3.

5.2 The “Fragmented Adaptation” Typology

We are consequently witnessing the rise of “Open-Weight Sovereignty.” For states outside the US–China binary—what we term “swing states” in AI geopolitics—reliance on closed-source US models (via API) presents a dilemma of digital sovereignty (Floridi, 2021). It implies structural dependency on US surveillance infrastructure, US-determined usage policies (AUPs), and the whims of US foreign policy (e.g., potential API cutoffs during geopolitical crises).

Open-weight models offer an escape hatch from this dependency. They allow nations to build “Sovereign AI” infrastructure: models hosted on domestic servers, aligned with local laws and cultural values, and immune to extraterritorial remote disablement. We propose a typology of “Fragmented Adaptation,” structured along two analytical dimensions: (1) the degree of *regulatory control* a state seeks over AI capabilities (preventive ↔ reactive), and (2) the degree of *sovereignty orientation* (dependency on external providers ↔ autonomous hosting).

This yields three ideal-typical blocs:

Bloc 1: The Containment Bloc (US & Close Allies). Characterized by “Know Your Customer” (KYC) cloud regulations, strict liability for model developers, and attempts to restrict the release of model weights above certain capability thresholds. Policy instruments include export controls (BIS Entity List), executive orders mandating safety evaluations for frontier models (EO 14110), and proposed legislation requiring licensing for weight releases. The normative frame is “AI Safety”—the prevention of catastrophic misuse.

Bloc 2: The Sovereign Bloc (EU Non-Aligned, India, UAE, Brazil, ASEAN). Characterized by the strategic adoption of open-weight models to build national AI infrastructure independent of both US and Chinese cloud ecosystems. Policy instruments include government-funded fine-tuning of open-weight models on national datasets, sovereign cloud initiatives, and selective participation in international AI governance forums without binding commitments. The normative frame is “Digital Sovereignty” and “Development Rights.”

Bloc 3: The Open Commons (Transnational, Decentralized). Characterized by the circulation of model weights via censorship-resistant protocols (torrents, IPFS, decentralized storage), anonymous fine-tuning and re-hosting, and a cultural ethos of radical openness inherited from the free software movement. This bloc has no formal state sponsors but is enabled by the technical architecture of the internet. The normative frame is “Information Freedom”—the belief that model weights, like cryptographic code, are a form of speech.

5.3 Observable Implications by Bloc

To make this typology empirically productive, we specify observable implications—predictions that future research can evaluate:

Containment Bloc predictions: Increasing investment in “KYC for compute” regulations requiring cloud providers to verify customer identity and intended use. Legislative proposals to restrict or license the release of model weights above specified capability thresholds. Diplomatic pressure on allied states to adopt compatible weight-release restrictions.

Sovereign Bloc predictions: Government funding for national fine-tuning of open-weight foundation models on domestic-language and domain-specific datasets. Establishment of “sovereign cloud” initiatives explicitly designed to host open-weight models outside US/Chinese infrastructure. Resistance to binding international agreements that would restrict access to open-weight models.

Open Commons predictions: Continued growth of decentralized model distribution channels (torrents, IPFS, peer-to-peer networks) that are technically resistant to state-level takedown. Proliferation of “uncensored” model variants (fine-tuned to remove safety guardrails) as a countercultural response to Containment Bloc restrictions. Emergence of anonymous communities specializing in model merging, quantization, and capability extraction from restricted models via distillation.

5.4 The End of the “Global AI Treaty” Dream

In this fragmented world, the prospect of a unified “Global AI Treaty” akin to the IAEA for nuclear energy or the OPCW for chemical weapons is unlikely under current diffusion conditions (Maas, 2023). The diffusion enabled by the DeepSeek Moment means that strategic capability is too widespread, and the capital barriers too low, for a single non-proliferation regime to hold. Unlike nuclear materials, model weights are infinitely copyable at zero marginal cost. Unlike chemical precursors, they leave no physical trace when transported across borders. The material conditions for a chokepoint-based non-proliferation regime simply do not exist.

This does not mean that international cooperation is impossible—but it means that cooperation must focus on *consequences* (harm mitigation, incident response) rather than *prevention* (restricting access to the

capability itself). The analogy shifts from the Nuclear Non-Proliferation Treaty to the Budapest Convention on Cybercrime: an imperfect but functional framework for coordinating *response* to misuse of an inherently distributed technology.

6 Conclusion: Beyond the Chokepoint

6.1 Summary of Argument

The DeepSeek Moment serves as a definitive wake-up call for the epistemic community of AI governance. The comfortable assumption that we can govern artificial intelligence by regulating a handful of San Francisco-based CEOs and controlling the flow of silicon through the Taiwan Strait is no longer tenable. The Chokepoint Era is eroding. The Efficiency Shock has lowered the price of intelligence; governance must now find ways to raise the cost of misuse, for it can no longer restrict the supply of the machine itself.

6.2 Policy Design Principles

Future governance must accept the reality of “Uncontrollable Diffusion.” If we cannot prevent the *existence* and *access* to powerful models, the focus of policy must shift from *prevention* to *resilience*. We propose three design principles for post-chokepoint governance:

Principle 1: From Supply Control to Impact Monitoring. Rather than attempting to restrict the distribution of the “tool” (which is now technically impossible for open-weight models), governance should focus on monitoring the “endpoint”—hardening financial systems, biological research infrastructure, critical national infrastructure, and electoral systems against the *attacks* AI might enable (Shevlane et al., 2023). This requires investment in detection capabilities (e.g., AI-generated content forensics, anomalous behavior detection in bio-labs) rather than compliance certifications for model developers.

Principle 2: From Provider Liability to Adaptive Usage Liability. Legal regimes must pivot from “Provider Liability” (which assumes a control chain that no longer exists) to calibrated “Usage Liability,” punishing the malicious actor rather than the toolmaker. This acknowledges the dual-use nature of open weights—the same model that assists in drug discovery can assist in synthesizing toxins—and allocates responsibility based on *intent and action* rather than *creation and distribution*. We note that this transition is not without risks: usage-based monitoring can itself become a vector for surveillance overreach (see Section 4.4, Model D).

Principle 3: Cognitive Security as Infrastructure. As the cost of generating high-persuasion content (text, audio, video) falls to near-zero, the defense of democratic discourse can no longer rely on content filtering (which is a supply-side strategy). Instead, governance must invest in “cognitive security” infrastructure: technologies and institutions that verify human authenticity (e.g., cryptographic proof-of-personhood, hardware-level watermarking of AI-generated content) and build societal resilience to manipulation (Weidinger et al., 2022).

6.3 Research Agenda

This article opens several avenues for future research. First, *empirical measurement of the Efficiency Shock*: operationalizing the Efficiency Shock Index (Box 1) through longitudinal data collection on training costs, capability benchmarks, hardware accessibility, and diffusion velocity, enabling *ex ante* identification of future “threshold collapse” events. Second, *comparative regulatory analysis*: systematic comparison of how different jurisdictions (EU, US, China, India, Brazil) are adapting their legal frameworks to open-weight models, with attention to the specific mechanisms of regulatory lag. Third, *open-weight liability regimes*:

legal scholarship developing viable alternatives to provider-centric liability for open-weight ecosystems (see Section 4.4), drawing on distributed agency theory and environmental law precedents. Fourth, *the political economy of the “Sovereign Bloc”*: detailed case studies of how “swing states” (UAE, India, France, Brazil) are strategically leveraging open-weight models to pursue digital sovereignty objectives. Fifth, *effectiveness of impact-based governance*: evaluating whether endpoint monitoring and usage liability actually deter harmful applications of open-weight models, or whether they produce their own forms of regulatory failure (e.g., surveillance overreach, chilling effects on open research).

6.4 Implications for Three Communities

This framework is intended to inform three distinct communities:

For regulators and policymakers: The chokepoint toolkit (export controls, provider licensing, compute thresholds) is losing efficacy. The shift to impact-based governance is not optional—it is being imposed by technological reality.

For legal scholars: The “provider” as a singular locus of liability is a legal fiction that the open-weight paradigm has exposed. New doctrines of distributed responsibility, drawing on product liability, environmental law, and intermediary liability, are urgently needed. The design space mapped in Section 4.4 provides a starting framework for this inquiry.

For international relations scholars: The assumption that AI can be governed through a non-proliferation model (restricting access to the capability) must be replaced by a harm-management model (coordinating response to the consequences). The relevant precedent is not the NPT but the Budapest Convention.

References

- AI Act Code of Practice. (2024). *General-Purpose AI Code of Practice: First Draft*. European AI Office.
- Albright, D., & Hinderstein, C. (2005). Unraveling the A. Q. Khan and future proliferation networks. *The Washington Quarterly*, 28(2), 111–128.
- Allen, G. C. (2022). *Choking off China’s access to the future of AI*. Center for Strategic and International Studies.
- Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.
- Bureau of Industry and Security. (2022). *Implementation of additional export controls: Certain advanced computing and semiconductor manufacturing items*. U.S. Department of Commerce.
- Brownsword, R., & Goodwin, M. (2012). *Law and the Technologies of the Twenty-First Century*. Cambridge University Press.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- Calabresi, G. (1970). *The Costs of Accidents: A Legal and Economic Analysis*. Yale University Press.
- Capoccia, G., & Kelemen, R. D. (2007). The study of critical junctures: Theory, narrative, and counterfactuals in historical institutionalism. *World Politics*, 59(3), 341–369.
- Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford.
- Danaher, J., Hogan, M. J., Noone, C., et al. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society*, 4(2). <https://doi.org/10.1177/2053951717726554>
- DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Ding, J., & Dafoe, A. (2023). Engines of power: Electricity, AI, and general-purpose military transformations. *European Journal of International Security*, 8(3), 319–340.
- European Commission. (2024). *General-Purpose AI Models in the AI Act: Questions and Answers*. Digital Strategy,

European Commission.

- Engler, A. (2022). The EU AI Act will have global impact, but a limited Brussels Effect. *Brookings Institution*.
- Erman, E., & Furendal, M. (2022). Artificial intelligence and the political legitimacy of global governance. *Political Studies*, 72(2), 421–441.
- European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689.
- Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security*, 44(1), 42–79.
- Federal Register. (2022). Implementation of additional export controls: Certain advanced computing and semiconductor manufacturing items. *Federal Register*, 87(197), 62186–62218.
- Floridi, L. (2021). The European legislation on AI: A brief analysis of its philosophical approach. *Philosophy & Technology*, 34(2), 215–222. <https://doi.org/10.1007/s13347-021-00460-9>
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, 8(4). <https://doi.org/10.14763/2019.4.1424>
- Levy, S. (2001). *Crypto: How the Code Rebels Beat the Government—Saving Privacy in the Digital Age*. Penguin.
- Locke, R. M. (2013). *The Promise and Limits of Private Power: Promoting Labor Standards in a Global Economy*. Cambridge University Press.
- Maas, M. M. (2023). Aligning AI regulation to sociotechnical change. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance*. Oxford University Press.
- Mahoney, J. (2000). Path dependence in historical sociology. *Theory and Society*, 29(4), 507–548.
- Miller, C. (2022). *Chip War: The Fight for the World's Most Critical Technology*. Scribner.
- Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.
- Scharre, P. (2018). *Army of None: Autonomous Weapons and the Future of War*. W.W. Norton.
- Shevlane, T., Farquhar, S., Garfinkel, B., et al. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.
- Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W. G. (2021). Framing governance for a contested emerging technology: Insights from AI policy. *Policy and Society*, 40(2), 158–177.
- UK Government. (2023). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023*.
- Weidinger, L., Mellor, J., Rauh, M., et al. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). ACM.
- Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4), 505–523.

Handling Editor: Changkui LI (cpcl@cpcl.hk)