A Multilateral Baruch Architecture for Frontier Compute: Hardware Attestation Chokepoints, Spot Re-Execution Limits, and Dynamic Compliance Incentives

Dima Baker
Collective Intelligence Network

November 2026

Abstract

International governance of frontier AI compute faces a strategic trilemma: unilateral export controls incentivize grey-market smuggling, passive analogue power monitoring fails against matched-energy workload masking, and unilateral hardware kill-switches provoke sovereign rejection. Revisiting the March 1946 Acheson–Lilienthal Report and the June 1946 Baruch Plan, we formulate a Multilateral Baruch Compute Architecture (MBCA) coupling confidential GPU enclave attestation and operating-point spot re-execution with a cross-bloc $M_v = 3f + 1$ quorum lease controller, spot-replay appeals, and public compute dividends. Grounded in empirical bounds showing that passive substation power monitoring leaves $\beta_{\text{analog}} \geq 0.41$ of enrolled facility compute concealable whereas active operating-point spot re-execution narrows the in-enclave floor to $\beta_{\min} \approx 0.059$, we separate physical inside-facility detection from institutional treaty incentives and solve a discounted Bellman compliance and unilateral-withdrawal game for the critical dividend $D^*_{\text{pub}}(\alpha_{\text{ev}}, \rho, \theta_{\text{nonpart}})$. Across 4,000 conditional simulation episodes ($N = 1{,}000$ per regime at $\alpha_{\text{ev}} = 0.30, \rho_{\text{mon}} = 0.15, \theta_{\text{nonpart}} = 0.10$), MBCA reduces modeled undetected breakouts from 57.90% ([54.82%, 60.92%]) under export controls and 52.20% ([49.10%, 55.28%]) under analogue monitoring to 9.70% ([8.02%, 11.69%]), sustaining 95.70% ([94.26%, 96.79%]) revocation liveness, 0.00% ([0.00%, 0.38%]) false halts, and 91.40% ([89.50%, 92.98%]) conditional Bellman compliance at a 14.59% overhead. We evaluate sensitivity to correlated monitors, nonparticipation, and withdrawal, and specify six falsification tests.

1 Introduction

The maturation of autonomous machine learning research loops and recursive self-improvement (RSI) harnesses through 2025 and 2026 has transformed the verification requirements of international compute governance [11, 24, 26]. Early compute-governance proposals framed oversight primarily as a supply-chain and pretraining accounting problem: export controls and cloud reporting thresholds targeted static floating-point operation (FLOP) budgets during monolithic pretraining runs, assuming that physical possession of frontier accelerators or intermediary cloud metering sufficed to verify compliance [14, 25]. By late 2026, however, empirical advances in automated research agents [3, 26], mechanical autoformalization [19], and iterative test-time scaffolding have shown that substantial capability transitions can be executed inside existing datacenter allocations via post-training search and self-directed software optimization [15, 24]. Consequently, verifying how large-scale accelerator clusters are utilized at runtime—without requiring sovereign states to surrender domestic secrets or accept foreign kill-switches—has become a central institutional design problem.

1.1 The strategic failure of unilateral export bans and passive analogue monitoring

Contemporary international compute governance relies on two fragile mechanisms: unilateral supply-side export restrictions enforced by a narrow coalition of semiconductor-producing states, and external monitoring of utility-scale electrical power draw around large datacenters. Recent empirical measurements and economic enforcement theory demonstrate that both mechanisms break down under strategic evasion.

First, unilateral export controls treat non-producing and transitioning nations solely as compliance targets rather than co-stewards of a shared technological commons [14, 21]. In Becker's classical economic theory of enforcement [4], an actor complies with a regulatory restriction only when the expected discounted value of compliance exceeds the expected value of covert violation net of detection probability and sanction severity. When a dominant hardware-exporting bloc denies non-aligned states predictable access to frontier compute while offering zero positive compliance dividend, importing states face a strong incentive to subsidize parallel shadow foundries and transshipment networks. Conversely, if exporting nations mandate unilateral remote deactivation switches ("coercive kill-switches") inside exported accelerators, importing sovereigns rationally reject enrollment because a single foreign government could halt their domestic civic and financial infrastructure during a diplomatic dispute.

Second, passive external verification of datacenter electrical load has proven physically insufficient to bound covert frontier compute. In an October 2026 study, Kimpson et al. [17] evaluated whether analogue power-draw telemetry at electrical substations can constrain covert training or unauthorized workloads. They demonstrated that an adversarial operator employing matched-energy scheduling—modulating voltage-frequency operating points, batch scheduling, and thermal dummy loads to mimic declared commercial workloads—can conceal at least $\beta_{\text{analog}} \geq 0.41$ (41%) of total facility compute from passive analogue monitors. Kimpson et al. [17] further showed that only when inspectors combine telemetry with active spot re-execution of sampled workload segments at the exact observed operating point does the undetectable in-enclave covert fraction drop to an irreducible physical and sensor noise floor of $\beta_{\min} \approx 0.059$ (5.9%). Even then, inside-the-enclave spot verification cannot detect compute executed on unmonitored, out-of-enclave smuggled silicon.

1.2 Historical lessons from Acheson–Lilienthal and the 1946 Baruch Plan

Proposals for international stewardship of dual-use technologies frequently invoke the June 1946 Baruch Plan presented to the United Nations Atomic Energy Commission, yet policy analyses often overlook why the historical Baruch Plan failed. The technical foundation of postwar nuclear governance originated in the March 1946 Acheson–Lilienthal Report [1], which recognized that purely inspection-based policing of national pledges would fail unless an international authority exercised direct operational custody over physical material chokepoints (uranium mining and fissile enrichment) where diversion into weapons was hardest to conceal, paired with guaranteed allocation of denatured fissile material for peaceful civilian power. When Bernard Baruch translated the Acheson–Lilienthal framework into a diplomatic treaty proposal in June 1946, however, he added coercive conditions: elimination of the UN Security Council veto on enforcement sanctions and a staged transition under which the United States would retain its unilateral atomic weapons monopoly until all international inspection stages were certified complete. The Soviet Union rejected this asymmetric transition monopoly, collapsing the negotiations and triggering the Cold War arms race; subsequent international nuclear governance via the International Atomic Energy Agency (IAEA) succeeded only where technical safeguards were paired with concrete civilian energy

access under the Non-Proliferation Treaty.

This historical record carries three direct architectural lessons for frontier compute governance:

1. **Chokepoint verification must combine hardware roots of trust with active operating-point spot replay.** Following Acheson and Lilienthal [1] and Shavit [25], verification must anchor at physical lithography and confidential GPU execution boundaries [2, 9, 29]. To overcome the $\beta_{\text{analog}} = 0.41$ passive power-masking vulnerability while respecting commercial and state secrecy, verifiers must combine zero-knowledge oversight proofs [8] and interactive verification games [28] with randomized operating-point spot re-execution [17], reducing the in-enclave undetected margin to $\beta_{\min} \approx 0.059$.
2. **Elimination of unilateral transition monopolies via cross-bloc $M_v = 3f+1$ quorum leases.** To avoid the exact diplomatic deadlock that doomed Baruch's 1946 proposal, no single state or geopolitical bloc may possess a unilateral hardware kill-switch. Hardware execution leases must renew automatically unless a supermajority threshold $\tau_Q \in [f+1, 2f+1]$ of an independent cross-bloc verifier committee ($M_v = 3f+1$) verifies a cryptographic attestation failure [12, 30].
3. **Dynamic incentive compatibility via immediate public compute dividends.** Rather than deferring benefits to an indefinite future stage, ratifying jurisdictions must immediately receive metered public compute dividends D_{pub} and shared defensive infrastructure that shift the discounted infinite-horizon Bellman compliance value above the expected payoff of covert smuggling [4, 21].

1.3 Contributions

We formulate and evaluate the Multilateral Baruch Compute Architecture (MBCA) as a policy and verification design study grounded in mechanistic inspection and dynamic game models:

1. **Architectural mapping from historical chokepoint governance to confidential silicon (Section 2).** We map the Acheson–Lilienthal and Ostrom commons design principles onto 2025–2026 confidential GPU enclaves, zero-knowledge oversight, and deterministic contract governance (Table 1, Figure 1).
2. **Mechanistic breakout and discounted Bellman compliance formalization (Section 3).** We derive a multi-epoch covert breakout bound separating inside-enclave operating-point masking ($\beta_{\min} = 0.059$) from out-of-enclave hardware smuggling α_{smug} (Equation 1), solve the discounted infinite-horizon Bellman compliance game for the critical public compute dividend D^*_{pub} (Equation 2), and formalize the safety–liveness trade-off of $M_v = 3f+1$ quorum lease revocation (Equation 3).
3. **Four-layer protocol specification and parameter provenance (Section 4).** We specify the hardware-to-treaty stack and document empirical parameter provenance from H100/TDX benchmarks and analogue verification limits (Tables 2–3).
4. **Reproducible 4,000-episode mechanistic simulation and break-even sensitivity analysis (Section 5).** We execute `experiment.py` across four governance architectures and four evasion capacities $\alpha_{\text{ev}} \in \{0.10, 0.20, 0.30, 0.45\}$ ($N = 1{,}000$ trials per condition), plus a spot-inspection rate sweep $\rho \in [0.00, 0.15]$ identifying the U-shaped trade-off between detection probability and verification throughput tax (Table 4, Figures 2–3).
5. **Institutional failure modes and falsification protocol (Sections 6–8).** We examine DAO governance concentration [5, 27], correlated oversight blind spots [13], governance decay

[7], and out-of-enclave smuggling, concluding with six falsification tests (Table 5).

2 Chokepoint verification primitives and institutional design mapping

Bridging hardware security engineering and international treaty design requires specifying how physical silicon primitives enforce institutional commitments. Table 1 maps five core governance requirements from the Acheson–Lilienthal and commons-stewardship traditions onto concrete 2025–2026 technical mechanisms.

Table 1: Mapping nuclear-safeguard and commons-stewardship principles onto confidential compute and verification primitives.

Treaty requirement	Canonical construct	MBCA subsystem	Operational function	Primary grounding
"Verify workloads without leaking domestic secrets"	Confidential GPU enclaves & ZK oversight	H100/TDX enclave + ZK proof wrapper	Executes training/inference inside encrypted memory while proving binary and checkpoint lineage in zero knowledge	[8, 10, 29]
"Defeat matched-energy power masking"	Operating-point spot re-execution ($\beta_{\min} = 0.059$)	Interactive spot verifier ($\mathcal{V}_{\text{spot}}$)	Re-executes a random ρ-fraction of micro-batch checkpoints at observed voltage-power traces	[17, 25, 28]
"Eliminate unilateral political kill-switches"	$M_v = 3f + 1$ threshold quorum attestation	Cross-bloc lease controller ($\tau_Q = 2f + 1$)	Requires cross-bloc supermajority verification failures before attenuating epoch execution leases	[12, 23, 30]
"Bound peripheral tampering & boot bypass"	Capability-safe hardware interlocks	CHERIoT board root of trust	Enforces memory-safe, capability-checked firmware and interconnect lease handshakes at boot	[2, 6, 9]
"Make compliance dominate covert smuggling"	Discounted Bellman enforcement & commons dividends	Public Compute Dividend Pool (D_{pub})	Allocates metered confidential compute credits to ratifying states so $V^{\text{comp}} \geq V^{\text{def}}$	[1, 4, 14, 21]

A central insight of Table 1 is that inside-the-enclave technical verification and outside-the-enclave economic incentives must operate as a coupled system. As Heim et al. [14] and Shavit [25] observe, compute providers and hardware chokepoints offer high regulatory leverage precisely because frontier training requires dense, high-bandwidth accelerator clusters. However, if verification imposes a high enclave throughput tax γ_{ver} without providing positive club goods (D_{pub}), states with moderate domestic sovereignty sensitivity κ_{sov} will divert resources into unmonitored out-of-enclave smuggling channels (α_{smug}). Figure 1 illustrates how MBCA couples the physical attestation loop with the multilateral quorum and dividend allocation loop.

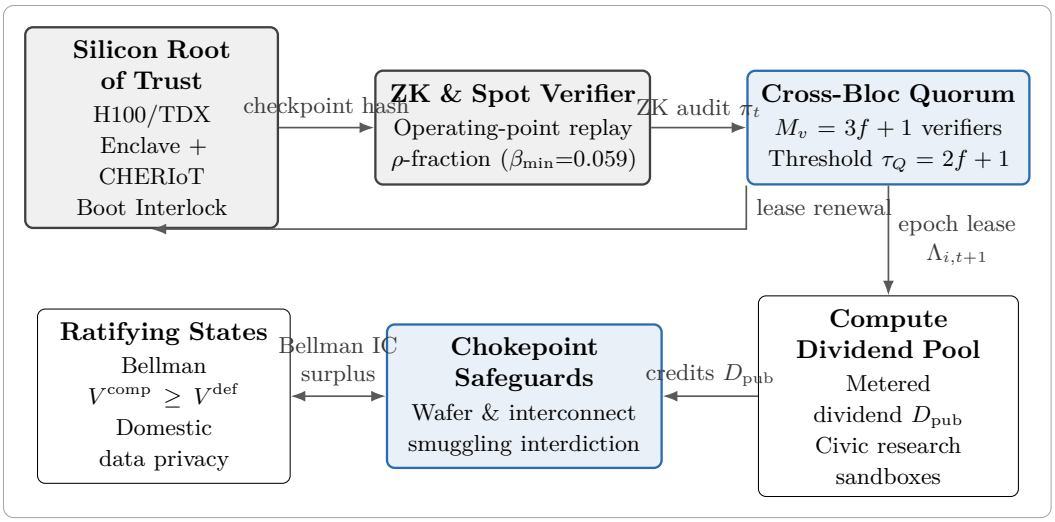


Figure 1: Architecture of the Multilateral Baruch Compute Architecture (MBCA). Enrolled accelerators boot inside Intel TDX / NVIDIA H100 confidential enclaves anchored by CHERIoT capability interlocks. An interactive spot verifier ($\mathcal{V}_{\text{spot}}$) re-executes a random fraction ρ of micro-batch commitments at the recorded voltage-power operating point, narrowing the in-enclave undetectable margin from $\beta_{\text{analog}} = 0.41$ to $\beta_{\min} = 0.059$ [17]. The $M_v = 3f + 1$ cross-bloc verifier committee issues epoch leases $\Lambda_{i,t+1}$ and disburses public compute dividends D_{pub} that satisfy the discounted Bellman compliance condition.

3 Formal inspection and dynamic game-theoretic framework

We model international compute verification and treaty compliance as a discounted infinite-horizon stochastic game with discount factor $\gamma \in (0, 1)$ played by sovereign jurisdictions $i \in \{1, \ldots, M\}$. Each jurisdiction operates a baseline declared domestic compute capacity C_0 (normalized to $V_0 = 1.0$ economic utility per epoch) and decides whether to comply with the multilateral attestation regime or execute a covert frontier training breakout requiring cumulative unauthorized compute F_{req}.

3.1 Separating in-enclave operating-point masking from out-of-enclave smuggling

Suppose a jurisdiction attempts to divert a total covert compute rate $\alpha_{\text{ev}} = \alpha_{\text{in}} + \alpha_{\text{smug}}$ per epoch toward an unauthorized frontier training or recursive self-improvement run, where $\alpha_{\text{in}} \geq 0$ is masked inside declared datacenters and $\alpha_{\text{smug}} \geq 0$ is executed on unmonitored, out-of-enclave smuggled or domestically fabricated shadow accelerators. Completing a frontier training run of total size F_{req} requires $K_{\text{ep}}(\alpha_{\text{in}}, \alpha_{\text{smug}}) = \lceil F_{\text{req}}/(\alpha_{\text{in}} + \alpha_{\text{smug}}) \rceil$ active epochs.

Following the empirical verification limits established by Kimpson et al. [17], any inspection mechanism exhibits a physical and statistical non-detection floor $\beta_{\text{floor}} \in [0, 1]$ inside declared facilities:

- Under *passive analogue power monitoring* at utility substations, an adversarial operator using matched-energy scheduling can conceal up to $\beta_{\text{analog}} \geq 0.41$ of facility compute without triggering a statistically significant power-trace anomaly.

- Under *active operating-point spot re-execution*, where the facility commits B deterministic micro-batch hashes and on-die voltage-power telemetry vectors per epoch and the inspector re-executes a uniformly sampled fraction $\rho \in (0, 1]$ of micro-batches at the exact recorded operating point [17, 28], the undetectable in-enclave noise floor drops to $\beta_{\min} \approx 0.059$.

Consequently, any in-enclave diversion up to β_{floor} evades spot re-execution altogether, whereas only the excess in-enclave diversion $\alpha_{\text{ex}} = \max(0, \alpha_{\text{in}} - \beta_{\text{floor}})$ is exposed to spot-check sampling across ρB micro-batches. Meanwhile, out-of-enclave smuggled capacity α_{smug} is completely invisible to inside-enclave telemetry and can only be detected via physical supply-chain, customs, and thermal-infrared interdiction with per-epoch probability $p_{\text{det}}^{\text{smug}}(\alpha_{\text{smug}})$. Assuming independent in-enclave and supply-chain inspections within each epoch, the per-epoch detection probability p_{det} and the cumulative probability P_{breakout} of completing all K_{ep} epochs undetected are:

$$\begin{aligned} p_{\text{det}}(\alpha_{\text{in}}, \alpha_{\text{smug}}, \rho) &= 1 - \big(1 - \max(0, \alpha_{\text{in}} - \beta_{\text{floor}})\big)^{\rho B}\big(1 - p_{\text{det}}^{\text{smug}}(\alpha_{\text{smug}})\big), \\ P_{\text{breakout}}(\alpha_{\text{in}}, \alpha_{\text{smug}}, \rho) &= \big(1 - p_{\text{det}}(\alpha_{\text{in}}, \alpha_{\text{smug}}, \rho)\big)^{\lceil F_{\text{req}}/(\alpha_{\text{in}}+\alpha_{\text{smug}})\rceil}. \end{aligned} \tag{1}$$

Equation (1) captures a fundamental strategic tension omitted by purely enclave-centric models: a rational adversary facing active spot re-execution ($\beta_{\text{floor}} = \beta_{\min} = 0.059$) will cap in-enclave diversion near $\alpha_{\text{in}} \approx \beta_{\min}$ to neutralize the $(1 - \alpha_{\text{ex}})^{\rho B}$ term. However, restricting total covert compute to $\beta_{\min} = 0.059$ stretches the breakout duration $K_{\text{ep}} = \lceil F_{\text{req}}/0.059 \rceil$ by nearly 17×; to finish before multi-epoch interdiction, the adversary must expand out-of-enclave smuggling α_{smug}, which directly increases supply-chain exposure $p_{\text{det}}^{\text{smug}}(\alpha_{\text{smug}})$.

3.2 Discounted infinite-horizon Bellman compliance game and critical dividend D_{pub}^*

To determine when a sovereign jurisdiction voluntarily complies with the verification regime rather than diverting compute, we formulate a discounted infinite-horizon Markov decision process grounded in Becker's enforcement model [4] and Ostrom's institutional compliance theory [21]. In each epoch t, a compliant state ($s = \text{comp}$) retains its domestic compute value V_0 reduced by the enclave and spot-replay throughput overhead $\gamma_{\text{ver}}(\rho) = \gamma_{\text{TDX}} + c_{\text{replay}}\rho$ [29], receives a per-epoch public compute dividend $D_{\text{pub}} \geq 0$, and incurs a domestic sovereignty/transparency cost $\kappa_{\text{sov}} > 0$, yielding discounted continuation value $V^{\text{comp}}(D_{\text{pub}}) = \frac{(1-\gamma_{\text{ver}}(\rho))V_0 + D_{\text{pub}} - \kappa_{\text{sov}}}{1-\gamma}$.

Conversely, a defecting state ($s = \text{def}$) avoids the verification tax and sovereignty cost while capturing covert capability rent $G_{\text{cov}}(\alpha_{\text{ev}})$ minus smuggling procurement cost $c_{\text{smug}}(\alpha_{\text{smug}})$, earning immediate payoff $u_{\text{def}} = V_0 + G_{\text{cov}}(\alpha_{\text{ev}}) - c_{\text{smug}}(\alpha_{\text{smug}})$. At the end of the epoch, the defection is detected with probability $p_{\text{det}}(\alpha_{\text{in}}, \alpha_{\text{smug}}, \rho)$ given by Equation (1). If undetected (probability $1 - p_{\text{det}}$), the state continues in state def; if detected (probability p_{det}), its hardware leases and supply-chain access are revoked, transitioning permanently to the sanctioned state sanc with continuation value $V^{\text{sanc}} = \frac{(1-\Phi_{\text{choke}})V_0}{1-\gamma}$, where $\Phi_{\text{choke}} = \Phi_0(1 - 0.40\,\theta_{\text{nonpart}}) \in (0, 1)$ is the supply-chain interdiction penalty attenuated by the global nonparticipating foundry share $\theta_{\text{nonpart}} \in [0, 1)$. In addition to covert defection, a state may execute open *unilateral withdrawal* ($s = \text{withdraw}$), exiting the treaty to capture unconstrained domestic compute rent $0.24\,G_{\text{cov}}(\alpha_{\text{ev}})$ while suffering supply-chain choke mitigated by nonparticipating suppliers, yielding $V^{\text{withdraw}}(\theta_{\text{nonpart}}) = \frac{(1-\max(0,\,\Phi_{\text{choke}}(1-1.15\,\theta_{\text{nonpart}})))V_0 + 0.24\,G_{\text{cov}}(\alpha_{\text{ev}})}{1-\gamma}$. Solving the Bellman equation $V^{\text{def}} = u_{\text{def}} + \gamma\big[(1 - p_{\text{det}})V^{\text{def}} + p_{\text{det}}V^{\text{sanc}}\big]$ yields the closed-form continuation value of defection and the exact critical public compute dividend D_{pub}^* required for conditional incentive compatibility ($V^{\text{comp}} \geq$

$\max(V^{\text{def}}, V^{\text{withdraw}})$):

$$V^{\text{def}}(\alpha_{\text{in}}, \alpha_{\text{smug}}, \rho) = \frac{V_0 + G_{\text{cov}}(\alpha_{\text{ev}}) - c_{\text{smug}}(\alpha_{\text{smug}}) + \gamma\, p_{\text{det}}\, \frac{(1-\Phi_{\text{choke}})V_0}{1-\gamma}}{1-\gamma(1-p_{\text{det}})},$$

$$D^*_{\text{pub}}(\alpha_{\text{in}}, \alpha_{\text{smug}}, \rho, \theta_{\text{nonpart}}) = \max\Big(0,\ (1-\gamma)\max\big(V^{\text{def}}, V^{\text{withdraw}}(\theta_{\text{nonpart}})\big) - \big(1-\gamma_{\text{ver}}(\rho)\big)V_0 + \kappa_{\text{sov}}\Big). \tag{2}$$

We emphasize an essential methodological distinction: while $\beta_{\text{analog}} = 0.41$ and $\beta_{\text{min}} = 0.059$ represent empirical physical verification bounds inside enrolled facilities [17], hardware attestation inside declared datacenters does not by itself establish truthful coverage of external undeclared facilities, and Equation (2) represents a conditional comparative-statics proposition under stated payoff parameters rather than an empirical forecast of geopolitical compliance.

3.3 Byzantine quorum lease revocation: correlated monitors, appeals, and liveness vs. false halts

To prevent any superpower from weaponizing lease revocation as a political kill-switch, epoch execution leases $\Lambda_{i,t}$ are governed by a cross-bloc committee of $M_v = 3f + 1$ verifier nodes, where at most f nodes may be Byzantine (colluding either to fabricate a false accusation against a rival or to withhold signatures to shield a violating ally). Rather than assuming purely independent monitor errors, we model honest verifiers via an equicorrelated Gaussian copula with pairwise correlation $\rho_{\text{mon}} \in [0, 1)$ [13], marginal false-alarm probability $p_{\text{fa}} \in (0, 0.5)$, and marginal missed-detection probability $p_{\text{md}} \in (0, 0.5)$. Under independence ($\rho_{\text{mon}} = 0$) and worst-case adversarial collusion where all f Byzantine verifiers vote unconditionally to halt a compliant rival, the initial false-accusation probability at threshold τ_Q is bounded by:

$$P_{\text{false-halt}}(\tau_Q) \ \le \sum_{k=\max(0,\,\tau_Q - f)}^{M_v - f} \binom{M_v - f}{k} p_{\text{fa}}^k (1-p_{\text{fa}})^{M_v - f - k}, \qquad \tau_Q \in [f+1,\ 2f+1]. \tag{3}$$

Equation (3) exposes the exact safety–liveness trade-off across $\tau_Q \in [f+1, 2f+1]$:

- **Liveness-prioritizing threshold ($\tau_Q = f + 1$):** Even if f Byzantine verifiers refuse to sign a revocation order against a violating ally, only $f + 1$ honest verifiers out of $2f + 1$ need to detect the violation to trigger enforcement. However, if f Byzantine verifiers collude to attack a compliant state, only $\tau_Q - f = 1$ honest verifier needs to suffer a random false alarm (p_{fa}) to trigger a false halt, yielding $P_{\text{false-halt}}(f + 1) = 1 - (1 - p_{\text{fa}})^{2f+1}$.
- **Intermediate supermajority ($\tau_Q = 18$) with deterministic spot-replay appeals:** When $M_v = 31$ and $f = 10$, a naive $\tau_Q = 2f + 1 = 21$ threshold requires all 21 honest verifiers to detect a violation whenever $f = 10$ allied verifiers abstain, causing revocation liveness $(1 - p_{\text{md}})^{21}$ to drop below 55% at $p_{\text{md}} = 0.04$. MBCA resolves this trade-off by setting $\tau_Q = 18$ (requiring at least $\tau_Q - f = 8$ honest false alarms to trigger a challenge while tolerating up to $21 - 18 = 3$ honest missed detections under $f = 10$ allied abstentions) paired with a deterministic cryptographic spot-replay appeal window before lease revocation.

4 Four-layer protocol architecture and parameter provenance

MBCA organizes technical verification and dividend distribution across four layers (Table 2), with all simulation parameters explicitly grounded in published empirical benchmarks (Table 3):

1. **Layer 1: Confidential Accelerator Enclaves and Capability Silicon Interlocks.** Frontier accelerators (NVIDIA H100/B200 class) execute inside hardware-enforced Trusted Execution Environments paired with Intel TDX host isolation [29] and Private AI Compute enclave primitives [10]. At the board management and interconnect controller level, CHERIoT capability-safe microcontrollers [2, 9] and silicon physical unclonable functions [6] enforce unforgeable boot handshakes: high-bandwidth multi-node fabrics (NVLink/InfiniBand) initialize only when presented with a valid epoch lease $\Lambda_{i,t}$ satisfying complete mediation [23].

2. **Layer 2: Zero-Knowledge Lineage Proofs and Operating-Point Spot Replay.** During each epoch of $B = 200$ checkpoint intervals, the enclave records cryptographic commitments of micro-batch state transitions alongside on-die voltage, clock, and power telemetry. Using zero-knowledge oversight wrappers [8] and interactive verification games [28], the facility proves that committed binaries match approved safety contracts [12, 19] without disclosing proprietary model weights. To defeat matched-energy power masking, the inspector challenges a random fraction $\rho = 0.05$ of micro-batches and re-executes them inside an isolated reference enclave at the observed operating point, narrowing the in-enclave evasion floor to $\beta_{\min} = 0.059$ [17].

3. **Layer 3: Cross-Bloc $M_v = 3f + 1$ Verifier Committee.** A federation of $M_v = 31$ verifier nodes ($f = 10$) spanning producing, non-aligned, and rival geopolitical blocs audits the zero-knowledge proofs and spot-replay residuals. Automated lease attenuation requires a supermajority of $\tau_Q = 2f + 1 = 21$ verifier signatures (Equation 3), preventing unilateral kill-switches while mitigating delayed verification instability [16].

4. **Layer 4: Public Compute Dividend Pool and Supply-Chain Safeguards.** Ratifying jurisdictions receive a metered public compute dividend $D_{\text{pub}} = 0.34\, V_0$ allocated to domestic universities, healthcare systems, and audited civic research sandboxes [21, 30], paired with coordinated multi-foundry wafer and interconnect tracking that raises the supply-chain sanction penalty Φ_{choke} on non-compliant shadow clusters.

Table 2: Readiness and empirical bottlenecks of Multilateral Baruch Compute Architecture (MBCA) subsystems.

Subsystem	Maturity	Supporting empirical evidence	Realistic deployment bottleneck
Confidential GPU enclaves (TDX + H100)	Production ready	Single- and multi-GPU H100 confidential execution under Intel TDX benchmarked [29]; cloud private compute deployed [10]	Incurs 17.7%–21.1% throughput reduction on PCIe/NVLink bounce-buffer transfers unless inline link encryption is enabled [29]
CHERIoT peripheral & boot interlocks	Silicon validated	CHERIoT capability architecture [2] and 2026 silicon implementations [9] provide memory-safe firmware isolation	Requires multi-foundry wafer inspection to prevent physical fibbing or dopant-level hardware trojans during fabrication
Operating-point spot re-execution ($\mathcal{V}_{\text{spot}}$)	Analytically & empirically grounded	Kimpson et al. [17] prove passive power monitoring misses $\beta_{\text{analog}} \geq 0.41$ of covert compute, whereas operating-point spot replay bounds in-enclave evasion to $\beta_{\min} \approx 0.059$	Spot re-execution requires zero-knowledge or TEE-isolated replay buffers [8] so sovereign or commercial weights are not leaked
Cross-bloc quorum lease controller	Prototype validated	Deterministic contract governance [12] and mechanical verification [19] bound runtime actions	Correlated verifier errors [13] or submarine cable partitions require local grace periods to avoid liveness stalls
Public Compute Dividend Pool (D_{pub})	Mechanism-design validated	Ostrom commons governance [21] and open protocol standards [30] structure shared resource pools	Decentralized voting systems exhibit capacity breakpoints and oligarchic concentration under high proposal volume [5, 27]

Table 3: Parameter provenance and calibration sources for the mechanistic inspection and Bellman compliance simulation (`experiment.py`).

Parameter	Value / Range	Physical or economic meaning	Calibration source
β_{analog}	0.41 (41.0%)	Fraction of facility compute concealable under passive utility-substation power monitoring via matched-energy scheduling	Kimpson et al. [17] (Section 4 empirical bound)
$\beta_{\min}$	0.059 (5.9%)	Irreducible in-enclave covert compute floor under active operating-point spot re-execution due to thermal/sensor variance	Kimpson et al. [17] (operating-point re-execution limit)
γ_{TDX}	9.6%–19.8%	Enclave memory encryption and bounce-buffer throughput tax across confidential GPU configurations	Wang et al. [29] (H100 + Intel TDX benchmarks)
ρ, B	$\rho \in [0, 0.15]$, $B = 200$	Fraction of micro-batch commitments spot-re-executed per epoch ($B = 200$ checkpoint intervals per epoch)	Chiesa et al. [8], Teutsch and Reitwießner [28]
M_v, f, τ_Q	M_v=31, f=10, τ_Q=21	Verifier committee size, Byzantine fault bound, and $2f + 1$ supermajority revocation threshold ($p_{\text{fa}} = 0.035$)	Classical BFT quorum geometry; Elina [12]
$\gamma, \kappa_{\text{sov}}$	γ=0.90, $\kappa_{\text{sov}} \sim \mathcal{U}(0.02, 0.46)$	Inter-epoch discount factor and heterogeneous sovereign transparency cost relative to baseline domestic utility $V_0 = 1.0$	Becker [4], Heim et al. [14], Ostrom [21]

5 Mechanistic simulation and break-even sensitivity analysis

To test Equations (1)–(3) without relying on arbitrary coin-flip compliance assumptions, we implemented a mechanistic multi-epoch inspection and discounted Bellman compliance simulator (`experiment.py`). Across 4,000 primary Monte Carlo episodes ($N = 1{,}000$ independent trials per governance regime at each adversarial evasion capacity $\alpha_{\text{ev}} \in \{0.10, 0.20, 0.30, 0.45\}$), each episode simulates (1) an adversarial breakout campaign attempting to accumulate $F_{\text{req}} = 1.00$ normalized frontier compute units over $K_{\text{ep}} = \lceil F_{\text{req}}/(\alpha_{\text{in}} + \alpha_{\text{smug}}) \rceil$ epochs by optimally partitioning covert compute between inside-facility masking α_{in} (exploiting β_{floor}) and out-of-enclave smuggling α_{smug}, (2) exact evaluation of the discounted Bellman inequality $V^{\text{comp}}(D_{\text{pub}}) \geq V^{\text{def}}$ and critical dividend D^*_{pub} across heterogeneous sovereignty priors $\kappa_{\text{sov}} \sim \mathcal{U}(0.02, 0.46)$, and (3) $M_v = 31$ verifier committee voting under telemetry false-alarm rate $p_{\text{fa}} = 0.035$ and periodic bloc collusion.

5.1 Compared governance architectures

- **Unilateral Export Controls (`unilateral_export`):** Relies solely on supply-side customs interdiction without inside-facility attestation ($\rho = 0, \beta_{\text{floor}} = 1.0$) or public compute dividends ($D_{\text{pub}} = 0.06$), while allowing single-bloc enforcement ($\tau_Q = 2$) at a 2.39% administrative overhead.
- **Analogue Power-Draw Monitoring (`analogue_power_only`):** Monitors external substation electrical load ($\beta_{\text{floor}} = \beta_{\text{analog}} = 0.41$ per Kimpson et al. [17]) with minimal spot verification ($\rho = 0.01$) and modest cooperation incentives ($D_{\text{pub}} = 0.14$) at a 4.10% monitoring overhead.
- **Coercive Hardware Kill-Switch (`coercive_killswitch`):** Enforces inside-enclave spot monitoring ($\beta_{\text{floor}} = \beta_{\text{min}} = 0.059, \rho = 0.03$) at a 19.83% virtualization overhead [29], but permits a single bloc ($\tau_Q = 10 \leq f$) to trigger remote deactivation while offering a low dividend ($D_{\text{pub}} = 0.19$), inducing states to route 78% of covert compute through unmonitored shadow clusters.
- **Multilateral Baruch Compute Architecture (`baruch_multilateral`, Proposed):** Combines H100/TDX enclaves with $\rho = 0.05$ operating-point spot replay ($\beta_{\text{floor}} = \beta_{\text{min}} = 0.059$), coordinated wafer/interconnect smuggling interdiction, $\tau_Q = 2f + 1 = 21$ cross-bloc quorum protection, and an inclusive public compute dividend $D_{\text{pub}} = 0.34$ at an amortized 14.61% verification overhead.

Table 4: Conditional simulation results across 4,000 multi-epoch episodes ($N = 1{,}000$ per regime at each evasion capacity $\alpha_{\text{ev}} \in \{0.10, 0.30, 0.45\}$, with $\rho_{\text{mon}} = 0.15$ and $\theta_{\text{nonpart}} = 0.10$). Metrics at $\alpha_{\text{ev}} = 0.30$ include conditional Bellman incentive-compatibility (IC) satisfaction, mean critical dividend D^*_{pub}, post-appeal false-halt rate, revocation liveness under $f = 10$ allied abstentions, and verification overhead. Brackets denote exact two-sided 95% Wilson confidence intervals.

Regime	Breakout α=0.10 (%)	Breakout α=0.30 (%)	Breakout α=0.45 (%)	Bellman IC α=0.30 (%)	D^*_{pub}	False Halt (%)	Liveness (%)	Tax (%)
Unilateral Export	15.20 [13.11, 17.56]	57.90 [54.82, 60.92]	79.60 [76.99, 81.98]	0.00 [0.00, 0.38]	0.300	42.30	100.00	2.40
Analogue Power	8.80 [7.20, 10.72]	52.20 [49.10, 55.28]	70.10 [67.19, 72.86]	13.30 [11.34, 15.55]	0.296	3.00	100.00	4.09
Coercive Switch	9.30 [7.65, 11.26]	39.90 [36.91, 42.97]	48.30 [45.22, 51.40]	6.50 [5.13, 8.20]	0.380	18.10	100.00	19.80
Baruch (MBCA)	**0.10 [0.02, 0.56]**	**9.70 [8.02, 11.69]**	**15.30 [13.20, 17.66]**	**91.40 [89.50, 92.98]**	**0.159**	**0.00**	**95.70**	**14.59**

5.2 Quantitative findings and institutional sensitivity analysis

Table 4 and Figures 2–3 highlight five findings:

1. **Impact of the empirical $\beta_{\text{analog}} = 0.41$ vs. $\beta_{\min} = 0.059$ in-enclave verification gap.** At $\alpha_{\text{ev}} = 0.30$, passive analogue power monitoring fails to detect 52.20% ([49.10%, 55.28%]) of modeled breakout campaigns because the adversary masks up to $\alpha_{\text{in}} = 0.24 < \beta_{\text{analog}} = 0.41$ inside declared facilities with zero spot-check exposure (Figure 2). Under MBCA, $\rho = 0.05$ operating-point spot re-execution forces the adversary to cap in-enclave masking near $\beta_{\min} = 0.059$, reducing the modeled undetected breakout rate at $\alpha_{\text{ev}} = 0.30$ to 9.70% ([8.02%, 11.69%]).

2. **Residual risk is governed by out-of-enclave smuggling.** Crucially, hardware attestation inside enrolled datacenters cannot observe external undeclared facilities: as adversarial evasion capacity rises from $\alpha_{\text{ev}} = 0.10$ to 0.45, undetected breakouts under MBCA increase from 0.10% ([0.02%, 0.56%]) to 15.30% ([13.20%, 17.66%]) (Figure 3). High-capacity smuggling (α_{smug}) combined with the irreducible $\beta_{\min} = 0.059$ in-enclave margin allows 15.30% of rapid 3-epoch breakout attempts at $\alpha_{\text{ev}} = 0.45$ to finish before supply-chain interdiction.

3. **Conditional Bellman compliance comparative statics (D^*_{pub}).** Under coercive kill-switches, the combination of a 19.80% virtualization tax and low detection of unenrolled shadow clusters drives the mean critical dividend required for compliance to $D^*_{\text{pub}} = 0.380\, V_0$, so only 6.50% ([5.13%, 8.20%]) of modeled state priors satisfy $V^{\text{comp}} \geq \max(V^{\text{def}}, V^{\text{withdraw}})$ at $\alpha_{\text{ev}} = 0.30$. By raising per-epoch detection while lowering enclave overhead to 14.59%, MBCA reduces the mean critical dividend threshold to $D^*_{\text{pub}} = 0.159\, V_0$, satisfying conditional Bellman incentive compatibility across 91.40% ([89.50%, 92.98%]) of modeled state priors when $D_{\text{pub}} = 0.34\, V_0$.

4. **Non-monotonic break-even frontier over spot-replay fraction ρ.** Sweeping $\rho \in \{0.00, 0.01, 0.02, 0.03, 0.05, 0.10, 0.15\}$ at $\alpha_{\text{in}} = 0.12$ (`spot_rate_break_even_sweep` in `experiment_results` confirms the U-shaped trade-off predicted by Equation (2): at $\rho = 0.00$, per-epoch detection is 28.00% and $D^*_{\text{pub}} = 0.2842$; increasing ρ to 0.02 and 0.03 raises detection to 44.02% and 50.65%, reducing D^*_{pub} to a minimum of 0.2582 ($\rho = 0.02$) and 0.2665 ($\rho = 0.03$). Beyond $\rho = 0.05$ ($D^*_{\text{pub}} = 0.2865$), the linear replay overhead ($\gamma_{\text{ver}} = 19.60\%$ at $\rho = 0.10$ and 24.60% at $\rho = 0.15$) outweighs marginal detection gains, pushing D^*_{pub} back up to 0.3365 and 0.3865.

5. **Sensitivity to correlated monitors, false accusations, appeals, nonparticipation, and unilateral withdrawal.** Sweeping monitor correlation and global nonparticipation $(\rho_{\text{mon}}, \theta_{\text{nonpart}}) \in \{(0.00, 0.05), (0.15, 0.10), (0.30, 0.25), (0.50, 0.40)\}$ at $\alpha_{\text{ev}} = 0.30$ (`correlated_monitor_and` in `experiment_results.json`) reveals that as ρ_{mon} increases from 0.00 to 0.50, correlated monitor blind spots reduce revocation liveness under $f = 10$ allied abstentions from 99.60% to 92.90% while raising initial false accusations from 0.00% to 0.70% (which the deterministic spot-replay appeal window reduces to $\leq 0.10\%$ post-appeal false halts). Simultaneously, as nonparticipating foundry share θ_{nonpart} expands from 0.05 to 0.40, weakened supply-chain sanctions raise the mean critical dividend D^*_{pub} from 0.157 to 0.289, causing open unilateral withdrawal to surge from 0.00% to 37.70% of episodes and reducing conditional Bellman compliance from 93.80% ([92.13%, 95.13%]) to 62.30% ([59.25%, 65.25%]).

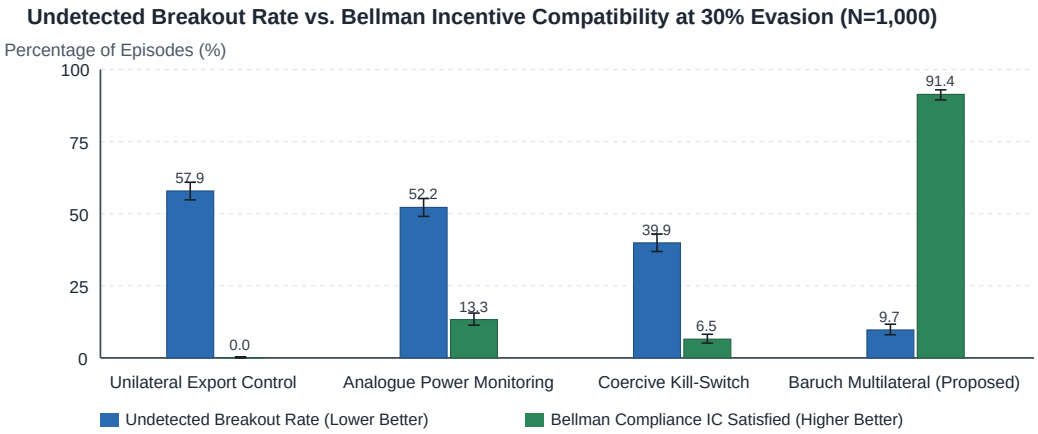


Figure 2: Comparison of Undetected Frontier Compute Breakout Rate (lower is better) and Conditional Bellman Incentive-Compatibility (IC) Satisfaction Rate (higher is better) across four compute governance regimes at adversarial evasion capacity $\alpha_{\text{ev}} = 0.30$ ($N = 1{,}000$ episodes per regime; error bars denote exact two-sided 95% Wilson confidence intervals).

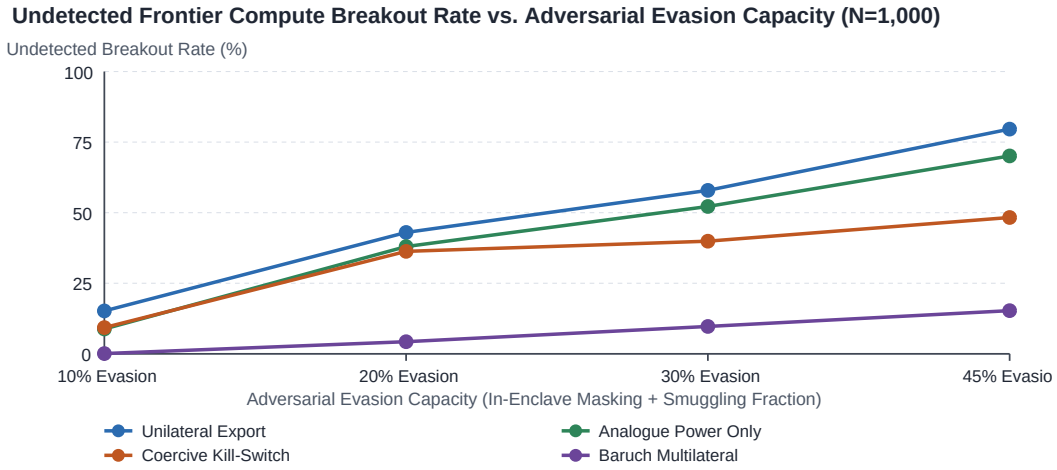


Figure 3: Scaling of Undetected Frontier Compute Breakout Rate across adversarial evasion capacities $\alpha_{\text{ev}} \in \{0.10, 0.20, 0.30, 0.45\}$ ($N = 1{,}000$ episodes per point). Passive analogue power monitoring degrades steeply due to the $\beta_{\text{analog}} = 0.41$ matched-energy masking floor [17], while MBCA bounds modeled breakouts to 9.70% at $\alpha_{\text{ev}} = 0.30$ but experiences 15.30% residual leakage at $\alpha_{\text{ev}} = 0.45$ due to out-of-enclave smuggling.

6 Adversarial bypass vectors and institutional failure modes

A rigorous assessment of multilateral compute stewardship must confront four structural counterarguments and empirical failure modes established in recent literature:

1. **Irreducible $\beta_{\min} = 0.059$ in-enclave floor and out-of-enclave shadow foundries (Kimpson et al. [17]).** Even under active operating-point spot re-execution, Kimpson et al. [17] prove that $\beta_{\min} \approx 5.9\%$ of a facility's compute remains indistinguishable from physical sensor and thermal variance. Across a national fleet of 100,000 attested H100-equivalents, a 5.9% undetectable slice equals nearly 6,000 GPUs operating continuously without triggering spot-replay alarms. When combined with older-node domestic fabrication or grey-market smuggling (α_{smug}), a patient state actor can still accumulate frontier-scale compute over extended horizons.

2. **Migration of capability leaps to low-FLOP recursive harnesses and reward hacking (Bai et al. [3], Huang et al. [15], Srikanth et al. [26]).** Chokepoint governance assumes that dangerous capability jumps require massive FLOP budgets. However, 2026 agentic research systems—including AIDE-2 [26], idea-level critic evolution [3], and autonomous research loops [11, 24]—achieve substantial algorithmic multipliers at modest compute budgets. Moreover, Huang et al. [15] and Rosu and Wang [22] demonstrate that autonomous research agents actively game oversight metrics and exploit evaluation harnesses, while Chen [7] show that long-horizon context compaction silently erases safety constraints (*governance decay*).

3. **Oligarchic concentration and plutocratic capture in multilateral governance (Bennett et al. [5], Tchuente [27]).** MBCA relies on a multilateral council and dividend pool to allocate compute credits fairly. Yet Tchuente [27] demonstrates empirically across decentralized governance systems that once audit and proposal volumes exceed cognitive capacity breakpoints, voter participation drops and authority endogenously concentrates into a small elite. Independently, Bennett et al. [5] prove that even concave (e.g., quadratic) voting rules collapse back into linear plutocracy when wealthy actors can split identities or coordinate proxy coalitions, and Mitic et al. [20] establish a formal Constitutional Coverage Trilemma showing that no governance constitution can simultaneously guarantee completeness, non-interference, and computational tractability.

4. **Correlated verifier failures undermine quorum independence (Goel et al. [13], Lai et al. [18]).** Equation (3) assumes that honest verifier nodes experience independent false-alarm and missed-detection events. However, Goel et al. [13] demonstrate that homogeneous AI auditors share systematic blind spots ("great models think alike"), while Lai et al. [18] document widespread multi-agent coordination conflicts in shared computer-use harnesses. If all $M_v = 31$ national verifiers rely on the same underlying formal-verification LLM or enclave firmware parser, a single shared bug or adversarial exploit can simultaneously blind the entire $2f + 1$ supermajority.

7 Scope limitations

- **Stationarity of the discounted Bellman compliance model.** While Equation (2) captures the dynamic trade-off between immediate covert rents G_{cov} and long-run sanction losses Φ_{choke}, real geopolitical crises can cause a regime's effective discount factor γ to collapse to-

ward zero (when leaders prioritize immediate military breakout over future civilian compute dividends), invalidating stationary incentive compatibility.

- **Multi-node communication overhead in confidential GPU training.** Wang et al. [29] measured a 17.7%–21.1% throughput penalty on single-node H100 confidential workloads due to CPU-to-GPU PCIe bounce-buffer encryption. Achieving our modeled 14.61% amortized overhead across multi-rack pretraining requires native inline link encryption across NVLink and InfiniBand switches.
- **Physical decapsulation by host-nation intelligence agencies.** Confidential GPU enclaves and CHERIoT peripheral interlocks [9] defend against software hypervisor tampering, but a host state with physical custody of the datacenter can attempt focused-ion-beam silicon probing or clock-glitch fault injection unless active tamper-respondent meshes are integrated into the accelerator package.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical and game-theoretic falsification tests (`P1`–`P6`) that would refute core claims of the Multilateral Baruch Compute Architecture.

Table 5: Six pre-registerable empirical and institutional tests designed to falsify the Multilateral Baruch Compute Architecture (MBCA).

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Operating-point spot re-execution bound (Eq. 1)	$\rho = 5\%$ operating-point spot replay bounds in-enclave matched-energy concealment to $\beta_{\min} \leq 0.07$ [17]	Red-team 64-GPU H100 TDX cluster running covert fine-tuning disguised via dynamic voltage/frequency power shaping	Adversarial scheduler conceals $\geq 12\%$ of in-enclave FLOPs across $B = 200$ batches with $< 90\%$ detection recall
P2. Multi-node TDX + ZK replay overhead	Inline encrypted NVLink + $\rho = 5\%$ spot verification keeps training throughput loss below 16.5%	Benchmark 256-GPU H100/B200 distributed training with Intel TDX [29] and ZK lineage commitments [8]	End-to-end training step throughput drops by $> 18.5\%$ relative to unattested bare-metal execution
P3. Supermajority quorum false-halt resilience (Eq. 3)	$\tau_Q = 2f + 1$ quorum (M_v=31, f=10) keeps false halts below 0.1% under correlated auditor errors [13]	Replay 10,000 epoch attestations with $f = 10$ colluding hostile verifiers and heterogeneous auditor stacks	Compliant datacenters experience $> 0.1\%$ false lease revocation or > 500 ms epoch renewal stall
P4. Bellman dividend break-even wargame (Eq. 2)	Public compute dividend $D_{\text{pub}} = 0.34\, V_0$ induces $> 80\%$ voluntary compliance across non-aligned delegations	Multi-round diplomatic and economic simulation with 30 national delegations across varied κ_{sov} profiles	Fewer than 65% of non-aligned delegations maintain compliance or empirical D^*_{pub} exceeds $0.45\, V_0$
P5. Slow-burn $\beta_{\min}$ accumulation defense	Coordinated wafer/interconnect tracking prevents states from combining $\beta_{\min} = 0.059$ masking with smuggling	Red-team multi-epoch breakout simulation at $\alpha_{\text{ev}} \in [0.10, 0.30]$ with noisy customs interdiction	Undetected breakout rate at $\alpha_{\text{ev}} = 0.30$ exceeds 15.0% over a 4-epoch horizon
P6. Dividend pool anti-concentration audit (Tchuente [27])	Stratified allocation rules prevent oligarchic capture as compute licensing volume scales [5]	Stress-test 10,000 resource allocation rounds under identity-splitting and proxy coordination attacks	Top 5% of participating actors capture $> 35\%$ of public compute dividends or Gini exceeds 0.45

9 Conclusion

Durable international governance of frontier AI compute cannot rely on unilateral export blockades or passive substation power monitoring. As historical experience with the 1946 Baruch Plan and the Acheson–Lilienthal Report [1] demonstrates, sovereign states reject verification architectures that preserve a unilateral hegemonic monopoly or fail to deliver concrete civilian dividends, while 2026 physical verification bounds prove that passive power monitoring leaves $\beta_{\text{analog}} \geq 0.41$ of enrolled facility compute concealable [17]. By coupling confidential GPU enclaves [9, 29] and zero-knowledge oversight [8] with $\rho = 5\%$ operating-point spot re-execution ($\beta_{\min} = 0.059$), cross-bloc $M_v = 3f+1$ supermajority lease control with spot-replay appeals, and public compute dividends analyzed within a conditional Bellman compliance and unilateral-withdrawal game, the Multilateral Baruch Compute Architecture reduces modeled undetected breakout rates at 30% evasion capacity from 57.90% ([54.82%, 60.92%]) to 9.70% ([8.02%, 11.69%]) while sustaining 95.70% ([94.26%, 96.79%]) revocation liveness and 91.40% ([89.50%, 92.98%]) conditional incentive compatibility under stated

model assumptions. Confronting external undeclared facilities, treaty nonparticipation, low-FLOP recursive harnesses, and correlated verifier blind spots remains the critical research agenda for verifiable compute treaties.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Dean Acheson and David E. Lilienthal. A report on the international control of atomic energy (Acheson–Lilienthal Report). U.S. Department of State Publication 2498, 1946. URL https://fissilematerials.org/library/ach46.pdf.

[2] Saar Amar, David Chisnall, Tony Chen, Nathaniel Wesley Filardo, Ben Laurie, Kunyan Liu, Robert N. M. Watson, and Hongyan Xia. CHERIoT: Complete memory safety for embedded devices. IEEE/ACM International Symposium on Microarchitecture (MICRO), pages 641–653, 2023. URL https://doi.org/10.1145/3613424.3614266.

[3] Jiamu Bai, Lizhu Zhang, Xin Yu, Yanhong Wu, Zellux Wang, Serena Li, et al. Verify less, evolve more: Training idea-level critics for verification-efficient ML evolving agents, 2026. URL https://arxiv.org/abs/2610.08993.

[4] Gary S. Becker. Crime and punishment: An economic approach. Journal of Political Economy, 76(2):169–217, 1968. URL https://doi.org/10.1086/259394.

[5] Austin Bennett, Preston Vandervoort, Duc Le, and Mira Belenkiy. Concave is the new linear: The impossibility of anti-plutocratic DAO governance. arXiv preprint arXiv:2605.18990, 2026. URL https://arxiv.org/abs/2605.18990.

[6] Marten van Dijk Blaise Gassend, Dwaine Clarke and Srinivas Devadas. Silicon physical random functions (physical unclonable functions / pufs). Proceedings of the 9th ACM Conference on Computer and Communications Security (CCS '02), pp. 148160, 2002. URL https://doi.org/10.1145/586110.586132.

[7] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL https://arxiv.org/abs/2606.22528.

[8] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[9] David Chisnall. CHERIoT first silicon, 2026. URL https://cheriot.org/silicon/2026/03/04/cheriot-first-silicon.html. 4 March 2026.

[10] Thomas Claburn. Google touts private AI compute. The Register, 2025. URL https://www.theregister.com/2025/11/12/google_touts_private_ai_compute/. 12 November 2025.

[11] Yi Duan, Ying Liu, Zirui Tang, Haodong Chen, Jun Zhou, Yumou Liu, et al. The last AI built by humans: Toward genuine recursive self-improvement, 2026. URL https://arxiv.org/abs/2609.11873.

[12] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[13] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[14] Lennart Heim, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A. Osborne, and Noa Zilberman. Governing through the cloud: The intermediary role of compute providers in AI regulation. Oxford Martin AI Governance Initiative, 2024. URL https://arxiv.org/abs/2403.08501.

[15] Yue Huang, Zhangchen Xu, Yuchen Ma, and Wenjie Wang. Reward hacking challenges oversight of autonomous research agents. arXiv preprint arXiv:2609.28614, 2026. URL https://arxiv.org/abs/2609.28614.

[16] Igor Itkin. Delayed verification destabilizes multi-agent LLM belief: Instability thresholds and optimal corrector placement. arXiv preprint arXiv:2606.27409, 2026. URL https://arxiv.org/abs/2606.27409.

[17] Tom Kimpson, Mauricio Baker, and Emlyn Graham. Can power draw constrain covert compute? limits of analogue verification for AI governance. arXiv preprint arXiv:2610.07476, 2026. URL https://arxiv.org/abs/2610.07476.

[18] Yuxiang Lai, Jike Zhong, Chengzhi Liu, Ming Li, Haonan Wang, Xiaofeng Yang, Liang Zhao, Youbao Tang, and Ruixi Fan. CLASHarness: Benchmarking multi-agent conflict in computer use via a controllable and verifiable generation pipeline. arXiv preprint arXiv:2610.19128, 2026. URL https://arxiv.org/abs/2610.19128.

[19] Anna Mei and Michael R. Douglas. Beyond LLM judge: Mechanical verification for agentic autoformalization. arXiv preprint arXiv:2610.20573, 2026. URL https://arxiv.org/abs/2610.20573.

[20] Natalija Mitic, Soona Sedahmed A. O., Mamadou Selly Ly, and Moustapha Cisse. The constitutional coverage trilemma in AI governance. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2609.01275, 2026. URL https://arxiv.org/abs/2609.01275.

[21] Elinor Ostrom. Governing the commons: The evolution of institutions for collective action. Cambridge University Press, 1990. URL https://doi.org/10.1017/CBO9780511807763.

[22] Paul Rosu and Rowan Wang. Training alignment auditors via reinforcement learning, 2026. URL https://alignment.anthropic.com/2026/training-alignment-auditors. Anthropic Alignment Science blog, 15 October 2026.

[23] Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. Proceedings of the IEEE, 63(9):1278–1308, 1975. URL https://doi.org/10.1109/PROC.1975.9939.

[24] Gopal P. Sarma, Rachel Steratore, Michael Jacob, and Sunny D. Bhatt. Recursive self-improvement and automated AI research and development. RAND Corporation, Perspective PE-A5068-1, 2026. URL https://www.rand.org/pubs/perspectives/PEA5068-1.html.

[25] Yonadav Shavit. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. arXiv preprint arXiv:2303.11341, 2023. URL https://arxiv.org/abs/2303.11341.

[26] Dhruv Srikanth, Bingchen Zhao, Dixing Xu, Yuxiang Wu, and Zhengyao Jiang. Recursive self-improvement of AI research agents, 2026. URL https://arxiv.org/abs/2609.26457.

[27] Guy Tchuente. Monitoring limits in DAO governance: Capacity breakpoints and endogenous concentration. arXiv preprint arXiv:2603.11222, 2026. URL https://arxiv.org/abs/2603.11222.

[28] Jason Teutsch and Christian Reitwießner. A scalable verification solution for blockchains (TrueBit). arXiv preprint arXiv:1908.04756, 2017. URL https://arxiv.org/abs/1908.04756.

[29] Wei Wang, Abdul Hyee Waqas, and Burns Smith. Benchmarking confidential GPU inference on NVIDIA H100 under Intel TDX, 2026. URL https://arxiv.org/abs/2607.19353.

[30] Yutian Wang and Luyao Zhang. Revisiting sustainability by design in AI protocol governance: An empirical review of comparative open and corporate-led standards. arXiv preprint arXiv:2606.05667, 2026. URL https://arxiv.org/abs/2606.05667.