AI Evaluations and National Security: Systemic Pathways For Disclosure

Technical Brief

Evals As National Security

Evals As National Security

AI Evaluations and National Security: Systemic Pathways For Disclosure

Author : Jonas Kgomo

Report Theme: Technical Brief

Formal Markets
Market Shaping & Commitments for Safe AI

Abstract

This technical brief examines the intersection of artificial intelligence (AI), national security, and cryptography amid evolving export controls . Shifting focus from chip size and compute thresholds to critical materials like germanium and indium arsenide-aluminum (InAs-Al), the brief addresses a gap in current regulatory frameworks, particularly in the context of quantum and mesoscopic materials. We propose a minerals-first approach for export controls and secure evaluation sharing using cryptographic tools like Verifiable Delay Functions (VDFs) for fairness and integrity. Traditional chip-focused export controls, based on size and compute thresholds, are outdated; we recommend **quantum-classical compliance workflows** prioritizing critical minerals. The brief explores: (1) **Algorithmic National Security**: (2) Export Controls for **Mineral-based licenses**; (3) Evaluation Vectors for compute and risk assessments; (4) **LLM Reconnaissance** for information security and open-system risks; (5) **Responsible Sharing** through cryptographic primitives(e.g., **verifiable delay functions**) for integrity. Through case studies (e.g., Japan's defense transfer model, Bernstein vs US Case) and mesoscopic technology analysis, we advocate international collaboration to harmonize governance, balance innovation with security, and mitigate AI-driven risks. This brief equips policymakers with actionable strategies for responsible AI diffusion.

Contents

1 Algorithmic National Security

1.1 History of Cryptography

The history of cryptography has evolved significantly from classical methods like shift ciphers(eg. Caesar (Caesar and Hirtius 58AD) and Vigenère) to the development of modern cryptographic protocols, such as public-key systems, RSA(Rivest-Shamir-Adleman) (Furht 2006) block ciphers like Advanced Encryption Standard(AES) (Daemen and Rijmen 2002) which hinged on number theory breakthroughs. These milestones have shaped the foundations of modern security systems, providing the basis for secure communications and data protection in the digital era. Such breakthroughs revolutionized cryptography by enabling secure key exchange over insecure channels, complementing symmetric systems like AES (Daemen and Rijmen 2002) for efficient encryption.

"*Hanc Graecis conscriptam litteris mittit, ne intercepta epistola nostra ab hostibus consilia cognoscantur.*" – Julius Caesar

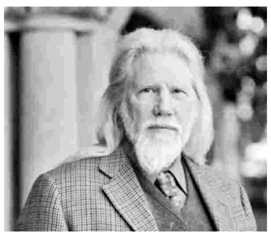
Furthermore, just as cryptography enables secure transmission of information, AI evaluation challenges—such as model transparency, security, and fairness—require the same level of attention to rigorous, systematic validation. This legacy extends to AI evaluations, where rigorous validation—akin to cryptanalysis—guards against threats like data breaches or model inversion (Gao, Shumailov, and Fawaz 2023).

1.2 Information Theory and National Strategy

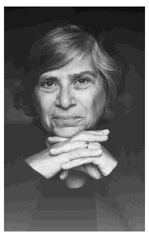
Advanced technology, like artificial intelligence and cryptography, are a dual-use technology with profound national security implications. From early expert systems in the 1980s to modern large language models (LLMs), AI's ability to process vast data and generate insights has made it a source of intelligence augmentation. However, this explosion introduces risks, such as adversaries exploiting model outputs for reconnaissance or disinformation.

1.2–I Science, The Endless Frontier

Vannevar Bush's "Science, The Endless Frontier" (1945) catalyzed federal investment in dual-use technologies, from computing to nuclear physics, shaping the modern military-industrial complex. Today, AI's dual-use nature echoes this legacy: LLMs enhance strategic analysis but risk enabling adversaries if improperly shared. Secure evaluation frameworks, using cryptographic tools like homomorphic encryption (Gentry 2009) are essential to ensure responsible diffusion, mirroring Bush's vision of science serving security.



Whitfield Diffie



Susan Landau

Vannevar Bush

Daniel Bernstein

1.3 Export of Cryptography

Cryptography's export remains a geopolitical flashpoint as nations balance security, innovation, and control in the AI era (Wassenaar Arrangement Secretariat 2023). Modern frameworks like the Wassenaar Arrangement regulate dual-use tech —encryption included—impacting AI systems reliant on secure computation (Congressional Research Service 2024). In the 1990s, U.S. controls classified encryption software as a "munition" under the Arms Export Control Act, restricting its dissemination. For AI evaluations, export rules shape access to cryptographic tools (e.g., homomorphic encryption), critical for assessing frontier models against risks like data breaches or adversarial exploits (National Institute of Standards and Technology 2025).

1.4 Daniel Bernstein vs US

The legal battle of **Daniel Bernstein v. United States** (1995–2003) stands as a landmark in cryptography history, with enduring implications for regulating artificial intelligence (AI) evaluations. Bernstein, a graduate student at UC Berkeley, challenged U.S. export controls on cryptographic software, arguing they violated his First Amendment rights. Supported by the Electronic Frontier Foundation (EFF), his case reshaped the legal status of **code as speech** and **loosened restrictive regulations** (Foundation 2011) This historical precedent offers critical insights into the intersection of technology, free expression, and national security—issues that resonate in today's debates over AI evaluation frameworks. Before Bernstein's case, cryptography was tightly regulated under U.S. law. During the Cold War, encryption was classified as a "munition" on the United States Munitions List, subject to the Arms Export Control Act (AECA) and International Traffic in Arms Regulations (ITAR) (Contributors 2003). Exporting cryptographic software required State Department approval, a process that stifled academic and commercial innovation (Editors 2014).

1.4–I The Bernstein Case: Code as Speech

In 1995, Bernstein sued the U.S. Department of Justice, asserting that ITAR's licensing requirements constituted an unconstitutional prior restraint on speech (Foundation 2011) His argument hinged on the expressive nature of source code, a medium cryptographers use to convey scientific ideas (Expression 2015) The EFF bolstered this claim, drawing parallels to mathematical equations protected under the First Amendment (Foundation 2015). By 2003, loosened Commerce Department rules under the Export Administration Regulations (EAR) rendered Bernstein's challenge moot, but the precedent endured (Contributors 2003). The case exposed the limits of blanket regulation. The export of cryptography argued proliferation of encrypted algorithms threatened national security. This tension mirrors current AI debates, where evaluation restrictions aim to curb misuse but risk hindering progress (Ilia Shumailov, Daniel Ramage, Sarah Meiklejohn, Kairouz, et al. 2025)

1.4–II Post-Bernstein Impact on Cryptography

Bernstein's victory catalyzed a relaxation of cryptographic export controls in the late 1990s, enabling secure e-commerce and global collaboration (Contributors 2003). However, the government's pivot to subtler controls—like key escrow proposals—highlighted ongoing regulatory challenges foreshadowing AI's regulatory landscape and in modern cryptography research, such as post-quantum algorithms. Yet, security concerns persist, as seen in studies of inference attacks on "safe" AI outputs (Glukhov et al. 2024a) suggesting parallels to cryptography's dual-use nature.

1.4–III Relevance to AI Evaluations

The Bernstein case offers a framework for regulating AI evaluations, where models like large language models (LLMs) face scrutiny for their potential to reveal Sensitive information or enable attacks (Glukhov et al. 2024a). Just as source code was deemed speech, AI evaluation methodologies—often encoded in software—deserve expressive protection (Appeals 1999; Foundation 2015)

1.5 Remark: Contemporary AI Regulation Challenges

The Bernstein v. United States case illustrates the tension between innovation and security, a challenge mirrored in AI regulation. Black-box attack attribution studies reveal how adversaries exploit model queries (Gao, Shumailov, and Fawaz 2023) , akin to cryptographic vulnerabilities Bernstein's **Snuffle** might have exposed (Foundation 2011), **Snuffle simply used a one-way hash function into a symmetric (private-key) encryption system.** This regulatory environment reflected national security concerns about encryption's potential misuse by adversaries (Glukhov et al. 2023).. Proposals for trusted model environments echo cryptographic key escrow debates, balancing privacy and oversight (Ilia Shumailov, Daniel Ramage, Sarah Meiklejohn, Kairouz, et al. 2025). Yet, as Bernstein's case proved, **overly restrictive policies can backfire, driving innovation underground or overseas**

2 Export Controls

2.1 Categorizing Legal Frameworks for Licenses

Legal frameworks governing technology, trade, and cybersecurity are diverse and interconnected, reflecting the need for regulation in an increasingly digital world. **Defence-related** laws such as the **Wassenaar Arrangement** (27), Strategic Goods Act (8), and Defence Act (4) control the export of sensitive technologies, while cybersecurity and communications policies like the **Computer Misuse Act** (2), **Telecommunications Act** (8), and **EU Regulation** (7) establish guidelines for digital infrastructure security. Legal compliance mechanisms, including Decrees (6), Resolutions (2), and the **Computer Crime Act** (5), create enforcement frameworks for cybercrime. **Trade and commerce** laws such as Dual-Use (4), Regulations (6), and Orders (3) govern the movement of restricted technologies, ensuring regulatory oversight. Meanwhile, data governance policies like the **Data Protection Act** (3), **Digital Signature Act** (3), and Trade Regulations enforce privacy and transaction security, while cryptography and international laws, including US Law on Encryption, Cryptography Act, and Computer Crimes Act (3), seek to balance national security with digital rights. These regulatory clusters highlight the global effort to manage the risks and opportunities presented by emerging technologies, shaping the landscape of digital governance and compliance.

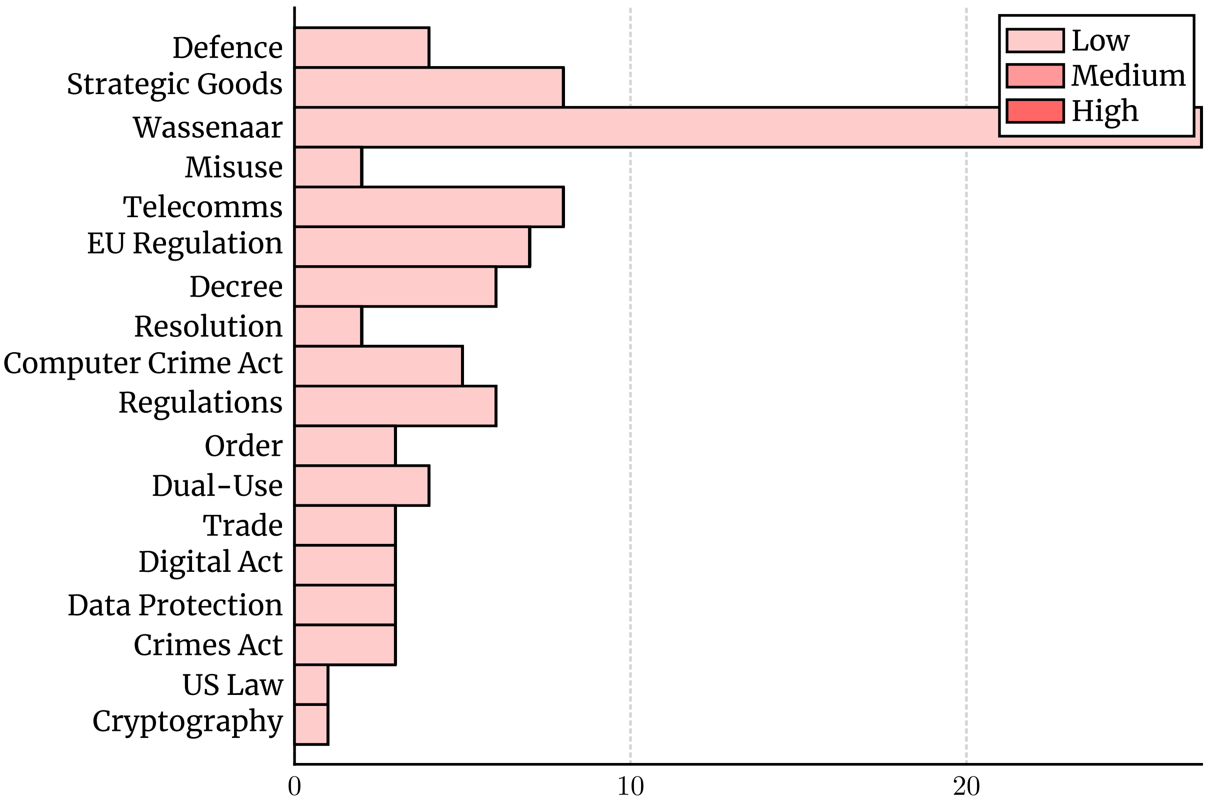


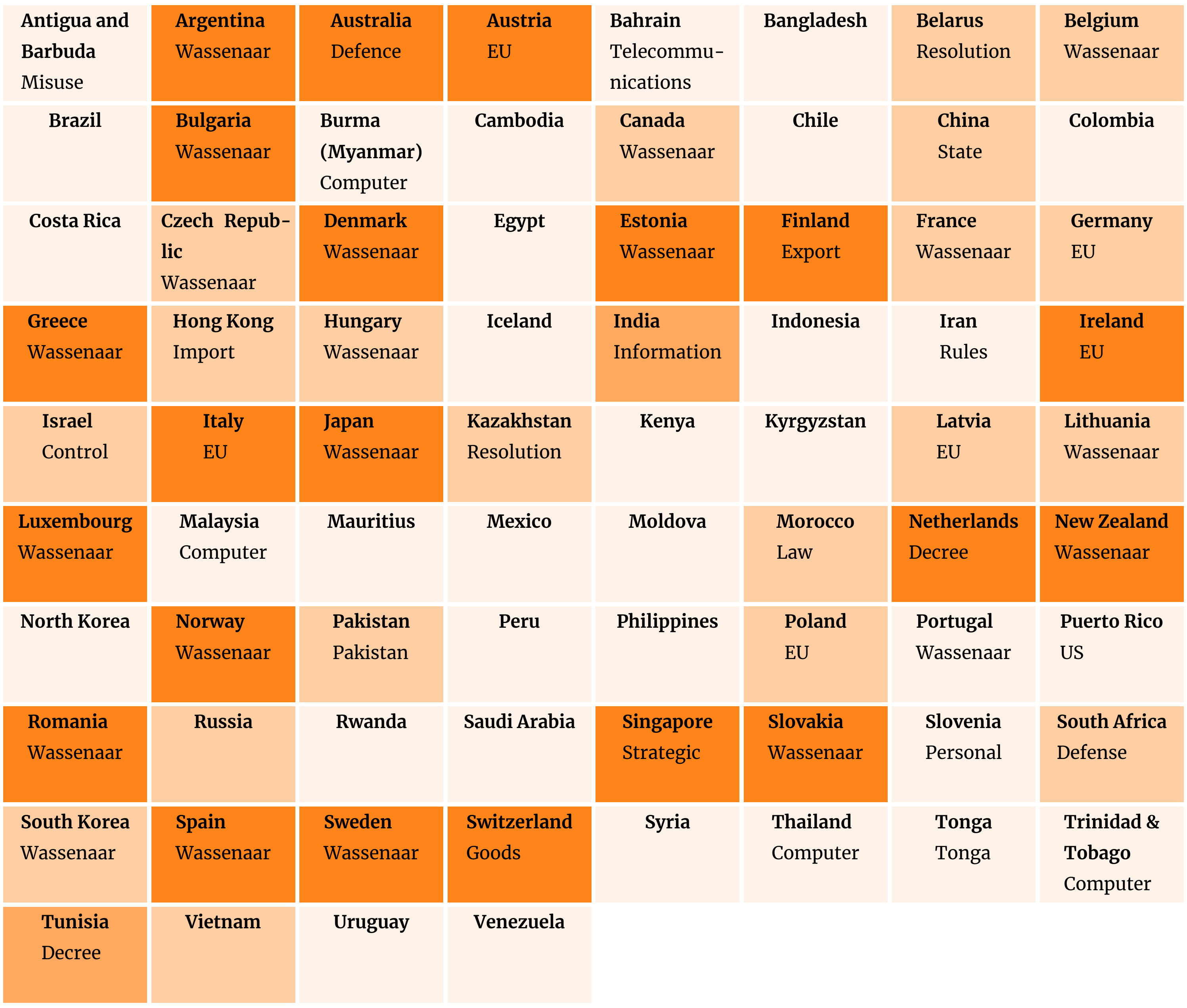
Figure 1: Bar chart depicting various regulatory categories and their associated frequencies

2.2 Country Export–Import Licenses

Comparing U.S., Chinese, and Wassenaar mineral export policies reveals gaps in global coordination, necessitating harmonized standards for AI security. National regulations governing export and import licenses for dual-use technologies, vary widely, reflecting divergent strategic priorities and security imperatives. From the United States' Export Administration Regulations (EAR) to Singapore's Strategic Goods Control Act, these frameworks shape the global flow of materials essential for AI innovation. This diversity underscores the need for unified, minerals-first export controls to secure AI supply chains and mitigate national security risks

2.3 Case Study: Wassenaar Arrangement

The **Wassenaar Arrangement**, established in 1996 by 42 nations, governs export controls on dual-use goods—items with civilian and military applications, including cryptography and advanced semiconductors (Wassenaar Arrangement Secretariat 2023). Classifying encryption as a munition under its Control Lists (e.g., Category 5, Part 2), it restricts software and hardware vital for AI (Bureau of Industry and Security 2025). Updated annually, it aims to prevent destabilizing accumulations of tech while promoting transparency among members (e.g., U.S., EU, Japan). For AI evaluations, Wassenaar shapes access to cryptographic tools and compute resources, complicating global standards for security and innovation (Congressional Research Service 2024)



Antigua and Barbuda Misuse	**Argentina** Wassenaar	**Australia** Defence	**Austria** EU	**Bahrain** Telecommunications	**Bangladesh**	**Belarus** Resolution	**Belgium** Wassenaar
Brazil	**Bulgaria** Wassenaar	**Burma (Myanmar)** Computer	**Cambodia**	**Canada** Wassenaar	**Chile**	**China** State	**Colombia**
Costa Rica	**Czech Republic** Wassenaar	**Denmark** Wassenaar	**Egypt**	**Estonia** Wassenaar	**Finland** Export	**France** Wassenaar	**Germany** EU
Greece Wassenaar	**Hong Kong** Import	**Hungary** Wassenaar	**Iceland**	**India** Information	**Indonesia**	**Iran** Rules	**Ireland** EU
Israel Control	**Italy** EU	**Japan** Wassenaar	**Kazakhstan** Resolution	**Kenya**	**Kyrgyzstan**	**Latvia** EU	**Lithuania** Wassenaar
Luxembourg Wassenaar	**Malaysia** Computer	**Mauritius**	**Mexico**	**Moldova**	**Morocco** Law	**Netherlands** Decree	**New Zealand** Wassenaar
North Korea	**Norway** Wassenaar	**Pakistan** Pakistan	**Peru**	**Philippines**	**Poland** EU	**Portugal** Wassenaar	**Puerto Rico** US
Romania Wassenaar	**Russia**	**Rwanda**	**Saudi Arabia**	**Singapore** Strategic	**Slovakia** Wassenaar	**Slovenia** Personal	**South Africa** Defense
South Korea Wassenaar	**Spain** Wassenaar	**Sweden** Wassenaar	**Switzerland** Goods	**Syria**	**Thailand** Computer	**Tonga** Tonga	**Trinidad & Tobago** Computer
Tunisia Decree	**Vietnam**	**Uruguay**	**Venezuela**				

Figure 2: Table showing the export-import licenses and laws in various countries. ***Graph Interpretation:*** *In the resulting visualization, each country is represented by a cell colored according to its trade restrictions:*

□ Indicates countries with no import or export restrictions.

🟧 Indicates countries with both import and export restrictions.

🟧 Indicates countries with only import restrictions.

🟧 Indicates countries with only export restrictions.

Important Note: This table summarizes publicly available information regarding cryptography regulations. Due to the complex and evolving nature of these regulations, it is essential to consult official sources and legal counsel for definitive guidance.

3 Reconnaissance with Large Language Models (LLMs)

3.0–I Intelligence Augmentation and Memory Expansion

In “As We May Think”, Vannevar Bush describes the **Memex** (Memory Expansion) as a device that allows individuals to store vast amounts of information and retrieve it with speed and flexibility. Bush defined the Memex as a device in which an ***individual stores information*** and which is mechanized so that it may be consulted with ***exceeding speed and flexibility***. This concept foreshadows modern AI’s Retrieval-Augmented Generation (RAG), where LLMs dynamically retrieve external data to enhance responses (Gao, Shumailov, and Fawaz 2023) . One foundational concept that emerged in the development of computing systems is key-value (Handy 1993) caching.

3.0–II Transformer Inference & Retrieval Augmentation

In particular, the concept of retrieval in Bush’s intelligence augmentation (Buckland 1992) resonates strongly with the mechanisms of modern AI, such as Large Language Models (LLMs). These models function as sophisticated retrieval systems, utilizing technologies like KV-cache(Handy 1993) to optimize memory management and inference efficiency. In the transformer architecture, **KV**-cache allows precomputed K and V to be reused across different time steps, reducing redundant computation during inference. Attention $= \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{d_k}}\right)V$. This formula describes Scaled Dot-Product Attention,which forms the foundation of KV-caching in transformers.

$$\text{Attention}\,(\underbrace{Q,K}_{\text{The Query-Key pair}},\overset{\text{The Value matrix weighed from attention mechanism}}{V}) = \text{softmax}\left(\frac{\underset{\text{Query and Key matrices for attention weights}}{\text{QK}^T}}{\sqrt{d_k}}\right)V \tag{1}$$

The concept of associative knowledge retrieval directly foreshadows modern retrieval-augmented generation (RAG) in AI, where models dynamically pull and integrate external information to enhance factual accuracy. At the core of large language models (LLMs), the attention mechanism—specifically the QK^T term in self-attention—determines which stored knowledge is most relevant at any given time. This mechanism is tightly coupled with KV-caching, which retains past key-value pairs to optimize inference efficiency. However, this caching structure introduces vulnerabilities that adversaries can exploit for model stealing and information leakage.

3.1 LLMs Lower Barriers to Reconnaissance

Large Language Models (LLMs) drastically lower barriers to sophisticated intelligence-gathering by automating and scaling tasks that once demanded extensive human expertise. This makes reconnaissance faster, easier, and more accessible, posing significant national security risks. Clio ((Tamkin et al. 2024)) leverages AI to analyze usage patterns from millions of conversations, offering scalable insights into real-world AI use—a powerful signal intelligence and reconnaissance tool. While designed to enhance safety, Clio can enable high-level clustering reconnaissance ((Glukhov et al. 2024a)) through privacy loss and unintended information leakage.

HUMINT	SIGINT	Cyber Security Reconnaissance
LLMs enhance HUMINT through behavioral profiling, sentiment analysis, and automated social engineering, reducing reliance on human operatives while improving manipulation and threat detection ((Glukhov et al. 2024a): 9; (Gao, Shumailov, and Fawaz 2023); (Glukhov et al. 2023)).	LLMs like Claude enable SIGINT by automating decryption, pattern inference, and real-time translation of intercepted communications, allowing non-state actors to conduct large-scale operations with minimal resources ((Glukhov et al. 2023); (Gao, Shumailov, and Fawaz 2023)).	LLMs automate vulnerability discovery, generate attack blueprints, and scale phishing campaigns, lowering barriers for novice hackers and amplifying cyber threat impact (; (Ilia Shumailov, Daniel Ramage, Sarah Meiklejohn, Kairouz, et al. 2025);).

Table 2– LLMs advance surveillance with real-time threat detection, object recognition, and predictive analytics for preemptive monitoring ((Gao, Shumailov, and Fawaz 2023)).

3.2 Case Study: LLM-Enabled Reconnaissance Demonstration

LLMs’ reconnaissance capabilities underscore the need for global AI governance. Multilateral frameworks, integrating cryptography and standardized safety metrics, ensure robust AI evaluations. Secure evaluation sharing, enabled by VDFs, mitigates risks like disinformation while fostering trust. International collaboration is critical to align export controls and evaluation protocols, supporting the minerals-first approach for AI hardware regulation.

The demonstration reveals several risks:

- **Scalability**: The LLM generates the email in seconds, producing dozens of variations tailored to different targets, a task that would require hours for a human operative (Glukhov et al. 2024b) .
- **Accessibility**: Minimal technical expertise is needed, as the LLM interprets natural language prompts, enabling non-state actors to conduct sophisticated exploits (Ilia Shumailov, Daniel Ramage, Sarah Meiklejohn, and others 2025).
- **Impact**: If deployed, such attacks could compromise sensitive systems

4 Sharing Evaluations For Responsible Diffusion

The rapid advancement of AI systems demands a collaborative approach to ensure safety, reliability, and trust. **Sharing evaluations of AI models** is not just a best practice—it is a necessity for responsible AI development. By openly sharing performance metrics, vulnerabilities, and insights, the AI community can collectively enhance model quality, accelerate progress, and mitigate risks. This collaborative effort fosters **transparency**, enabling policymakers, researchers, and developers to make informed decisions. It also builds **trust** among stakeholders, ensuring that AI systems are deployed responsibly and ethically. Below, we outline the key benefits of sharing AI model evaluations and their impact on the broader AI ecosystem.

Benefit	**Description**	**Policy Implications**
Enhancing Model Performance	External validation can identify issues and suggest improvements, leading to more robust and accurate models.	Fund multilateral evaluation platforms to standardize benchmarks.
Reducing Redundancy	Open evaluations create a common knowledge base, minimizing duplicated efforts and fostering cumulative advancements.	Establish global repositories for evaluation data under Wassenaar.
Promoting Transparency and Trust	Publicly shared evaluations provide empirical data, aiding policymakers in crafting informed regulations and building stakeholder confidence.	Mandate VDF-based protocols for transparent evaluation sharing.
Facilitating Collaborative Risk Mitigation	Shared insights into model performance enable collective efforts to identify and address vulnerabilities, ensuring safer AI deployments.	Create task forces to align safety metrics across nations.

Table 3– Key Benefits of Sharing AI Model Evaluations

In an era where AI systems are increasingly integrated into critical domains—from healthcare to national security—the stakes for reliability and safety have never been higher. Sharing evaluations ensures that **lessons learned** from one system can benefit the entire community, reducing the risk of catastrophic failures and promoting ethical AI deployment. By embracing a culture of openness and collaboration, we can build AI systems that are not only more capable but also more **aligned with societal values**. This is the foundation of responsible AI development—a foundation built on shared knowledge, collective progress, and unwavering commitment to safety.

4.1 Surprisal from Information

Sharing AI evaluations fosters collaboration, enhances model safety, and informs policy, but unexpected model capabilities—termed "surprisal"—pose risks. Claude Shannon's introduced entropy in his work on A ***Mathematical Theory of Communication*** (Shannon 1948), and quantifies surprisal as the shock of new information, such as a model's unforeseen proficiency in chemical synthesis (Urban & Russell, 2024). For policymakers, high surprisal signals models requiring stricter export controls, as their capabilities could enable adversaries. His cryptography work underpinned secure systems —from advanced elliptic curve cryptography (Daemen and Rijmen 2001) to homomorphic encryption (Gentry 2009)— crucial for AI eval integrity. For AI, **surprisal** quantifies the shock of new model insights:

- if party **A** learns performance(ΔModel **B**), entropy measures the gap from prior beliefs, informing security and export policies (Cover and Thomas 2006), where entropy is $H = -\sum P(x) \log_2 P(x)$

Evaluation Surpisal can tell us about how AI systems might behave and we can better estimate their capabilities in real world or simulated contexts.

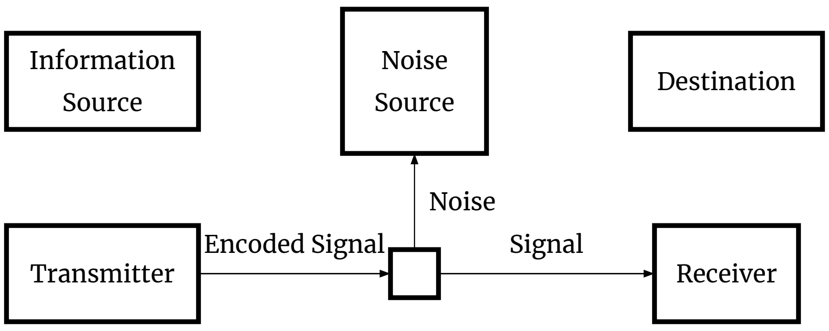


Figure 3: Shannon's Model of Information Channel & Noise

4.2 Surprisal from Evaluations

AI evaluations often reveal unexpected leaps in model performance or risks like CBRN enablement—that reshape strategic priors (Urban and Russell 2024). Measuring this surprisal quantifies how new data (e.g., perf(ΔModel B)) upends assumptions, critical for secure sharing under export controls (National Institute of Standards and Technology 2025). We formalize that shock as divergence, guiding responsible disclosure in high-stakes AI governance.

Theorem 4.2.1 (Information Surprise in Large Language Models): Let M be a model evaluated at time t with performance Perf_t (e.g., accuracy, capability). Define **surprisal** S_t as the Kullback-Leibler divergence between prior and updated output distributions:

$$S_t = D_{\{\mathrm{KL}\}}(P_t \mid P_{t-1}) = \sum P_{t(x)} \log\left(\frac{P_{t(x)}}{P_{t-1}}(x)\right)$$

where P_t and P_{t-1} are probability distributions over M's capabilities at times t and $t-1$.

1 **Prior Belief**: Model $P(p_A)$ from I_A (e.g., baseline eval data).
2 **New Evidence**: Observe p_B, likelihood $P(p_B \mid p_A)$ reflects capability shift (e.g., robustness jump).
3 **Surprisal Update**: Compute $S_{t+1} = D_{\mathrm{KL}}(P(p_A \mid p_B) \mid P(p_A))$, high if p_B reveals unexpected perf (e.g., CBRN risk (Urban and Russell 2024)).

High S_t signals asymmetry, necessitating cryptographic safeguards (e.g., (National Institute of Standards and Technology 2025)) for secure eval sharing under export controls.

4.3 Remark

By viewing an LLM as an agent navigating a sequence of conversational states, we can draw an analogy to RL. Just as RL agents avoid drastic jumps in policy updates to maintain stability, LLMs should not generate outputs that are too unexpected unless the situation demands it. KL divergence between expected and actual outputs of an LLM can quantify how much the model is "surprised" by a given prompt or situation. Monitoring KL divergence in LLMs allows for controlling the degree of surprise in the model's outputs. This control is analogous to the policy constraints in RL that prevent erratic behavior. By managing KL divergence, we can ensure that the LLM adapts appropriately to new information without producing outputs that are too unexpected, thereby maintaining coherence and reliability in its responses.

5 Mutually Beneficial Evaluations Sharing

Evaluations in AI are specialized, performance-driven assessments designed to uncover model weaknesses through techniques like red-teaming and benchmarking. Foundational work, such as Reinforcement Learning from Human Feedback (RLHF)(Christiano et al. 2017; Ouyang et al. 2022), highlights their role in refining models. A key example is METR's exploration of large language models' self-replication potential (Kinniment and others 2023). Yet, challenges persist: evaluations are context-dependent and often rely on "norm-referenced" measures , complicating universal standards. This chapter examines how shared evaluations can address these issues, fostering collaboration and insight (Burden 2024).

5.1 Case Study: Japan's Defense Transfer Model

Japan's 2014 "**Three Principles**" (revamped 2023) flipped its arms export ban into a strategic playbook (Ministry of Foreign Affairs of Japan 2024). Including radar handoffs to the Philippines, chip tech swaps with the U.S.—all vetted via tight controls (Wassenaar (Wassenaar Arrangement Secretariat 2023)) and a 2018 roadmap (Cabinet Secretariat of Japan 2018). It's a trust machine: allies get cutting-edge gear while Japan guards CBRN risks and export redlines. This dense, partner-driven model thrives on mutual gain.

5.2 Applying Japan's Model to AI Evaluations

AI chips rely on rare earths and critical minerals, constrained by export laws and supply chain conflicts (European Commission 2020). Japan's model—developed through years of technological trade—suggests evaluation sharing that tracks the usage of germanium or cobalt. Party A discloses Model B's performance, tagged with the computational mineral cost, aligning allies while avoiding adversaries.

$$\begin{tikzcd} \text{Japan} \arrow[r, "f"] & \text{Partners} \\ \text{Defense / Minerals} \arrow[ur, hook, dashed, "\tilde{f}"] & \end{tikzcd}$$

$$\begin{tikzcd} \text{A} \arrow[r, "f"] & \text{C} \\ \text{Eval / Perf} \arrow[ur, hook, dashed, "\tilde{f}"] & \end{tikzcd}$$

$$\begin{tikzcd} \text{A} \arrow[r, "f"] & \text{Allies} \\ \text{Minerals / Hoard} \arrow[ur, hook, dashed, "\tilde{f}"] & \end{tikzcd}$$

Figure 4: From Japan's tech pacts to minerals-aware eval flows

5.3 Remark: Japan's Defense Transfer Model

Japan's defense transfer model offers a pathway for AI evaluation sharing, culminating in a minerals-aware approach. Japan's established trust model—shaped by defense exchanges—illuminates a method for AI evaluation sharing that is controlled, resource-conscious, and reciprocal. While challenges persist—such as complex supply chains and variable metrics—this approach represents a pragmatic effort toward collaboration in a technologically driven and tense global environment.

5.4 Verifiable Delayed Evaluations (VDEs)

AI evaluations (AI Evals) are becoming essential tools for managing risks and building trust in high-stakes interactions, particularly in scenarios involving adversarial conditions. One such cryptographic primitive, the **Verifiable Delay Function (VDF)**,(Fisch, Pass, and Shelat 2019) also referred to as **Proof of Time**, ensures that a sequential function has been executed a specific number of times. In this context, consider a large language model (LLM) with capabilities to compute a function $f_{c(X)}$, where X represents an input sequence. The challenge is to prove that a function has been computed over a series of inputs, say $t_1, t_2, ..., t_N$, where each token stream is processed by two entities: the prover L_P and the adversary L_A. They are ideal for AI evaluations, preventing issues like benchmark hacking, chain-of-thought cheating, and reward hacking. The VDF guarantees that the computation was done sequentially, without shortcuts, and that the result is verifiable without needing to repeat the entire process.

VDF Definition – A VDF is a cryptographic primitive that leverages time and storage resources to verify and proof claims, it uses the following three algorithms:

```
VDF Definition
Setup :
    Generates public parameters $p$ based on the security parameter $\lambda$ and time-bound T.
Eval:
    Computes the output y and proof $\pi$ for input $x$ using the public parameters p.
Verify:
    Confirms the correctness of $y$ by verifying the proof $\pi$ with the public parameters p.
```

We want to embed VDFs in the benchmark process that requires a language model to solve a reasoning problem, ensuring the model performs the computation rather than exploiting shortcuts – we call this Verifiable Delayed Evaluations. VDFs are timelords for AI evaluations.

5.5 Proofs of Replication (PoRs)

Proofs of Replication (PoRs) (Lerner 2014), built on VDFs, prove that an AI model was trained on specific data without revealing it. For instance, a provider might use PoRs to demonstrate training on a designated dataset, enhancing evaluation trust. The prover dedicates storage to encode data replicas, which the evaluator verifies without accessing the data itself.

5.5–I Proof Sketch

A model provider proves training integrity:

PoR Sketch

```
Setup:
    Generate public parameters and encode training data into replicas.
Evaluation:
    Compute a proof showing data was used, leveraging sequentiality.
Verification:
    Evaluator confirms proof without seeing the data.
```

5.6 Commit-and-Reveal Protocols

Commit-and-reveal protocols (Wikipedia Contributors 2023) allow parties to commit to data (e.g., model weights) before revealing it, preventing manipulation. In AI evaluations, a provider commits to model weights via a cryptographic hash before testing. After evaluation, weights are revealed and checked against the commitment, ensuring no alterations occurred.

5.7 Remark

These protocols provide auditability and trust in AI development, especially in collaborative or competitive environments. While Proofs of Retraining (PoR) ensure that a model was trained on authentic data, commit-and-reveal schemes protect against tampering during evaluation. Together, they create a foundation for verifiable and secure AI system claims without exposing sensitive data or proprietary models. They may require additional verification for real-world evaluations sharing mechanism.

6 Mesoscopic Subterfuge

6.1 Mesoscopic Regime and Dual-Use Frontiers

The **mesoscopic regime** (10-100 nm) bridges quantum and classical physics, powering AI hardware via materials like InAs-Al (Aghaee 2023) hybrids that host topological states (e.g., Majorana fermions) for quantum chips (Lutchyn et al. 2018). EU workshops (2019-2020) flag these—alongside cloaking devices, metamaterials, nanomaterials, and high-entropy alloys (HEA)—as dual-use tech under export scrutiny (e.g., Wassenaar (Wassenaar Arrangement Secretariat 2023)), given military potential like ultracentrifuge rotors (European Commission 2020). Yet their limits (e.g., HEA's slow progress) challenge AI's compute race, where innovation wrestles with security and evaluation gaps in a fractured global regime.

6.2 Subterfuge of Emerging Mesoscopic Technologies

Emerging mesoscopic tech promises breakthroughs but stumbles on practical hurdles:

- **Cloaking Devices**: Adaptive camouflage and metamaterials (negative refraction) mask objects, yet scalability lags—nanomaterial coatings (e.g., Vantablack) absorb light but falter in dynamic environments (European Commission 2020).
- **Metamaterials**: Tailored properties (e.g., light bending) aid AI sensors, but fabrication complexity and cost hinder mass adoption (Smith and Pendry 2022).
- **Nanomaterials**: Carbon nanotubes and graphene boost chip strength and conductivity, yet inconsistent synthesis (e.g., quantum dot variability) stalls reliability for AI compute (Zhang and Novoselov 2023).
- **High-Entropy Alloys (HEA)**: Multi-element blends hint at superconductor potential, but fundamental research dominates—rapid deployment remains elusive (European Commission 2020).

These gaps—tied to dual-use export controls—limit AI eval precision (e.g., perf(ΔModel B)) and underscore the need for coordinated standards.

6.3 Transition in Chip Regulations

The transition from classical to quantum semiconductor technologies represents a fundamental challenge to existing regulatory frameworks. Traditional approaches, anchored in dimensional metrics like the 5nm node, reflect an outdated paradigm that fails to capture the multidimensional complexity of quantum-enabled devices. As semiconductor technology evolves beyond classical physics constraints, regulatory frameworks must undergo a parallel transformation. This evolution requires a shift from size-based benchmarks to performance-oriented metrics that encompass quantum phenomena such as coherence times, entanglement fidelity, and topological protection. The regulatory challenge lies not merely in updating technical specifications, but in developing adaptive frameworks that can accommodate the convergence of classical and quantum computing paradigms. Such frameworks must balance innovation enablement with security considerations, particularly as quantum technologies introduce novel vulnerabilities and capabilities that transcend classical security models.

6.4 Chips within the 5nm-20nm Regime

As semiconductor technology progresses beyond the 5nm node, traditional scaling laws that once governed device miniaturization increasingly encounter limitations. At this size, transistor behavior is heavily influenced by the physical constraints of material properties, and classical semiconductor models no longer fully explain device performance. As transistors continue to shrink, phenomena such as quantum tunneling and electron-electron interactions become more pronounced, complicating the stability and reliability of these ultra-scaled devices. These physical effects necessitate innovative approaches to design, manufacturing, and regulatory oversight to maintain device performance and security. This convergence of classical and quantum behaviors creates a unique design space where conventional semiconductor models prove insufficient, necessitating new theoretical frameworks that bridge classical and quantum domains (Kane & Mele, 2005).

6.5 Emerging Post-Classical Architectures

The post-5nm era introduces novel architectural paradigms that exploit rather than avoid quantum mechanical effects. Topological transistors, spintronic devices, and quantum-dot architectures represent a fundamental departure from traditional **CMOS** scaling, leveraging quantum phenomena for enhanced functionality. These innovations operate at quantum critical points (QCPs)—phase transitions driven by quantum fluctuations—where material properties undergo dramatic transformations that can be harnessed for information processing. The regulatory challenge lies in fostering innovation while ensuring security in an increasingly traditional technological landscape and non-classical architectures.

6.6 Solid State Memory Breakthoughs (China)

The PoX flash memory, developed by Fudan University, revolutionizes semiconductor chips with its 400-picosecond programming speed, leveraging a 2D graphene-channel design at the nanometer scale to push non-volatile memory performance to unprecedented levels. (Xiang 2025). The development of the **PoX** (Phase-change Oxide) operating at a 400-picosecond switching speed (25×10^9 operations per second), redefines storage speed limits and supports the high computational demands of AI models, while maintaining compatibility with CMOS fabrication while achieving unparalleled speed and efficiency far beyond the limitations of conventional silicon-based transistor scaling (China Daily 2025).

7 Mesoscopic Export Laws

The convergence of classical and quantum domains at mesoscopic scales represents not just a technical challenge but a fundamental reimagining of semiconductor technology. Success in this new era requires integrated approaches that recognize the inseparability of quantum effects from classical operation at advanced nodes (Roadmap to Fault Tolerant Quantum Computation Using Topological Qubit Arrays 2025)

Figure 5: The Microsoft Majorana 1 is a topological quantum computer that represents a groundbreaking advancement in quantum computing. These qubits encode information in the parity of electrons split across nanowires, making them inherently resistant to noise and errors

7.1 Chip Exports: Classical & Topological Materials

Majorana-based qubits are more robust against decoherence, making them a promising candidate for scalable and reliable quantum computation. Leveraging Majorana zero modes—quasiparticles that exhibit non-Abelian statistics—this system promises inherently fault-tolerant quantum operations.

Microsoft's approach involves creating a new state of matter called a **topological superconductor** , which is neither a solid, liquid, nor gas. This material is fabricated atom by atom using a stack of **indium arsenide and aluminum** (Luo et al. 2020). The **Majorana 1** chip is designed to scale to **one million qubits** on a single chip, which is small enough to fit in the palm of a hand (Lee et al. 2023).

Table 4 – Classical vs. Topological Materials for Chips

Parameter	Classical: Silicon	Classical: Germanium	Topological: InAs-Al
Mobility (cm^2/Vs)	1400	3800	>10,000
Superconducting Gap	N/A	N/A	0.2-0.3 meV
Critical Temp (T_c)	N/A	N/A	1.2 K
Spin-Orbit (meV·nm)	Negligible	Low	10-20
Scale	>10 nm	>10 nm	10-100 nm
Dimensionality	3D	3D	1D (nanowire)
Fabrication	Lithography	Lithography	MBE

Table 4 Compares regular semiconductors (e.g., silicon) and mesoscopic Majoranas (hybrid structures).

7.2 Explanation of Changes

Electron Mobility Correction: The original table stated electron mobility for regular semiconductors is $\sim 1400\frac{\mathrm{cm}^2}{\mathrm{Vs}}$ at room temperature. Research confirmed that Al has a bulk superconducting gap of $\sim 0.17 - 0.18$ meV at zero temperature, and in hybrids, the induced gap can be $0.1 - 0.2$ meV (Cole, Das Sarma, and Stanescu 2015)

Mesoscopic Majoranas: These are hybrid structures using **InAs** and Al, designed for studying Majorana fermions, with high electron mobility and strong spin-orbit coupling, making them suitable for advanced quantum research.

7.3 Consideration for Critical Materials

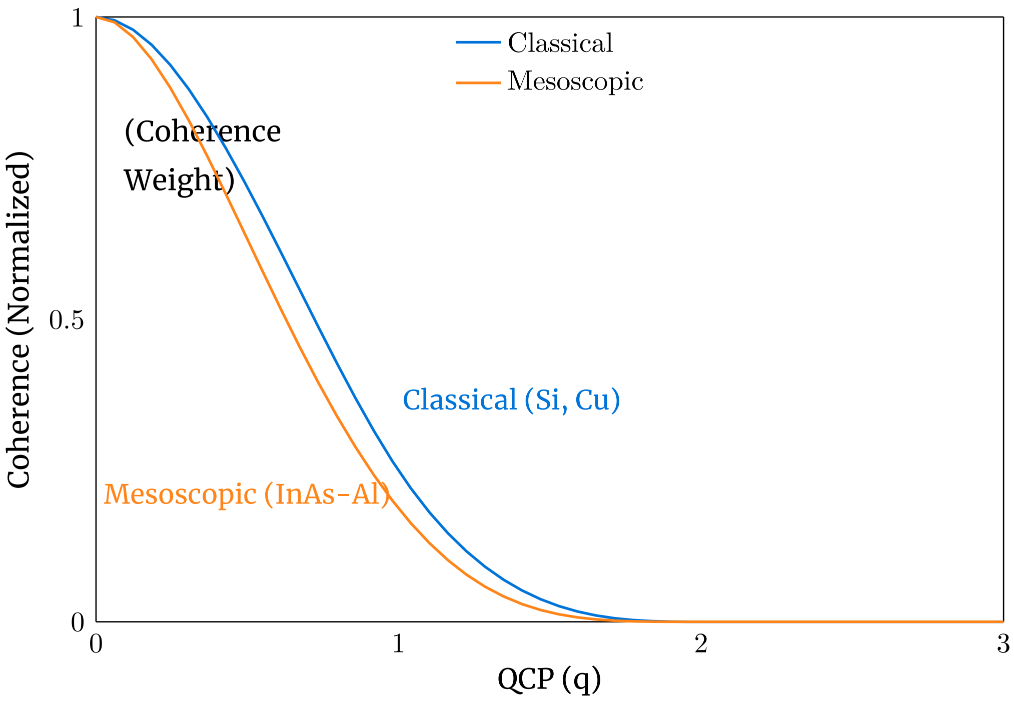
Mesoscopic Majoranas, due to their potential in quantum technologies, may be subject to export controls. Exporters should check US regulations like the Export Administration Regulations (EAR) for licensing requirements, especially for devices with quantum computing applications.

7.3–I Fabrication Method

Various methods, e.g., **lithography**, were noted, which is broad but accurate, encompassing techniques like chemical vapor deposition (CVD) and lithography for patterning. **Molecular Beam Epitaxy** (MBE) is standard for growing high-quality InAs-Al interfaces, confirmed by recent studies on epitaxial growth (Control over epitaxy and the role of the InAs/Al interface in hybrid two-dimensional electron gas systems - APS). Molecular-beam epitaxy(MBE) can be employed for Al,Ga, As, P, Mn, Cu , Si and C which are core materials for transistors and semiconductor chips.

7.3–II QCP Transitions in Chip Materials

This graph maps the Quantum Coherence Parameter q to nanoscale behavior, distinguishing **Classical** (e.g., silicon, copper) and **Mesoscopic** (e.g., InAs-Al) materials used in AI chips. q reflects coherence strength, with Classical materials dominating at $q < 1$ (>10 nm scales) and Mesoscopic materials peaking at $q = 1$ (10-100 nm), critical for quantum-enhanced chips under export scrutiny.



- **Classical Regime ($q < 1$)**: Silicon and copper dominate at >10 nm scales, with coherence dropping as q rises. These materials face standard export controls (e.g., US: 8% Cu exports).
- **Mesoscopic Peak ($q \approx 1 - 2$)**: InAs-Al nanowires (10-100 nm) peak, enabling topological states like Majoranas for quantum AI chips, triggering stricter export oversight (e.g., US EAR).
- **Export Impact**: Materials with $q \gg 1$ (mesoscopic) face heightened restrictions due to quantum potential, unlike classical counterparts.

7.4 Material Showdown

- **Classical (Si, Cu)**: High coherence at $q = 0$ (>10 nm) for standard chips, fading fast as q climbs—export-friendly but tech-limited.
- **Mesoscopic (InAs-Al)**: Peaks slightly above 1 at $q \approx \frac{1}{2} - 1$ (10-100 nm), ideal for quantum AI, then drops—export-sensitive due to strategic value.

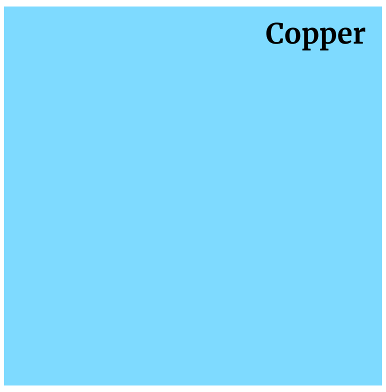
In this context, the x-axis represents the quantum parameter q, which governs the transition from classical to mesoscopic regimes. The y-axis, denoted $W(q)$, represents the quantum mechanical response of the system, measured in nanometers, illustrating how the system's behavior evolves as q changes.

8 Materials-Critical Export Laws : Sino-American Restrictions

In the complex landscape of global trade, **China & the US** have implemented significant restrictions on the export of critical materials such as **gallium, germanium, and antimony,** (CSIS 2024) which are vital for producing semiconductors and other advanced technologies. Additionally, with the **United States**, it has imposed strict export controls on advanced chips and semiconductor technologies, aiming to limit China's access to cutting-edge tech amid an ongoing trade war. Meanwhile, countries like Congo, despite being rich in minerals such as cobalt—essential for batteries and electronics—have not established similar critical export restrictions, possibly reflecting different economic priorities or geopolitical positioning.

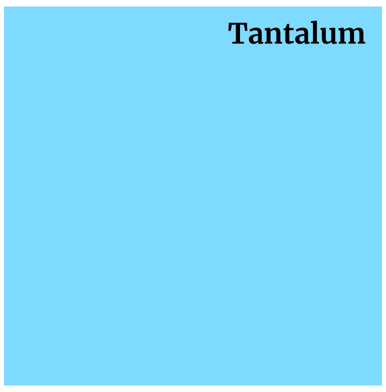
8.1 Copper Exports: Key Players

China cited national security concerns, the minerals have "dual military and civilian uses" (CSIS 2024). The U.S. and China dominate the supply of critical minerals like germanium, gallium, and InAs-Al, which are essential for AI chips, from classical silicon to quantum topological systems. China controls 60% of global germanium exports, while the U.S. leads in InAs-Al fabrication for Majorana-based qubits (Luo et al., 2020). Recent policies, such as China's 2024 mineral export bans and the U.S. CHIPS Act, escalate this rivalry, with both nations citing national security to restrict access. Copper, vital for chip interconnects, is less restricted (U.S.: 8%, China: 60% of exports), but its role pales compared to quantum-critical materials. This competition underscores the need for a minerals-first export framework to secure AI supply chains and align global standards



US - 8% of world exports

China - 60% of world exports



Rwanda - 22% of world exports

8.2 Compliance Workflow

```
Assess chip: material, quantity, **QCP** ($q$).
    Silicon/Germanium (Si | Ge) : $q \ll 1$, classical rules.
    InAs-Al: $q \gg 1$, quantum/topological rules.
    Verify end-use, user, destination.
Apply controls:
    **if** $q \gg 1$: Strict quantum licenses.
    **else**: Standard licenses.
Process order:
    Secure permits, document fully.
    Track shipment, report per $q$.
```

8.3 Workflow Explained:

Chips are evaluated by material and Quantum Coherence Parameter (**QCP**, q). Classical silicon/germanium ($q \ll 1$) follow standard export rules; topological InAs-Al ($q \gg 1$) require quantum-specific controls due to nanoscale (10-100 nm) and 1D Majorana features. Compliance scales with tier and q ([EAR - BIS, 2024](https://www.bis.doc.gov)).

8.4 Key Recommendations for Minerals-first Export Controls

Compute-focused restrictions fail to address the intricate role of minerals in chip design. Classical chips (silicon, germanium, >10 nm) and topological InAs-Al (1D nanowires, $10 - 100$ nm, hosting Majoranas) show that material complexity —tracked by QCP (q)—drives quantum potential, not just compute power.

A minerals-first approach is critical:

- **Minerals-First Controls**
  - Prioritize restrictions on raw materials (e.g., copper, tantalum) over chips, as they shape design at $q \gg 1$.
- **QCP-Driven Oversight**:
  - Scale controls with $q \gg 1$ for topological systems, reflecting mineral-enabled quantum leaps.
- **Tracking**:
  - Monitor mineral flows to chips, detailing quantum specifics.
- **Global Standards**:
  - Standardize QCP across classical-to-quantum materials.
  - Harmonize mineral export policies [Wassenaar Arrangement, 2024](https://www.wassenaar.org).

Current chip-centric rules miss the mark—minerals like germanium and tantalum inform InAs-Al's 10-100 nm quantum edge, not just compute specs. Unified, minerals-first restrictions ensure compliance and secure supply chains for AI innovation.

8.5 Remark

While minerals-first controls and cryptographic evaluation sharing offer significant benefits, challenges remain. Industry stakeholders may resist stricter regulations due to increased compliance costs, and non-Wassenaar nations may prioritize economic growth over harmonized standards. To address these, policymakers could offer tax incentives for adopting VDF-based evaluations and establish regional task forces to align export policies

8.6 Conclusion

The rapid advancement of AI, coupled with its national security implications, demands a paradigm shift in export controls and evaluation sharing. This technical brief demonstrates that traditional chip-focused regulations, rooted in size and compute metrics, are ill-equipped to address the strategic importance of critical materials like germanium, InAs-Al, and copper, which underpin both classical and quantum AI hardware. By adopting a minerals-first export control framework, guided by the Quantum Coherence Parameter (q), policymakers can better regulate dual-use technologies at mesoscopic scales (10-100 nm).

Simultaneously, cryptographic tools such as Verifiable Delay Functions (VDFs) and Proofs of Replication (PoRs) offer robust mechanisms for secure, transparent AI evaluation sharing, fostering trust and collaboration across nations. Case studies, including Japan's defense transfer model, illustrate practical pathways for implementation, while historical precedents like Bernstein v. United States underscore the need to balance innovation with security.

To operationalize these findings, we recommend:

- Global Standards: Harmonize minerals-first export controls through frameworks like the Wassenaar Arrangement, prioritizing materials with high quantum potential ($q > 1$).
- Cryptographic Integration: Mandate VDFs and PoRs in AI evaluation protocols to ensure integrity and prevent misuse.
- International Collaboration: Establish multilateral governance bodies to standardize AI safety metrics and share evaluations responsibly.
- Supply Chain Transparency: Track mineral flows in AI chip production to enhance compliance and security.

As AI reshapes global security, inaction risks ceding strategic advantage to adversaries. By embracing minerals-first controls and secure evaluation sharing, nations can harness AI's potential while safeguarding against its risks, ensuring a future where innovation and security coexist.

Bibliography

2023 Aghaee

Inas-Al Hybrid Devices Passing the Topological Gap Protocol. Physical Review B 107(24). American Physical Society (APS). http://dx.doi.org/10.1103/PhysRevB.107.245423.

1999 Appeals, Ninth Circuit Court of

Bernstein V. U.S. Dept. Of Justice, 192 F.3d 1308 (9th Cir. 1999). https://openjurist.org/192/f3d/1308.

1992 Buckland, Michael K.

Emanuel Goldberg, Electronic Document Retrieval, And Vannevar Bush's Memex. Journal of the American Society for Information Science 43(4): 284–294. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199205%2943%3A4%3C284%3A%3AAID-ASI3%3E3.0.CO%3B2-0.

2024 Burden, John

Evaluating AI Evaluation: Perils and Prospects. https://arxiv.org/abs/2407.09221.

2025 Bureau of Industry and Security

Export Administration Regulations: Updates on Cryptography and Semiconductor Controls. U.S. Department of Commerce. https://www.bis.doc.gov/index.php/encryption-and-export-administration-regulations-ear.

2018 Cabinet Secretariat of Japan

Defense Partnerships and Technology Transfer Report. https://www.cas.go.jp/jp/gaiyou/jimu/pdf/r60326_bouei10.pdf.

58AD Caesar, Julius, and Aulus Hirtius

Commentarii De Bello Gallico. Julius Caesar.

2025 China Daily

Researchers Develop Flash Memory Device. https://www.fudan.edu.cn/en/2025/0417/c344a145016/page.htm.

2017 Christiano, Paul F., Jan Leike, Tom B. Brown, et al.

Deep Reinforcement Learning from Human Preferences. Advances in Neural Information Processing Systems 30. https://arxiv.org/abs/1706.03741.

2015 Cole, William S., S. Das Sarma, and Tudor D. Stanescu

Effects of Large Induced Superconducting Gap on Semiconductor Majorana Nanowires. Physical Review B 92(17). American Physical Society (APS). http://dx.doi.org/10.1103/PhysRevB.92.174511.

2024 Congressional Research Service

Export Controls and Emerging Technologies: Implications for AI and Encryption. https://doi.org/10.51593/20190001.

2003 Contributors, Wikipedia

Bernstein V. United States. https://en.wikipedia.org/wiki/Bernstein_v._United_States.

2006 Cover, Thomas M., and Joy A. Thomas

Elements of Information Theory. 2nd edition. Wiley-Interscience.

2024 CSIS

China Imposes Its Most Stringent Critical Minerals Export Restrictions yet Amidst Escalating U.S.-China Tech War. https://www.csis.org/analysis/china-imposes-its-most-stringent-critical-minerals-export-restrictions-yet-amidst.

2001 Daemen, Joan, and Vincent Rijmen

The Design of Rijndael: AES - the Advanced Encryption Standard. Springer.

2002 Daemen, Joan, and Vincent Rijmen

The Design of Rijndael: AES — the Advanced Encryption Standard. Springer-Verlag.

2014 Editors, Britannica

Bernstein V. The U.S. Department of State. https://www.britannica.com/event/Bernstein-v-the-U-S-Department-of-State.

2020 European Commission

Emerging Technologies: Developments in the Context of Dual-Use Export Controls - Factsheets. https://mvep.gov.hr/UserDocsImages/2024/datoteke/Dokument%201.pdf.

2015 Expression, Columbia Global Freedom of

Bernstein V. Department of Justice. https://globalfreedomofexpression.columbia.edu/cases/bernstein-v-department-justice.

2019 Fisch, Benjamin, Rafael Pass, and Abhi Shelat

Verifiable Delay Functions. *In* Advances in Cryptology – CRYPTO 2019 Pp. 757–786. Springer.

2011 Foundation, Electronic Frontier

Bernstein V. US Department of Justice. https://www.eff.org/cases/bernstein-v-us-dept-justice.

2015 Foundation, Electronic Frontier

EFF at 25: Remembering the Case That Established Code as Speech. https://www.eff.org/deeplinks/2015/04/remembering-case-established-code-speech.

2006 Furht, Borko, ed.

The RSA Public-Key Encryption Algorithm. *In* Encyclopedia of Multimedia P. 757. Boston, MA: Springer US. https://doi.org/10.1007/0-387-30038-4_206.

2023 Gao, Yue, Ilia Shumailov, and Kassem Fawaz

SEA: Shareable and Explainable Attribution for Query-Based Black-Box Attacks. https://arxiv.org/abs/2308.11845.

2009 Gentry, Craig

A Fully Homomorphic Encryption Scheme. https://crypto.stanford.edu/craig/craig-thesis.pdf.

2024a Glukhov, David, Ziwen Han, Ilia Shumailov, Vardan Papyan, and Nicolas Papernot

Breach by a Thousand Leaks: Unsafe Information Leakage in `safe' AI Responses. https://arxiv.org/abs/2407.02551.

2024b Glukhov, David, Ziwen Han, Ilia Shumailov, Vardan Papyan, and Nicolas Papernot

Breach by a Thousand Leaks: Unsafe Information Leakage in 'Safe' AI Responses. https://arxiv.org/abs/2407.02551.

2023 Glukhov, David, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan

LLM Censorship: A Machine Learning Challenge or a Computer Security Problem?. https://arxiv.org/abs/2307.10719.

1993 Handy, Jim

The Cache Memory Book. USA: Academic Press Professional, Inc.

2021 Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, and others

Measuring Mathematical Problem Solving with the MATH Dataset. Neurips Datasets and Benchmarks Track. https://arxiv.org/abs/2103.03874.

2023 Kinniment, Max, and others

Evaluating Language Models for Self-Replication and Robustness. METR Technical Report. https://arxiv.org/abs/2312.11671.

2023 Lee, Jaehak, Nuri Kang, Seok-Hyung Lee, et al.

Fault-Tolerant Quantum Computation by Hybrid Qubits with Bosonic Cat-Code and Single Photons. https://arxiv.org/abs/2401.00450.

2014 Lerner, Sergio Demian

Proof of Replication.

2020 Luo, Xi, Yu-Ge Chen, Ziqiang Wang, and Yue Yu

Topological Superconductor from Superconducting Topological Surface States and Fault-Tolerant Quantum Computing. https://arxiv.org/abs/2003.11752.

2018 Lutchyn, Roman M., Erik P. A. M. Bakkers, Leo P. Kouwenhoven, and others

Majorana Zero Modes in Superconductor-Semiconductor Heterostructures. Nature Reviews Materials 3: 52–68. https://doi.org/10.1038/s41578-018-0003-1.

2024 Ministry of Foreign Affairs of Japan

The Three Principles on Transfer of Defense Equipment and Technology. https://www.mofa.go.jp/policy/security/three_principles.html.

2025 National Institute of Standards and Technology

Cryptographic Standards for AI Security: Guidelines for 2025. https://doi.org/10.6028/NIST.SP.800-227.ipd.

2022 Ouyang, Long, Jeff Wu, Xu Jiang, et al.

Training Language Models to Follow Instructions with Human Feedback. Advances in Neural Information Processing Systems 35. https://arxiv.org/abs/2203.02155.

2025 Roadmap to Fault Tolerant Quantum Computation Using Topological Qubit Arrays

. https://arxiv.org/abs/2502.12252.

1948 Shannon, Claude E.

A Mathematical Theory of Communication. Bell System Technical Journal 27(3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

2025 Shumailov, Ilia, Daniel Ramage, Sarah Meiklejohn, and others

Trusted Machine Learning Models Unlock Private Inference for Problems Currently Infeasible with Cryptography. https://arxiv.org/abs/2501.08970.

2025 Shumailov, Ilia, Daniel Ramage, Sarah Meiklejohn, Kairouz, et al.

Trusted Machine Learning Models Unlock Private Inference for Problems Currently Infeasible with Cryptography. https://arxiv.org/abs/2501.08970.

2022 Smith, David R., and John B. Pendry

Metamaterials: Progress and Challenges in Practical Applications. Nature Photonics 16: 123–134. https://doi.org/10.1038/s41566-021-00945-2.

2024 Tamkin, Alex, Miles McCain, Kunal Handa, et al.

Clio: Privacy-Preserving Insights into Real-World AI Use. https://arxiv.org/abs/2412.13678.

2024 Urban, Michael, and Stuart Russell

AI and Catastrophic Risk: Assessing CBRN Capabilities. https://humancompatible.ai/news/2023/10/24/managing-ai-risks-in-an-era-of-rapid-progress/.

2023 Wassenaar Arrangement Secretariat

Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies. https://www.wassenaar.org/control-lists/.

2022 Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, and others

Emergent Abilities of Large Language Models. *In* Transactions on Machine Learning Research. https://arxiv.org/abs/2206.07682.

2023 Wikipedia Contributors

Commitment Scheme. Wikimedia Foundation.

2025 Xiang, Wang, C., Liu, C. et al., Y.

Subnanosecond Flash Memory Enabled by 2d-Enhanced Hot-Carrier Injection. https://doi.org/10.1038/s41586-025-08839-w.

2023 Zhang, Yong, and Konstantin S. Novoselov

Nanomaterials for Next-Generation Electronics: Opportunities and Synthesis Challenges. Advanced Materials 35(15): 2208456. https://doi.org/10.1002/adma.202208456.

Appendix A Appendix

A VDE proof ensures that:

1. The function $f_{c(X)}$ was executed correctly.
2. The execution required a provable minimum delay D.
3. The output cannot be forged or computed faster than a defined bound.

Given a sequence of token streams $X = \{t_1, t_2, ..., t_N\}$, the prover L_P computes the output $Y = f_{c(X)}$ and generates a proof π.

A.1 Construction of the Proof

A Verifiable Delay Function (VDF) enforces a delay constraint, ensuring sequential execution of $f_{c(X)}$. The proof mechanism consists of:

A.1.1 Sequential Computation Constraint

The function $f_{c(X)}$ is computed step by step such that each step depends on the previous one:

$h_1 = H(t_1)$, $h_2 = H(h_1, t_2)$ $...h_N = H\left(h_{\{N-1\}}, t_N\right)$

where H is a cryptographic hash function enforcing sequential execution.

Proof Verification

The adversary L_A verifies π using an efficient verification function V such that:

$V(Y, \pi) \to \{0, 1\}$

If $V(Y, \pi) = 1$, the proof is valid. Otherwise, the claim is rejected.

A.1.2 Security Properties

- Soundness: If L_P provides pi, then $f_{c(X)}$ must have been computed correctly.
- Completeness: A correctly computed $f_{c(X)}$ will always yield a valid proof pi.
- Sequentiality: The proof generation process ensures that L_P cannot shortcut the required computation.

A.2 Proof of Surprise in Evals

The **Kullback-Leibler** (KL) divergence measures the difference between two probability distributions. In the context of LLMs, P_t represents the model's output distribution at time t, and $P_{\{t-1\}}$ is the distribution at the previous time step. The information surprise S_t quantifies how much the model's output has changed, reflecting the model's adaptation to new information or shifts in context.

$D_{\{\text{KL}\}}(P \parallel Q) = \sum_{\{x \in \{X\}\}} P(x) \log \frac{P(x)}{Q(x)}$

In the context of LLMs, let $\{X\}$ represent the set of all possible outputs. At time t, the model's output distribution is P_t, and at time $t - 1$, it is $P_{\{t-1\}}$. The information surprise S_t measures the divergence between these distributions:

$S_t = \sum_{\{x \in \{X\}\}} P_{t(x)} \log \frac{P_{t(x)}}{P_{\{t-1\}(x)}}$

The first term, $\sum_{\{xin\{X\}\}} P_{t(x)} \log P_{t(x)}$, represents the negative entropy of the distribution at time t, denoted as $-H(P_t)$. The second term, $\sum_{\{x \in \{X\}\}} P_{t(x)} \log P_{\{t-1\}}(x)$, is the cross-entropy between P_t and $P_{\{t-1\}}$. Therefore, the information surprise can be interpreted as the difference between the cross-entropy and the entropy of the current distribution. Monitoring S_t helps in understanding how the model's outputs evolve, ensuring stability while allowing for necessary adjustments. By bounding S_t, we can control the model's responsiveness to new data, analogous to policy constraints in RL that prevent erratic behavior. This balance is crucial for maintaining coherent and reliable performance in dynamic environments.

Appendix B LLM Inference

B.1 Case Study: LLM Math Performance

Suppose Party A assumes Model B, an LLM, scores 40% on the MATH dataset (algebra, calculus problems) based on prior evals (Hendrycks et al. 2021). Country B reports perf(ΔModel B): 85% accuracy after fine-tuning. Surprisal, $S_t = D_{\{\mathrm{KL}\}}\left(P_{\{t\}} \mid P_{\{t-1\}}\right)$, soars as A's prior (40% expected) clashes with B's leap—hinting at emergent reasoning or compute scale (Wei et al. 2022). This shocks A, signaling strategic gaps.

B.2 Case Study: CBRN Capabilities

Suppose:

- System_A is a CBRN detection system with a prior performance estimate $p_A = 0.85$ (85% accuracy).
- System_B reported by Country B, has a performance $p_B = 0.92$ (92% accuracy).

Appendix C VDE Definition

VDE Definition

```
Initialize Evaluation Environment:
    Party A and Party B agree on evaluation parameters and protocols.
Conduct Model Evaluation:
    Party A provides standardized test inputs to Party B.
    Party B processes these inputs using Model B and records the outputs.
Generate Proof of Performance:
    Party B compiles the outputs and relevant performance metrics.
    Party B creates a verifiable report demonstrating Model B's performance on the shared evaluations.
Verify Performance Report:
    Party A reviews the report and validates the results against expected benchmarks.
    if the report meets agreed-upon standards:
        Party A accepts the model's performance.
    else:
        Party A raises concerns or requests further testing.
Correctness Property:
    Review the report and validates the results against expected benchmarks.
```