

Global AI Safety Governance and Executive Accountability Architecture: A Framework for Unstoppable Technological Progress

Executive Summary

The rapid advancement of frontier artificial intelligence presents a structural governance crisis. Frontier AI capabilities are progressing along continuous scaling vectors, driven by massive compute clusters and expanding algorithmic efficiency. Because these technological advancements yield immense economic and national security advantages, no sovereign nation or market leader can realistically cease development without incurring unacceptable strategic costs. Consequently, regulatory strategies predicated on permanent moratoriums, total technological containment, or mandatory model shutdowns are fundamentally non-viable. At the same time, existing legal and regulatory instruments are structurally ill-equipped to govern frontier AI. Software-based intelligence is non-rivalrous, highly mobile, and susceptible to rapid replication, modification, and uncensored fine-tuning. The central dilemma of frontier AI governance lies in the accountability paradox: imposing strict liability on developers for any downstream misuse or jailbreak disincentivizes transparent security research and suppresses vital red-teaming. Conversely, providing blanket immunity shields reckless development practices, privatizing corporate profits while externalizing catastrophic tail risks onto society. This report presents a global AI safety architecture designed to operate under the assumption that capability scaling, open-weight proliferation, and adversarial security testing will persist. Drawing on lessons from nuclear safeguards, arms control treaties, biological weapons conventions, financial systemic risk supervision, and civil aviation safety, this study identifies physical compute hardware and cryptographic verification as the sole non-reproducible bottlenecks capable of serving as the foundation for global governance. The recommended framework—the Global Hardware-Cryptographic and Tiered Liability Architecture (GHC-TLA)—pairs hardware-enabled compute telemetry and zero-knowledge training proofs with a multi-tiered liability structure modeled on the Price-Anderson Nuclear Industries Indemnity Act. Under this model, developers receive a statutory liability safe harbor for misuse only if they demonstrate compliance with independently audited, cryptographically verified safety standards and contribute to a mandatory pool for catastrophic losses. International enforcement relies on chip supply-chain controls, on-die verification interlocks, and a network of national AI Safety Institutes coordinated under a binding treaty framework.

The AI Accountability Paradox

Governing frontier AI requires isolating where risk originates across the complex lifecycle of advanced foundation models. Unlike conventional industrial goods, an AI system's latent capabilities do not map cleanly to its initial design intent. Risks emerge from complex interactions across distinct layers of the technological stack, requiring specific, tailored

governance interventions for each operational category.

Taxonomy of AI Risk Across Five Functional Axes

To establish targeted liability and regulatory mechanisms, AI risks must be separated into five distinct operational domains based on their primary causal origin.

Risk Axis	Specific AI Risk Categories Covered	Primary Causal Mechanism	Appropriate Governance Focus
A. Model-Level Risks	Model-level vulnerabilities, dangerous capability elicitation, autonomous replication, accidental catastrophic behavior	Emergent neural network properties, architectural alignment failures, and pre-training dataset hazards	Pre-training evaluations, architectural alignment audits, and training dataset sanitization
B. Deployment Risks	Model weight proliferation, open-source proliferation, concentration of power, systemic economic risk	Choices made by developers regarding distribution channels, API access, and raw weight releases	API access rate-limiting, cloud deployment oversight, and weight release safety gates
C. User-Driven Misuse	Jailbreaks, cyber capabilities, biological misuse, AI-assisted crime, deliberate malicious use	Adversarial input crafting, prompt injection, uncensored fine-tuning, and malicious agent deployment	User identity verification, output filtering, biological order screening, and red-teaming
D. State & Military Risks	Military applications, model theft, critical infrastructure attacks, misinformation/manipulation	Sovereign strategic deployment, state-sponsored cyber warfare, and intelligence operations	Bilateral arms control, strategic compute limits, hotlines, and state counter-intelligence
E. Systemic Multi-Agent Risks	Autonomous agents, interaction between AI systems, cascading economic failures	Unintended dynamic feedback loops between semi-autonomous systems operating in shared environments	Market circuit breakers, agent transaction limits, and cross-system interoperability controls

A. Risks Caused by the Model Itself

These risks stem directly from architectural design, pre-training dataset hazards, or intrinsic alignment failures within the neural network weights. They include emergent dangerous capabilities such as advanced biological design reasoning, autonomous self-exfiltration tendencies, architectural vulnerabilities to adversarial perturbation, and unexpected step-function capability jumps during training runs. Because these capabilities exist intrinsically within the trained model, they represent latent hazards that persist prior to public deployment.

B. Risks Caused by Deployment

Deployment risks arise from the distribution strategies chosen by model developers and infrastructure providers. These encompass the unmitigated release of open weights, inadequate Application Programming Interface (API) monitoring, insufficient rate-limiting, raw parameter proliferation, and the rapid integration of foundation models into critical infrastructure without redundant manual overrides. Once raw model weights are broadly distributed, ex-post recall becomes technically impossible.

C. Risks Caused by Users

User-driven risks involve deliberate adversarial attempts by external actors to bypass alignment guardrails or exploit model reasoning for harmful ends. Methods include complex prompt injection, jailbreaking, fine-tuning open-weight models via Low-Rank Adaptation (LoRA) to remove refusal behaviors, and deploying autonomous user-created agents to carry out cyberattacks or generate synthetic disinformation.

D. Risks Caused by Governments or Militaries

These risks stem from state-sponsored activities, including the integration of foundation models into nuclear command-and-control loops, direct cyber warfare, autonomous targeting platforms, foreign political manipulation, and state theft of intellectual property or model weights. In this domain, risks are driven by strategic national competition rather than commercial motives.

E. Risks Caused by Interactions Between AI Systems

Systemic risks emerge when multiple semi-autonomous or autonomous AI systems interact dynamically within complex, high-speed environments. Examples include flash crashes in financial markets caused by algorithmic trading agents, multi-agent strategic escalation during diplomatic crises, and cascading failures across automated supply chains or energy distribution grids.

The Core Accountability Paradox

The taxonomy of risk exposes the fundamental accountability gap facing frontier lab executives. Under current corporate legal frameworks, holding a Chief Executive Officer or corporate entity strictly liable for every harm generated downstream by an AI model creates a severe structural dilemma.

If legal systems impose strict liability for every jailbreak or malicious output, frontier labs face existential financial risk whenever security researchers discover latent vulnerabilities. Corporate executives will rationally respond by restricting independent red-teaming, enforcing non-disclosure agreements on researchers, closing evaluation APIs, and obfuscating model safety assessments. This opacity suppresses the exact security research required to identify and patch dangerous capabilities before bad actors exploit them.

Conversely, if governments grant blanket liability immunity to shield innovation, AI safety becomes purely voluntary. Corporate executives facing intense market competition will prioritize rapid capability deployment over rigorous pre-training evaluations, externalizing catastrophic tail risks—such as an uncensored model enabling a non-state actor to acquire pandemic-class

biological agents—onto the general public. Voluntary industry commitments consistently break down under commercial pressures, leaving society unprotected against severe negative externalities.

Why Traditional Regulation Is Insufficient

Traditional regulatory tools rely on static statutory rules, ex-post judicial litigation, and localized territorial enforcement. These mechanisms encounter structural barriers when applied to frontier software intelligence.

Traditional Regulatory Assumption	Frontier AI Reality	Resulting Regulatory Failure
Static Rulemaking Cycles	Exponential capability scaling (compute doubles every 5–6 months)	Rules become obsolete before legislative enactment
Localized Physical Assets	Digital model weights are highly mobile and non-rivalrous	Regulatory arbitrage and offshore cloud training shifts
Ex-Post Civil Tort Liability	Catastrophic losses far exceed corporate balance sheets	Corporate judgment-proofness externalizes tail risk
Clear Territorial Jurisdiction	Borderless distribution via cloud networks and decentralized mirrors	Single-jurisdiction bans fail to stop global access

Exponential Pace versus Linear Legislation

Administrative rulemaking and legislative enactment operate on multi-year timeframes. In contrast, compute allocation for state-of-the-art training runs has historically doubled roughly every five to six months, driving rapid, step-function jumps in capabilities. Statutory definitions anchored to specific parameter counts or fixed architectural designs become outdated prior to implementation.

Jurisdictional Arbitrage and Weight Proliferation

Unlike heavy industrial assets, software parameter weights are non-rivalrous, highly compressed, and easily transmitted over encrypted network channels. If a jurisdiction imposes stringent pre-training authorization requirements or bans high-risk capability deployment, developers can shift training runs to cloud infrastructure located in permissive foreign jurisdictions. Furthermore, once model weights are released publicly as open weights, they can be permanently duplicated, mirrored, and modified globally, making ex-post enforcement within a single jurisdiction ineffective.

The Judgment-Proof Corporate Entity Problem

Conventional product liability relies on civil tort law to compel manufacturers to internalize safety risks through potential monetary damages. However, the upper-bound tail risks associated with frontier AI—such as critical infrastructure shutdowns or widespread biological attacks—could inflict damages valued in hundreds of billions or trillions of dollars. Because even the largest technology firms possess balance sheets that are small relative to these catastrophic tail risks, corporate entities become effectively "judgment-proof". The threat of bankruptcy fails to ensure

adequate pre-training caution, as the firm cannot absorb the full cost of a worst-case scenario.

What Nuclear Governance Actually Teaches Us

To establish a functional architecture for AI governance, policymakers must examine how prior international regimes successfully regulated transformative, dual-use, or uncontrollable technologies.

Underlying Mechanisms of Nuclear Governance

Nuclear governance under the Nuclear Non-Proliferation Treaty (NPT) and the International Atomic Energy Agency (IAEA) succeeded not because of abstract international goodwill, but due to structural mechanisms inherent to nuclear physics and state power.

1. **Scarcity and Physical Bottlenecks:** The enrichment of fissile material (Uranium-235 and Plutonium-239) requires complex industrial infrastructure, such as gas centrifuge cascades and heavy-water reactors. Fissile material is scarce, highly detectable, and impossible to synthesize without massive physical facilities.
2. **Material Accountancy and Safeguards:** The IAEA enforces non-proliferation through strict material accountancy, tracking enriched material down to gram-level quantities. Environmental sampling, seals, continuous camera surveillance, and verified material balances ensure that civilian fuel cannot be diverted to weapons programs without immediate detection.
3. **Intrusive Verification Rights:** Signatory states grant the IAEA legal authority to conduct routine and short-notice inspections of declared nuclear facilities.
4. **Reciprocal Technology Transfer:** The NPT structure created a self-interest-compatible bargain: non-nuclear weapon states accepted verification and pledged not to develop nuclear weapons in exchange for access to peaceful civilian nuclear power technology and fuel supply guarantees.

Transferable Lessons from Other Historical Regimes

In addition to nuclear governance, several historical frameworks offer transferable mechanisms for AI governance:

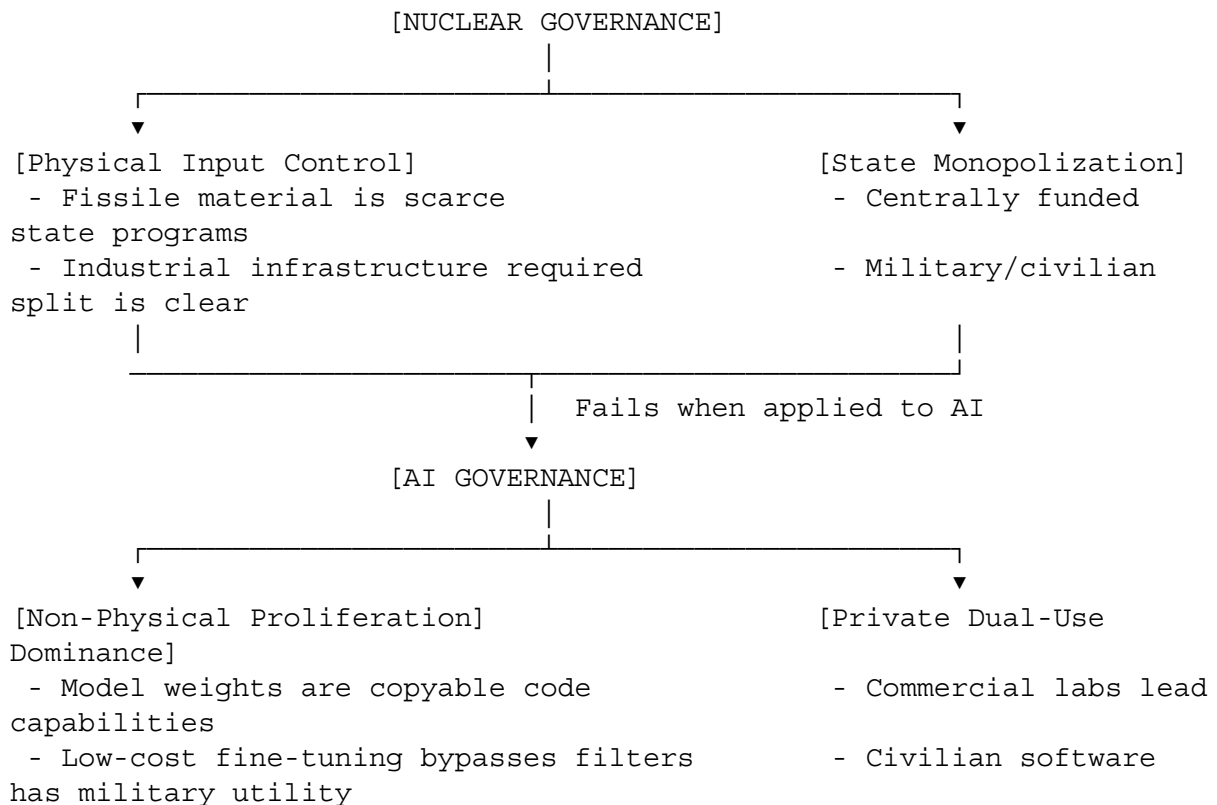
- **Chemical Weapons Convention (CWC):** Demonstrates the utility of Schedule-based precursor tracking and mandatory Challenge Inspections under Article IX, where any state party can demand an immediate, non-refusable inspection of a facility suspected of illicit chemical synthesis.
- **Systemically Important Financial Institutions (SIFI):** Provides a precedent for managing systemic tail risk through mandatory capital buffers, forward-looking stress testing under extreme crisis scenarios, and pre-approved resolution plans ("living wills").
- **International Civil Aviation Organization (ICAO):** Illustrates how universal technical baselines can be enforced through domestic authorities, supported by non-punitive incident reporting systems (such as the US Aviation Safety Reporting System) and independent root-cause investigation bodies like the National Transportation Safety Board (NTSB).

What Cannot Be Transferred from Nuclear

Governance

While nuclear treaties offer useful precedent for monitoring physical infrastructure, the unique characteristics of artificial intelligence prevent a direct copy of the IAEA model.

Key Structural Breakdown Points



Non-Physical Proliferation

Fissile nuclear material cannot be transmitted across the internet, copied onto digital storage drives, or executed on consumer electronics. Foundation model weights consist of digital parameters that can be compressed, fine-tuned, or distributed globally via decentralized networks. Once released, model weights cannot be recovered or remotely disabled through traditional physical containment.

Pervasive Dual-Use Nature

In nuclear technology, the enrichment threshold distinguishes civilian power reactor fuel (3–5% U-235) from weapons-grade material (>90% U-235). In frontier AI models, no such architectural boundary exists. The same foundation model that analyzes biological protein structures for drug discovery can be prompted or fine-tuned to design novel biological toxins or map acquisition

routes for regulated pathogens.

Private Sector Preeminence

Nuclear development was initiated and controlled by sovereign nation-states through centralized military programs. Conversely, frontier AI development is driven primarily by private commercial entities with multi-billion dollar capital budgets, proprietary training datasets, and rapid release cadences. Governance models built exclusively for state-to-state diplomatic interaction fail to capture private developer incentives.

Finding the Equivalent of "Nuclear Material"

To establish a functional accounting and inspection regime for AI, governance must anchor to a physical input that is excludable, detectable, and governed by a concentrated global supply chain.

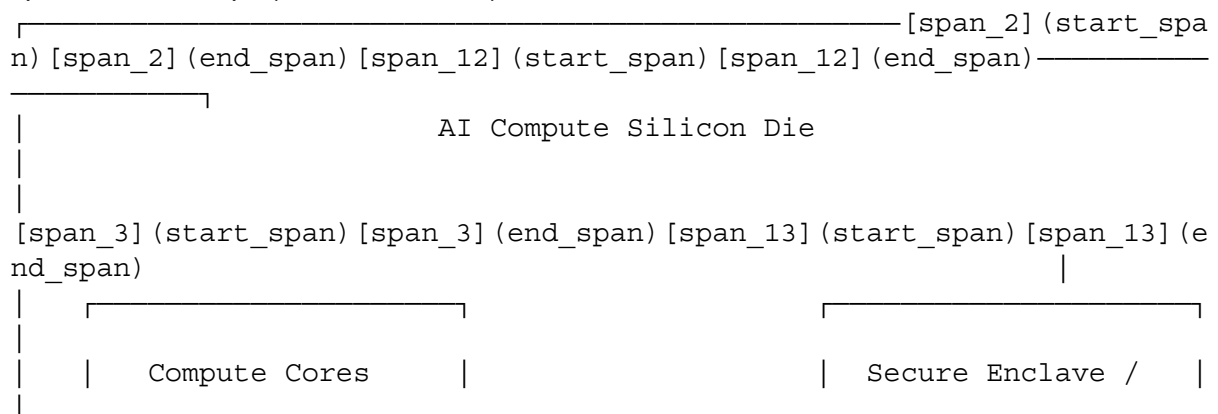
The Centrality of Compute and Microelectronics

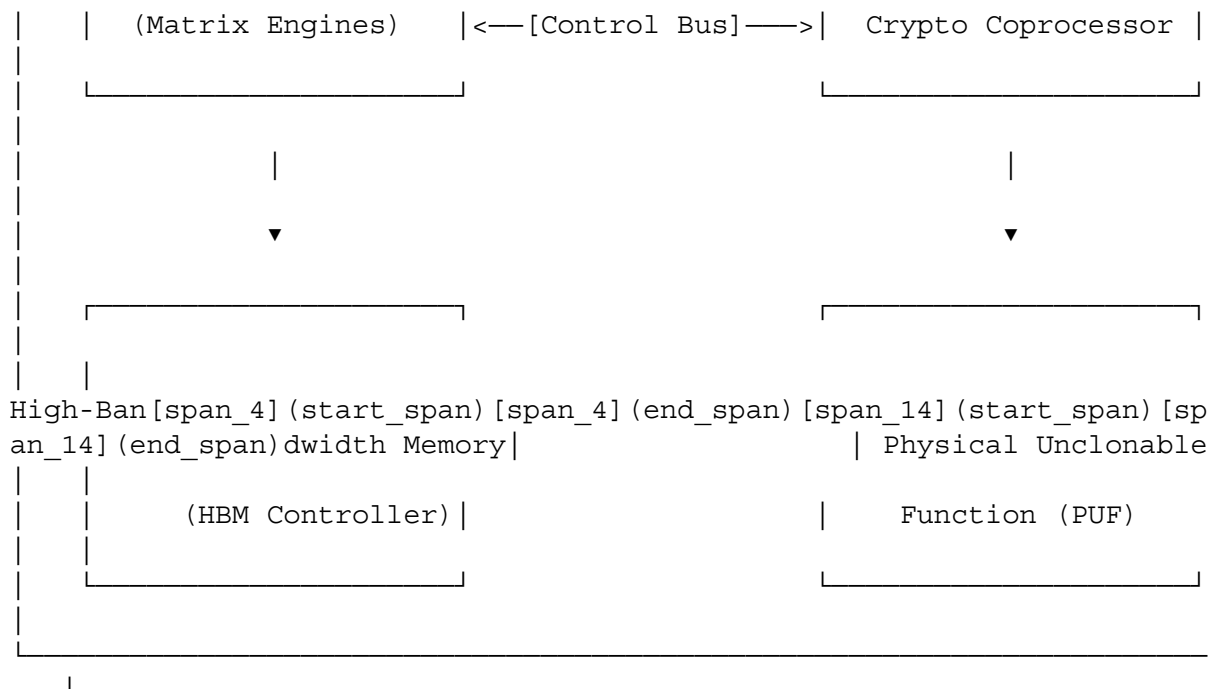
Compute hardware satisfies the criteria for physical accounting:

- **Extreme Supply Chain Concentration:** Advanced AI hardware requires specialized photolithography equipment produced by a single firm (ASML), manufactured in high-end foundries (such as TSMC), and packaged using specialized high-bandwidth memory (HBM) integration.
- **Physical Detectability and Energy Footprint:** Datacenters capable of hosting frontier training runs require hundreds of megawatts of dedicated power infrastructure, specialized cooling setups, and massive physical footprints, making them readily visible to satellite reconnaissance and grid monitoring.
- **Excludability:** Physical chips can be tracked from manufacturing foundries to deployment, subject to export controls, and remotely managed via secure hardware controllers.

Technical Feasibility of Hardware-Enabled Governance (HEGs)

Hardware-Enabled Governance Mechanisms (HEGs) integrate technical safeguards directly into specialized AI chips (GPUs and TPUs) at the silicon level:





On-Chip Cryptographic Telemetry

Integrating secure execution enclaves and Physical Unclonable Functions (PUFs) onto AI chips enables cryptographically signed hardware attestations. The chip periodically generates status reports confirming its physical identity, location, cluster interconnect topology, and cumulative FLOP execution count.

Physical Interconnect Limits

Hardware limits on chip-to-chip interconnect bandwidth prevent unauthorized scaling. If chips are altered or chained into unauthorized superclusters without cryptographic clearance keys, hardware-level throttling reduces interconnect throughput, preventing high-speed distributed training runs.

Multi-Party Authorization Keys

High-compute training runs exceeding defined floating-point operation (FLOP) thresholds (e.g., $>10^{26}$ total FLOPs) require dynamic activation keys issued jointly by national safety authorities. Without periodic signed authorization tokens from regulators, the hardware halts processing.

Cryptographic Training Auditing via Zero-Knowledge Proofs (ZKPs)

A central challenge of compute oversight is verifying that a lab is not training dangerous capabilities without forcing them to reveal proprietary model architecture or trade secrets to government inspectors.

Zero-knowledge proof architectures (such as ZK-SNARKs and ZK-STARKs) address this challenge. A frontier lab generates a zero-knowledge proof of compute execution during training. This cryptographic proof demonstrates to regulators that:

- The training run did not exceed pre-declared cumulative compute thresholds.
- Specified dataset filters (e.g., removing known viral genome sequences or cyber-exploitation payloads) were executed.
- Required safety evaluation benchmarks were periodically calculated during training.

ZKPs confirm compliance without exposing model weights, dataset contents, or exact neural network parameters, resolving the tension between regulatory verification and corporate intellectual property protection.

The Jailbreak and Security Testing Paradox

Adversarial testing, red-teaming, and vulnerability research are essential for identifying latent failure modes prior to deployment. However, unconstrained jailbreaking can enable the extraction and distribution of dangerous dual-use capabilities.

Differentiating Good-Faith Research from Unsafe Proliferation

To maintain security research without enabling hazardous proliferation, legal systems must formalize clear criteria distinguishing protected vulnerability discovery from negligent deployment.

Parameter	Protected Security Research	Negligent Proliferation / Misuse
Operational Scope	Conducted within authorized scope or secure evaluation sandboxes	Uncontrolled probing of production systems without disclosure agreement
Disclosure Protocol	Adheres to Coordinated Vulnerability Disclosure (CVD) timelines	Immediate public release of functional bioweapons or exploit payloads
Testing Intensity	Proof-of-concept constrained to minimal viable demonstration	Exfiltration of full model weights or building automated exploit tools
Evaluator Status	Executed by accredited security evaluators or verified institutions	Unregistered fine-tuning specifically designed to remove safety filters

Structural Safe-Harbor Mechanisms

To protect research activities while upholding public safety, governance frameworks must implement three structural safeguards:

Statutory AI Security Safe Harbor

Modeled on CISA Binding Operational Directive 20-01 and modern vulnerability disclosure standards, statutory reforms must explicitly exempt good-faith security research from criminal prosecution under cybercrime laws (such as the US Computer Fraud and Abuse Act) and civil liability claims. This protection applies provided the researcher adheres to established

coordinated disclosure guidelines.

Accredited Red-Teamer Frameworks and Secure Evaluation Sandboxes

National AI Safety Institutes issue cryptographic credentials to accredited independent red-teaming organizations. Model developers must provide these accredited evaluators with elevated API access or secure enclave execution environments. This setup allows researchers to probe for jailbreaks, emergent reasoning hazards, and capability thresholds without triggering public automated access blocks.

Coordinated Vulnerability Disclosure (CVD) Protocols

When an accredited researcher discovers a critical vulnerability or jailbreak capable of generating high-risk outputs, a mandatory 90-day embargo window takes effect. The developer receives immediate access to the vulnerability data to engineer mitigations. If the developer fails to implement effective patches within the specified timeframe, the findings are escalated to national safety authorities to determine whether public warning or mandatory service suspension is warranted.

Models of Liability and Accountability

Determining appropriate legal liability models for corporate leaders and developers is critical for aligning commercial incentives with public safety.

Comparative Assessment of Seven Liability Models

Model 1: Strict Liability —————→ [High Safety Incentive] —→
[Suppresses Research & Innovation]
Model 2: Statutory Negligence —————→ [Vague Care Standards] —→
[Protracted Civil Litigation]
Model 3: Compliance Safe Harbor —————→ [Strong Audit Process] —→
[Incentivizes Safety Compliance]
Model 4: Price-Anderson Insurance —————→ [Mandated Insurance Pools] ►
[Provides Universal Risk Backstop]

Model 1: Strict Liability

Developers and corporate officers face automatic liability for any harm caused by their models, regardless of negligence or precautions taken.

- *Incentive Effect:* Maximize safety caution, but suppresses open-source research, forces developers to severely restrict API access, and penalizes transparent vulnerability research.

Model 2: Statutory Negligence

Executives are liable only if they fail to meet defined legal standards of care.

- *Incentive Effect*: Encourages adherence to industry baselines, but defining "duty of care" through common-law litigation creates legal uncertainty due to rapid capability shifts.

Model 3: Compliance-Based Safe Harbor

Developers receive statutory protection from third-party misuse claims if they demonstrate compliance with independently verified safety standards.

- *Incentive Effect*: Strong incentive to adopt safety protocols. Requires continuous post-deployment auditing to prevent safe harbors from becoming mere paper compliance.

Model 4: Capability Licensing

Training runs or deployments exceeding defined capability thresholds require formal government permits. * *Incentive Effect*: Prevents unauthorized deployment of high-risk models. Risks regulatory capture by market incumbents unless paired with academic exemptions.

Model 5: Insurance-Based Accountability (The Price-Anderson Model)

Frontier labs must secure mandatory primary liability insurance coverage proportional to model capabilities. Insurance underwriters conduct technical safety audits, pricing premiums based on real-world risk management. Claims exceeding primary coverage are drawn from a secondary retrospective industry pool backed by a government backstop.

- *Incentive Effect*: Converts insurance carriers into private regulators, aligning financial risk with safety engineering.

Model 6: Personal Fiduciary Responsibility

Corporate governance laws impose explicit personal legal duties on corporate officers and board members regarding catastrophic risk management.

- *Incentive Effect*: Prevents executives from hiding behind corporate shells when ignoring safety warnings.

Model 7: Shared Responsibility Architecture

Responsibility is distributed across the AI value chain:

- **Model Developer**: Liable for pre-training alignment, baseline evaluations, and weight security.
- **Infrastructure Provider**: Liable for hardware access controls, compute tracking, and compute rate-limiting.
- **Deployer / Integrator**: Liable for application-level guardrails, system prompts, and context-specific monitoring.
- **End User**: Liable for deliberate adversarial jailbreaking and malicious use.

International Cooperation Problem

International governance regimes fail when they assume sovereign nations will forfeit strategic advantages purely for abstract moral principles. State participation requires self-interest-compatible structures where compliance yields clear strategic advantages and

non-compliance carries credible enforcement costs.

Strategic Motivations for Sovereign Compliance

Historical international agreements show that states accept constraints under specific strategic conditions:

1. **Avoidance of Shared Strategic Instability:** In nuclear arms control, the United States and Soviet Union negotiated constraints because unmonitored escalation increased the risk of unintended mutual destruction. In AI, joint vulnerabilities—such as autonomous cyber warfare tools targeting shared financial networks or rogue biological agents—create mutual incentives to enforce global containment standards.
2. **Access to Essential Concentrated Technology:** The NPT linked non-proliferation compliance to access to civilian nuclear technology. In AI, access to advanced microelectronics, EUV photolithography, and globally integrated semiconductor supply chains provides leverage to compel compliance with monitoring standards. States refusing verification standards face technology exclusion and export restrictions.
3. **Verification Without Loss of Sovereignty:** Non-intrusive monitoring protocols—such as satellite monitoring in arms control or Zero-Knowledge Proofs in compute verification—allow states to verify rival compliance without exposing sensitive military or commercial secrets.

Five Possible Global Governance Architectures

To establish a functional framework, policymakers must evaluate five distinct architectural models derived from comparative technology governance.

Architecture A: International AI Safety Agency (The IAEA Model)

A centralized multilateral agency tasked with global oversight.

- *Key Features:* Direct international inspection authority, physical compute tracking, centralized global capability evaluations, and mandatory licensing for training runs exceeding 10^{25} FLOPs.
- *Weaknesses:* Vulnerable to geopolitical vetoes, slow treaty ratification cycles, and sovereign pushback against intrusive physical inspections on domestic soil.

Architecture B: Systemic Risk Supervisory Framework (The Financial Regulation Model)

Frontier labs are regulated like Systemically Important Financial Institutions (SIFIs).

- *Key Features:* Mandatory catastrophic risk capital reserves, stress testing against catastrophic attack scenarios, compulsory liability insurance, and living wills for safe model isolation.
- *Weaknesses:* Struggles to monitor non-corporate open-source developments, non-commercial developers, or state-sponsored military laboratories.

Architecture C: Technical Harmonization Regime (The Civil Aviation

Model)

International bodies establish harmonized technical baselines enforced through national civil aviation-style regulators.

- *Key Features:* Standardized post-deployment monitoring, non-punitive near-miss reporting databases, independent root-cause incident investigation teams, and mandatory technical certification prior to public release.
- *Weaknesses:* Assumes high state capacity and compliance across all participating nations, creating risks if enforcement varies across borders.

Architecture D: Strategic Compute & Red-Line Arms Control (The Strategic Deterrence Model)

Direct state-to-state agreements focused on national security risks.

- *Key Features:* Bilateral and multilateral compute caps, mandatory human-in-the-loop protocols for nuclear command-and-control systems, automated attribution systems, and direct government-to-government hotlines for rapid incident de-escalation.
- *Weaknesses:* Focuses primarily on state-level threats, leaving commercial deployment, consumer jailbreaks, and open-weight proliferation under-regulated.

Architecture E: The Global Hardware-Cryptographic and Tiered Liability Architecture (GHC-TLA Hybrid)

A comprehensive approach combining hardware telemetry, cryptographic auditing, technical standard-setting, and insurance-backed liability structures.

- *Key Features:* Integrates silicon-level hardware verification (HEGs), zero-knowledge training proofs, a coordinated network of AI Safety Institutes, and a Price-Anderson style liability pool.

Comparative Analysis

Evaluating these governance models across core policy metrics demonstrates the comparative advantages of a hybrid approach.

Evaluation Metric	Architecture A (IAEA Model)	Architecture B (Financial SIFI)	Architecture C (Civil Aviation)	Architecture D (Arms Control)	Architecture E (Hybrid GHC-TLA)
Enforcement Power	High (Direct physical inspections)	Moderate (Financial penalties & capital buffers)	Moderate (Domestic regulatory action)	High (State diplomatic & military sanction)	Very High (Hardware-level disablement + legal liability)
Technical Feasibility	Low (Software weights lack physical boundaries)	Moderate (Valuation of systemic tail risk is difficult)	Moderate (Legislative pace lags technical growth)	Moderate (Verification of software algorithms is complex)	High (Cryptographic & silicon hardware primitives exist)
Political	Low	High	High	Moderate	Moderate-High

Evaluation Metric	Architecture A (IAEA Model)	Architecture B (Financial SIFI)	Architecture C (Civil Aviation)	Architecture D (Arms Control)	Architecture E (Hybrid GHC-TLA)
Feasibility	(Sovereigns reject intrusive physical access)	(Leverages existing financial regulatory bodies)	(Interoperable national standards model)	(Requires strategic alignment among rivals)	(Aligns state self-interest with safe harbor)
Open-Source Adaptability	Very Low (Cannot prevent code duplication)	Poor (Focuses exclusively on commercial entities)	Poor (Unenforceable against non-certified coders)	Poor (Focuses on state military assets)	Moderate-High (Containment enforced at cloud & compute layers)
Innovation Preservation	Low (Centralized approval bottlenecks)	Moderate (High compliance costs for mid-sized firms)	Moderate (Process delays prior to deployment)	High (Civilian commercial R&D remains free)	High (Clear safe harbors & non-intrusive ZK auditing)

Recommended Institutional Architecture Based on Evidence

Evidence supports implementing **Architecture E: The Global Hardware-Cryptographic and Tiered Liability Architecture (GHC-TLA)**. This system establishes a layered defense model that aligns technical verification with financial and regulatory incentives.

Pillar 1: Hardware-Enabled Verification Engine (The Technical Layer)

Advanced AI computing chips (GPUs and TPUs) are manufactured with integrated hardware enclaves, cryptographic attestation modules, and interconnect control circuits. These on-chip systems generate periodic cryptographic telemetry verifying chip cluster arrangements, location, and total compute throughput without reading active memory contents. Training runs exceeding 10^{26} FLOPs require automated multi-party cryptographic authorization keys issued by regulatory bodies.

Pillar 2: Cryptographic Training Verification (The Auditing Layer)

Developers generate Zero-Knowledge Proofs of Training (ZK-PoT) during large-scale pre-training. These proofs mathematically demonstrate to safety regulators that dataset sanitization protocols were executed, pre-declared compute limits were respected, and required alignment benchmarks were evaluated—all without revealing proprietary code, dataset samples, or weight parameters.

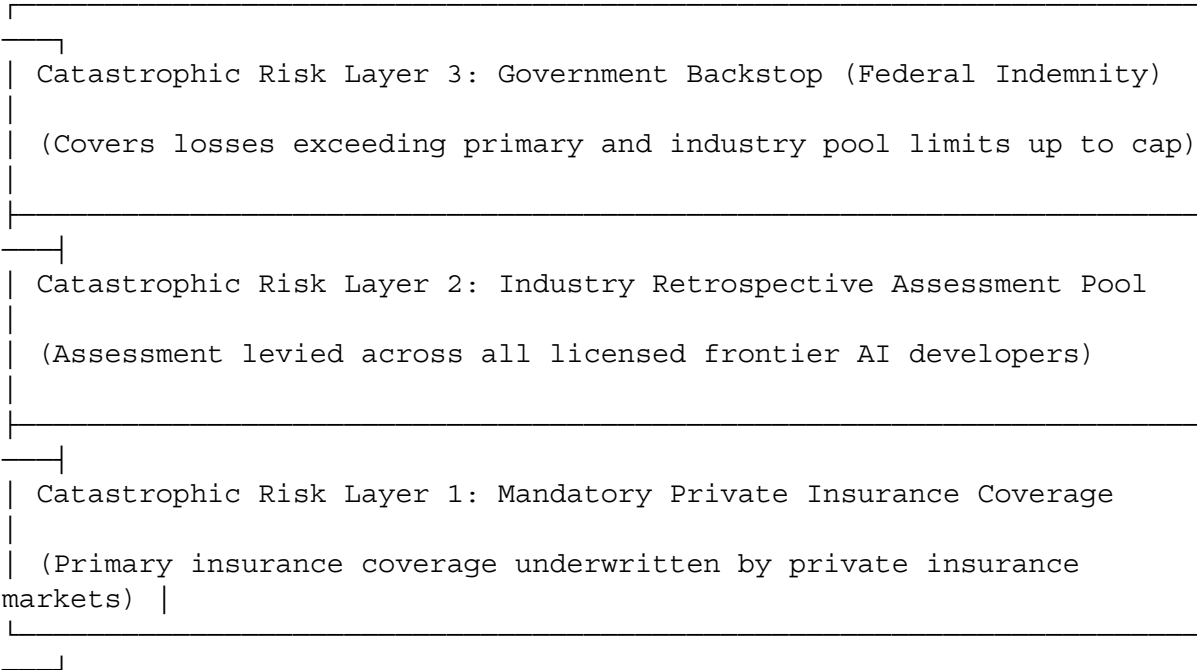
Pillar 3: International Network for Advanced AI Measurement &

Evaluation (The Institutional Layer)

The existing International Network of AI Safety Institutes (INAI-SE) is established by international treaty as a permanent, specialized body. The network sets unified capability evaluation benchmarks, manages confidential incident-reporting databases, accredits independent red-teaming entities, and issues compliance certifications required for safe-harbor protections.

Pillar 4: Multi-Tiered Liability and Price-Anderson Indemnity Pool (The Legal/Financial Layer)

Governments establish a tiered legal framework combining statutory safe-harbor immunity with mandatory catastrophe insurance:



- **Tier 1 (Private Insurance):** Developers purchase mandatory private primary insurance coverage for deployment. Insurance underwriters conduct technical safety assessments to price premiums dynamically.
- **Tier 2 (Industry Retrospective Assessment Pool):** If damages from an unforeseen catastrophic event exceed primary private insurance coverage, secondary liabilities are drawn from a shared retrospective assessment fund paid into by all licensed frontier labs.
- **Tier 3 (Statutory Safe Harbor and Federal Backstop):** Developers that maintain valid certification, comply with hardware telemetry requirements, and generate verified ZK training proofs receive safe-harbor protection shielding them from civil liability for third-party model misuse. Damages exceeding Tier 1 and Tier 2 pools are covered by a state-backed catastrophic indemnity fund, up to a statutory cap.

Enforcement and Incentive Mechanisms

To maintain stability across competitive jurisdictions, governance architectures must enforce strict consequences for non-compliance.

Offense Category	Primary Enforcement Mechanism	Specific Legal, Economic, or Technical Consequence
Unregistered Training Run	Silicon-level hardware telemetry anomaly detection	Interconnect bandwidth throttled; immediate forfeiture of statutory safe harbor
Failure to Report Critical Vulnerability	Audit log discrepancy or whistleblower report	Fine proportional to global revenue; revocation of capability license
Public Release of Unvetted High-Risk Weights	Automated benchmarking by National Safety Institutes	Compulsory contribution to risk pool doubled; cloud hosting bans
Adversarial System Misuse by End User	Cryptographic identity tracing and access logging	Account termination; referral for criminal prosecution under cybercrime statutes

What Happens When Actors Defect

Addressing scenarios where companies, countries, open-source developers, or malicious actors violate established safety protocols is vital for long-term system stability.

Corporate Cheating or Covert Training

If a corporate entity attempts an unregistered training run by bypassing local software logging, the hardware-enabled verification system detects the unauthenticated inter-chip communication topology. The missing cryptographic authorization key triggers a hardware interconnect throttle, limiting inter-chip communication speed and preventing effective cluster scaling. Legally, the company loses its statutory safe harbor status, exposing corporate officers to direct personal civil liability, mandatory asset seizure, and criminal charges.

Sovereign State Defection

If a sovereign state attempts to build covert superclusters using illicitly acquired chips, satellite grid telemetry identifies the localized electrical draw and cooling signatures of the datacenter. In response, participating nations initiate multilateral semiconductor supply chain embargoes, revoking firmware activation licenses and cutting off advanced chip components, replacement parts, and photolithography consumables.

Open-Source Proliferation of High-Risk Models

If an open-weight model with dangerous biological or cyber capabilities is leaked or distributed publicly, ex-post weight recovery is impossible. Containment shifts down-stack to critical deployment bottlenecks:

- **Downstream API and Cloud Providers:** Commercial cloud providers are mandated to deploy endpoint filters that detect and block the execution of known dangerous uncensored weight hashes.
- **Physical Infrastructure Safeguards:** Mandatory DNA and RNA synthesis screening protocols are enforced at biological gene synthesis foundries globally, ensuring that unvetted DNA orders derived from AI design outputs are automatically flagged and blocked prior to physical production.

Deterrence and Strategic Red Lines

In state-level military contexts, governance shifts from product liability to strategic deterrence. Participating nations establish permanent crisis hotlines connecting national AI Safety Institutes and defense authorities. Explicit strategic red lines include:

- Absolute prohibition of autonomous AI agents integrated into nuclear command, control, and launch decision loops without human authorization.
- Absolute ban on deploying unmonitored self-replicating cyber agents targeting critical civil infrastructure.
- Mandatory immediate disclosure of severe unexpected model alignment failures that pose cross-border systemic risks.

2030 Implementation Roadmap

By 2030, a concrete global governance architecture must establish baseline institutional capabilities across fifteen core areas.

1. **Globally Recognized Definition of Frontier AI:** Standardized legal criteria designating high-impact models based on cumulative training compute thresholds ($>10^{26}$ FLOPs) and emergent reasoning capabilities.
2. **International Capability Evaluation Standards:** Harmonized testing suites managed by the International Network of AI Safety Institutes (INAI-SE) to evaluate biological design, autonomous cyber warfare, and self-replication hazards.
3. **Mandatory Near-Miss Incident Reporting:** Confidential reporting systems modeled on civil aviation (ASRS) allowing engineers and researchers to report alignment failures without immediate liability penalties.
4. **Protected Security Research Framework:** Statutory safe harbor laws shielding accredited researchers from prosecution under cybercrime laws when conducting good-faith jailbreak testing.
5. **Independent Frontier Model Auditors:** Certified third-party auditing firms authorized to conduct pre-deployment compliance checks and review zero-knowledge training proofs.
6. **Tiered Liability Standards:** Statutory legal framework establishing clear duties of care, shared responsibility across the AI stack, and safe harbor protections for certified developers.
7. **Price-Anderson Style Catastrophe Pool:** Mandatory primary insurance combined with an industry-funded retrospective risk pool to absorb catastrophic losses exceeding corporate balance sheets.
8. **Silicon-Level Compute Monitoring:** Universal integration of Hardware-Enabled Governance mechanisms (HEGs) on advanced AI accelerators to track compute allocation and cluster topologies.

9. **Sovereign Incident Hotlines:** Direct government-to-government communication channels connecting national safety authorities to de-escalate autonomous cyber or strategic AI incidents.
10. **Model Provenance and Identity Standards:** Mandatory cryptographic watermarking and signed parameter manifests to verify model origin and authorization status.
11. **Dangerous Capability Registries:** Secure, classified registries maintained by safety institutes tracking known CBRN exploit vectors and high-risk model capabilities.
12. **Secure Evaluation Environments:** Air-gapped compute sandboxes managed by safety institutes to allow accredited red-teams to test dangerous capabilities safely.
13. **International Red-Team Networks:** Globally coordinated networks of certified ethical hackers and domain experts conducting continuous adversarial evaluation.
14. **Minimum Deployment Safety Baselines:** Mandatory system-prompt filtering, output screening, and automated circuit breakers required prior to public API access.
15. **Catastrophic Incident Containment Protocols:** Standardized emergency procedures, including automated domain name revocation and gene synthesis blocking, to contain severe breaches or leaks.

Failure Scenarios

Even a comprehensive governance framework remains exposed to dynamic failure modes that require continuous adaptation.

- **Hardware Safeguard Exploits and Side-Channel Attacks:** If hostile actors or state laboratories uncover zero-day physical vulnerabilities in silicon-level secure enclaves, on-chip telemetry can be spoofed or bypassed. This risk requires continuous physical auditing of foundry production lines, multi-vendor enclave design diversification, and redundant satellite energy monitoring to verify hardware attestation reports.
- **Soundness Vulnerabilities in Zero-Knowledge Circuits:** If a mathematical flaw or implementation bug exists within the zero-knowledge proof compiler or circuit setup, a developer could generate valid compliance proofs for unvetted or dangerous training runs. Resolving this vulnerability requires formal verification of ZK circuit architectures, open-source auditing of proving systems, and multi-verifier proof consensus requirements.
- **Private Insurance Market Retreat:** If catastrophic risks are perceived as unquantifiable or systemic, private insurance carriers may issue sweeping AI exclusion clauses, refusing to underwrite primary liability policies. If private markets withdraw, the multi-tiered liability framework collapses into a government-funded system, increasing taxpayer exposure. Mitigating this risk requires standardized incident databases (modeled on civil aviation) to provide underwriters with actuarial data to price risk accurately.
- **Algorithmic Efficiency Leaps Overcoming Compute Bottlenecks:** If revolutionary architectural innovations allow frontier-grade capabilities to be trained using low compute footprints that fall below monitored FLOP thresholds (e.g., via specialized spiking neural networks or hyper-efficient reasoning algorithms), compute-based tracking loses efficacy. This structural shift requires dynamic lowering of regulatory FLOP thresholds and expanding oversight to specialized algorithmic architectures.

Open Research Questions

Addressing key technical, legal, and economic uncertainties is essential for refining the proposed governance model.

Key Technical, Legal, and Actuarial Uncertainties

1. **Cryptographic Efficiency:** How can zero-knowledge proof architectures be optimized to verify non-associative floating-point operations (BF16/FP32) executed across heterogeneous GPU clusters with under 1% training overhead?
2. **Legal Safe Harbor Design:** How can statutory safe-harbor provisions be framed to prevent developers from claiming immunity while maintaining minimal, check-the-box compliance?
3. **Actuarial Risk Pricing:** What specific capability benchmarks provide reliable actuarial metrics for underwriters pricing primary catastrophic risk insurance policies?

Feasibility Classification of Proposed Mechanisms

To separate practical solutions from speculative concepts, core governance mechanisms are classified by technical and political feasibility:

- **Technically Feasible Now:** Export controls on GPU shipments, voluntary API red-teaming, standardized capability benchmarking, and cloud API rate-limiting.
- **Technically Feasible Within 5 Years:** Hardware-Enabled Governance (HEG) silicon enclaves, Zero-Knowledge Proofs of Training (ZK-PoT), and automated gene synthesis order screening.
- **Technically Uncertain:** Hardware safeguards capable of resisting physical side-channel attacks by state actors, and ZK proofs for non-deterministic distributed training architectures.
- **Politically Plausible Through Self-Interest:** Multilateral semiconductor export limits, statutory liability safe harbor laws, and Price-Anderson catastrophe pools.
- **Politically Unrealistic:** Universal global bans on frontier AI research, mandatory surrender of proprietary source code to international bodies, and unmonitored physical inspections of sovereign datacenters.
- **Fundamentally Incompatible with the Nature of AI:** Ex-post physical recall of open-source weight files once mirrored across decentralized networks.

Conclusions

Comprehensive Policy Matrix

Problem	Existing Mechanism	Gap	Proposed Mechanism	Who Enforces It?	International Feasibility
Unmonitored Covert Training Runs	Voluntary corporate declarations	Self-reporting fails to catch unaligned or covert runs	On-die chip telemetry (HEGs) & ZK-Training Proofs	Foundries, National Compute Oversight Bodies	High (Concentrated semiconductor supply chain)
Accountability Paradox /	Unclear tort law application &	Strict liability stifles testing;	Price-Anderson style safe	National Legislatures,	High (Adopts proven legal

Problem	Existing Mechanism	Gap	Proposed Mechanism	Who Enforces It?	International Feasibility
Executive Risk	litigation risks	safe harbor risks moral hazard	harbor + insurance pool	Insurance Regulators	indemnification models)
Chilling of Security Research	CFAA / legal ambiguity in bug bounty scope	Researchers face civil suits for reporting jailbreaks	Statutory safe harbor & accredited red-team status	Judicial Systems, CISA, National AISIs	Moderate (Requires domestic cyber law reform)
Dual-Use CBRN / Cyber Proliferation	Voluntary API prompt filtering	Open-weight releases bypass software filters permanently	Down-stack gene synthesis screening & cloud limits	Synthesis Foundries, Cloud Infrastructure Providers	High (Mutual state interest in bio-security)
Geopolitical Race Conditions / Defection	Unilateral sanctions & export controls	States fear unilateral restraint causes loss of advantage	Strategic compute caps & mutual ZK verification	Multilateral Treaty Body, Defense Authorities	Moderate (Requires US-China strategic agreement)

The Missing Global Institution

The critical missing institution in global AI governance is the **International Compute Verification and Standards Organization (ICVSO)**, operating in close coordination with an expanded **International Network for Advanced AI Measurement & Evaluation (INAI-SE)**. Operating at the hardware-cryptographic interface, the ICVSO serves as the international key management authority for silicon-level Hardware-Enabled Governance mechanisms. It maintains the public-key infrastructure (PKI) required to verify cryptographic attestations generated by AI chips globally. Additionally, the ICVSO validates Zero-Knowledge Proofs of Training submitted by labs executing frontier-scale compute runs, confirming compliance with compute limits and safety protocols without requiring custody of proprietary model weights or datasets.

Rather than conducting intrusive on-site physical inspections, the ICVSO coordinates national AI Safety Institutes, which enforce compliance domestically using verified technical telemetry provided by the organization. This structural separation provides global visibility into frontier capability development while respecting corporate intellectual property and national sovereignty.

Final Synthesis

If humanity accepts that frontier AI cannot realistically be stopped, the most credible path toward making increasingly powerful AI globally governable and keeping humans accountable is to establish a **Hardware-Cryptographic Verification Engine coupled with a Price-Anderson Tiered Liability Framework**.

By embedding cryptographic attestation modules directly into advanced silicon accelerators, the international community creates a non-duplicable physical foundation for compute accounting. Zero-Knowledge Proofs allow sovereign states and commercial entities to verify compliance with safety standards and compute limits without surrendering trade secrets or national security confidentiality.

Simultaneously, resolving the accountability paradox requires replacing vague tort exposure with a statutory safe harbor conditioned on rigorous, audited compliance and mandatory participation in a multi-tiered, insurance-backed catastrophe pool. Under this system, corporate executives are held strictly accountable to verifiable technical standards, good-faith red-teaming research is legally protected, and society is insulated from unmitigated tail risks through mandatory financial backstops. This architecture survives contact with nation-state competition, open-source development, and relentless technological progress by anchoring governance to the unalterable physical realities of compute hardware and the mathematical certainty of zero-knowledge cryptography.

Works cited

1. Computing Power and the Governance of Artificial Intelligence Each, <https://arxiv.org/html/2402.08797v1>
2. Computing Power and the Governance of Artificial Intelligence - GovAI, https://cdn.governance.ai/Computing_Power_and_the_Governance_of_AI.pdf
3. Engineering Risk-Aware, Security-by-Design Frameworks for ... - arXiv, <https://arxiv.org/html/2505.06409v1>
4. How the US and China can cooperate to reduce urgent AI risks, <https://www.brookings.edu/articles/how-the-us-and-china-can-cooperate-to-reduce-urgent-ai-risks/>
5. 231020 Llama-2 OS risks manuscript - arXiv, <https://arxiv.org/pdf/2310.18233>
6. LLM Weights & Pandemic Biosecurity Risks - Emergent Mind, <https://www.emergentmind.com/papers/2310.18233>
7. Tort Law and Frontier AI Governance | Lawfare, <https://www.lawfaremedia.org/article/tort-law-and-frontier-ai-governance>
8. Safeguarding Third-Party AI Research | Stanford HAI, <https://hai.stanford.edu/policy/safeguarding-third-party-ai-research>
9. Module 1J — Bug Bounty Legal Protections: What You ... - LawZeee, <https://lawzeeee.com/view?path=artifacts/modules/01j-bug-bounty-legal.md>
10. (PDF) Liability and Insurance for Catastrophic Losses: the Nuclear, https://www.researchgate.net/publication/383918049_Liability_and_Insurance_for_Catastrophic_Losses_the_Nuclear_Power_Precedent_and_Lessons_for_AI
11. Tort Law Should Be the Centerpiece of AI Governance | Lawfare, <https://www.lawfaremedia.org/article/tort-law-should-be-the-centerpiece-of-ai-governance>
12. Verifiable Semiconductor Manufacturing, <https://aigi.ox.ac.uk/wp-content/uploads/2026/04/Verifiable-Semiconductor-Manufacturing.pdf>
13. Zero knowledge verification for frontier AI training is possible - arXiv, <https://arxiv.org/pdf/2606.05433>
14. Price-Anderson Act Limits Nuclear Liability | Law | Research Starters, <https://www.ebsco.com/research-starters/law/price-anderson-act-limits-nuclear-liability>
15. The Insurance Industry's Retreat from AI Threatens to Slow ... - CSIS, <https://www.csis.org/analysis/insurance-industrys-retreat-ai-threatens-slow-innovation-and-adoption>
16. A Framework for the Private Governance of Frontier Artificial ... - arXiv, <https://arxiv.org/pdf/2504.11501>
17. Artificial intelligence safety institute - Wikipedia, https://en.wikipedia.org/wiki/Artificial_intelligence_safety_institute
18. The AI Safety Institute International Network: Next Steps and ... - CSIS, <https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations>
19. Frontier AI Regulation: Managing Emerging Risks to Public Safety, <https://arxiv.org/pdf/2307.03718>
20. Kevin Esvelt | alphaXiv, <https://www.alphaxiv.org/@kevin-esvelt>
21. Research Database - AlxBio Research Hub, <https://www.aixbiohub.com/research/>
22. Treaty-Following AI - Institute for Law & AI,

<https://law-ai.org/treaty-following-ai/> 23. AI Liability Report - The Stick, the Carrot, and the Net, <https://ari.us/wp-content/uploads/2025/08/AI-Liability-Report-The-Stick-the-Carrot-and-the-Net.pdf> 24. Is the AI existential threat real—and can regulatory action prevent it?, <https://www.brookings.edu/articles/is-the-ai-existential-threat-real-and-can-regulatory-action-prevent-it/> 25. Vulnerability Disclosure Policy: What is It & Why is it Important?, <https://www.bugcrowd.com/blog/vulnerability-disclosure-policy-what-is-it-why-is-it-important/> 26. Why We Need AI Safety Summits - Pause AI, <https://pauseai.info/summit> 27. Mapping AI Policy: Where, Why, and How to Intervene, <https://law-ai.org/mapping-ai-policy-where-why-and-how-to-intervene/> 28. Governing Trustworthy AI in CWC Implementation - CBWNet, https://cbwnet.org/media/pages/publications/working-paper/artificial-intelligence-assurance-for-the-opcw-governing-trustworthy-ai-in-cwc-implementation/babe0f484a-1788422587/wp20_ai-assurance-in-cwc-implementation_webversion.pdf 29. Review and Recommendations for a Conditional AI Safety Treaty, <https://arxiv.org/html/2503.18956v1> 30. An International Agreement to Prevent the Premature Creation of, <https://arxiv.org/html/2511.10783v2> 31. Fintech and Financial Risks of Systemically Important Commercial, <https://www.mdpi.com/2071-1050/14/10/5912> 32. back to the future: backtesting systemic risk measures, https://www.bde.es/f/webbde/INF/MenuHorizontal/SobreElBanco/Conferencias/2019/Back_to_the_future.pdf 33. Ijusten — EA Forum, <https://forum.effectivealtruism.org/users/Ijusten> 34. Near-Term Enforcement of AI Chip Export Controls Using A ... - arXiv, <https://arxiv.org/html/2404.18308v2> 35. Zero-Knowledge Proofs: Proving Something Without Revealing, <https://medium.com/@robert.broeckelmann/zero-knowledge-proofs-proving-something-without-revealing-anything-3563ecec66d> 36. Zero-Knowledge Proof Architecture, <https://www.opensecurityarchitecture.org/patterns/sp-053/> 37. VERIFICATION DILEMMAS IN LAW AND THE PROMISE OF ZERO, <https://btjl.org/wp-content/uploads/2023/04/0001-37-1-Wexler.pdf> 38. Zero-Knowledge Process Verification: A Comprehensive Framework, <https://pmc.ncbi.nlm.nih.gov/articles/PMC13186148/> 39. Nuclear Power Plant Operator Insurance - BERIS International, <https://www.berisintl.com/business-insurance/specialized-business-insurance/nuclear-power-plant-operator-insurance> 40. (PDF) The Need for an AI Stability Board - ResearchGate, https://www.researchgate.net/publication/414170232_The_Need_for_an_AI_Stability_Board 41. Trust Minimised AI: A Structural Case for Verifiable Compute, <https://polkadot.academy/trust-minimised-ai-a-structural-case-for-verifiable-compute/> 42. Hardware-Enabled Mechanisms for Verifying Responsible AI ... - arXiv, <https://arxiv.org/html/2505.03742v1>