The Verification Model: Bounding What a Council Cannot See

Alex Stremsky · The Convoy Papers, No. 4 · v1.1 · September 2026 · CC-BY 4.0

DOI (all versions): 10.5281/zenodo.22838800. Companion to Preventing a Singleton in AI Development: A Council of Rivals *(The Convoy Papers, No. 3, 10.5281/zenodo.22019542), whose §4.1 this paper delivers. Supplements on the same record: A — the normative Reference Algorithm specification (v3.19); B — the reference implementation and its assertion output; C — the compute-model archive.*

Keywords: verification; AI governance; compute governance; arms control; anti-divergence; mechanism design; concealment; reference algorithm

Abstract. A council that limits any member's lead over the rest must be able to see how large each member's capability is — and the footprints that verification has always relied on are shrinking, not because anyone hides them but because increased efficiency, leased and recycled hardware, and the silent copying of know-how remove them. This paper replaces the classical question, *is this member verified?*, with a budget: the council does not need to see everything — it needs to bound what it cannot see, and to keep that bound below what its margin arithmetic tolerates. We price nine observation channels by their blindness, catalogue eight concealment plays with their counters and residues, and show that the gap between the margin at which the council engages a diverging member and the margin it defends serves two purposes at once — a concealment allowance and a reaction buffer — which fixes the gap's bounds. We then present a reference algorithm for the council's continuous decision: a per-member allowance that, by one line of arithmetic, enforces a coalition-level safety bound without the council ever having to detect a coalition; a remedy menu in which the member computes its own least-costly compliance; clocks that are computed rather than negotiated; and an escalation ladder that ends outside the council. The numbers come from models that can be run, and the models were built by failing: a compute floor fails at every level, a rank-based throttle fails instructively, and an individual, continuous, margin-anchored throttle holds. A hardware-supply neutrality architecture closes the cheaper attack — starvation — with obligations, statistical detection, relief, engagement, and deterrence of force. One attack is stated as unsolved: the hidden efficiency flip, which the design bounds to a single evaluation cycle and does not prevent. The specification, a conformance implementation, and the model archive accompany the paper as supplements.

Relation to The Convoy Papers No. 3. *A Council of Rivals* §4.1 promised a companion on the watch: how the council verifies capable rivals when remnants vanish, what the instruments are, how openness and work substitute for surveillance, what spying does and does not buy, and what an efficiency flip does to all of it. This paper is that companion. It supplies the verification model (§§1–4), the reference algorithm the council runs (§5), the hardware neutrality architecture (§6), the constitution of the sensor system itself (§7), the weaknesses and the falsifiers (§8), and the models the numbers come from (§9). Where No. 3 states a design decision, this paper carries the arithmetic behind it; where No. 3 cites *The Reference Algorithm*, *The Verification Model*, or *Hardware Neutrality* by name, this is a section here.

Version note. v1.1 (September 2026): corrected cross-references and two typos; §9 and Supplement C now carry only the eight models behind this paper, the four serving later companions having moved to their records; the DOI line now gives the version-agnostic identifier.

Acknowledgment. Drafting, literature anchoring, and structural review of this paper were carried out in collaboration with Claude (Anthropic). The claims, the design, the choice of what to say, and any errors are the author's.

Contents

1 The reframe: verification is a budget, not a binary

Every arms-control regime that has ever worked relied on footprints. A weapons programme needs material, facilities, power, people, and time. Each of these leaves a remnant that an inspector can count, sample, or photograph. The verification literature is, at bottom, a literature about remnants: what they are, how long they persist, how expensively they can be hidden.

The footprints of a frontier AI programme are shrinking, for reasons that have nothing to do with concealment. Efficiency advances let the same capability be reached with less compute. Workloads can be moved onto hardware that was never counted as frontier. Leased capacity leaves no purchase. Recycled and non-standard hardware reaches the field through channels no export register tracks. The decisive input — know-how — copies without leaving anything behind. A programme that once required a visible datacentre may soon require a declaration and little else. Verification built on the assumption that a decisive capability must show can be defeated not by a determined concealer but by ordinary progress.

The classical framing makes this worse than it needs to be. It asks whether a member is "verified", as if verification were a state one enters or fails to enter, and it treats every undetected remnant as a failure of the regime. That framing collapses the moment any channel is imperfect, which every channel is.

This paper replaces it. Verification is a budget: the Council does not need to see everything — it needs to bound what it cannot see, and to keep that bound below what the margin arithmetic of the Council's design tolerates. Each observation channel is priced by its blindness — the fraction of capability a concealer could hold behind it. The channels are combined until the residual, the part that could be hidden across all of them at once, is small enough that a member holding it would still be inside its allowance. What the budget buys is not certainty. It is a published, quantified statement of what a hidden programme could at most be, and a reaction buffer sized to that number. The rest of the paper is the accounting.

2 What must be measured

2.1 Observables — an initial list

- compute stocks and acquisitions
- data stocks and acquisitions
- efficiency
- capability
- influence holdings (class F)
- the sharing graph
- development timing
- release-terms compliance

The actual list will be maintained by the Council. As capability inputs evolve, it will be constantly extended, reweighted and pruned, under §7.3's standing mandate and §6 of the reference specification (Supplement A)'s change process.

2.2 The mandate limit

The meter reads acquirable compute adjusted by declared efficiency, and the observables above. It does not read graded intelligence — the writ ends where member-recomputability ends.

2.3 The declaration's verification

(Closing the gap is confessed as W2 below.)

A declared efficiency is a falsifiable claim, not a testimony. Metered compute (the hardware channels) and measured capability (the evaluation channels) bound it from both sides.

The consistency test holds every declaration, within a published tolerance, against four independent readings:

- evaluation-suite performance
- released-artifact quality
- the member's own declaration history (trajectories are smooth — discontinuity flags, the biological-passport pattern)
- the peer band for the hardware generation

Breach contests the declaration with the burden on the declarer, adjudicated by the estimation lane. The specification's capability-per-declared-compute anomaly tracking (Supplement A §4) supplies the standing statistic.

Sustained frontier-scale input with persistently little declared result intensifies audit on the hazard-indexed ladder (Supplement A §5) — the concealer must sandbag every suite, forgo product revenue, and explain the megawatts at once, a cost rising with the hidden march's size.

Bar-clustering around the Developer seat threshold is published as its own statistic (the bunching estimator); persistent gaming is conduct — sanction class 28 of ACoR's sanction table, the attestation-fraud twin. The full accountancy elaboration (CUF, the significant-quantity calibration) is left to future work.

W2 below stands as the residual: compute-to-capability is not conserved, and the remainder lands in the scorecard's detection-dependent row.

2.4 Retirement declarations join the same apparatus

A declared retirement is a declaration, and rides the consistency test with three readings:

- the published *depreciation band* per hardware generation and asset class (retirement volume beyond the band flags)
- *disposal attestation* against the serial-level registry, physically verifiable by sampled inspection — the IAEA decommissioning pattern: disposal must be verified, not merely declared
- *timing correlation* — retirement clustering at engagement, at TEST, or upon an adverse offer is the fraud signature (the insurance-forensics and wash-sale precedents: the pattern flags without a need for intent inquiry)

The Measurement Organ reads indicators, the Inspectorate samples on anomaly, the Adjudication Organ takes contests with the burden on the declarer — one apparatus, using self-declared quantities.

This apparatus carries a heavier office too: it defines verified exit for the specification's capability accountancy. Unverified exits remain on the holder's ledger, for the holder's own test only (the ghost binds its maker), while every field statistic reads verified-present capability — the book binds its keeper, the field reads reality.

Class 29 narrows to disposal-attestation fraud, an attack on the apparatus itself.

2.5 Two further offices

The apparatus's readings now also trigger the compensation lane's *revival-by-surfacing* (Supplement A §5b):

- a capability appearing in the world — declared, used, deployed — within a debt period plus its class tail pays the outstanding tranches. The trigger for that is the sensors' ordinary work, not a new collection.
- the sharing graph computes coverage: each item's field coverage (whence $\xi = 1 -$ coverage) and each member's per-item lack, feeding the advantage stock, the opening credits, and credit-follows-lack — again from readings already made

2.6 The floor side introduces no new collection

The anti-attrition mechanism (the spec's §5b) consumes the same measurements — C_i, the median, the count — that the ceiling side already publishes. Support adds no sensor, no channel and no intrusiveness.

3 The observation channels and the concealer's playbook

3.1 The channels, each against its concealer

Plainly: the ways the world leaks information about who holds what — today's channel list — and, for each, what a determined hider can and cannot do about it.

The channel portfolio below is Council-maintained on §7.3's standing schedule. Channels decay, merge, and are added as architectures and chains evolve. The numbering is this document's, not a constant of the design.

3.1.1 Supply-chain remnants — the diverse chain, not just the frontier chain.

The frontier chain is narrow (§6's chokepoint map serves as the sensor grid and observer-suppliers serve as attestors). The capability-relevant chain however is wider, and the model prices every one of its branches:

3.1.1.a Trailing-edge fabs.

Quantity partially substitutes for frontier quality — at a rising power-and-interconnect cost per capability. Channel §3.1.1.b catches more of what this branch hides. Mass balance extends trailing-edge with **capability weighting**: chips weighted by capability contribution, so dark trailing stock counts proportionally less.

3.1.1.b New non-frontier fabs.

The Chinese buildout is the live case.

A non-cooperating production is the ledger's hard case. Non-cooperation makes verification hard, and excluding the capability would make the ledger useless.

The resolution is to stop pretending that the ledger closes. External production enters as an estimated term with published error bars, not as zero.

Fab construction is among the most observable industrial activities on earth. This is satellite-visible infrastructure that takes years to build, and the world tracks the current buildouts publicly. So, *capacity* is estimable at statistical grade while *allocation and quality* stay dark.

The budget becomes geographically structured: c_inside (cooperating chains - small) + c_outside (estimated, wide bars). The conservative end of the sum drives μ_eng.

Please note that as an unaccounted vertical completes, effective c climbs toward c*. At this point verification cannot substitute for membership. The answer to a completed non-cooperating stack is not better sensors - it is recruitment or the endgame instruments.

3.1.1.c Hidden contract production.

Unreported runs attack the production ledger itself. And some suppliers might not be members - that is, they are not compellable. Especially during a supply crunch, when the leverage runs the other way. Escalation above the Council is heavy machinery that hidden production, by definition, denies documentary proof for. The instruments that actually exist are:

- Contracting at accession, in whatever market weather prevails. The AMC pattern buys a multi-year demand floor, and the attestation covenant travels with the contract.

A supplier that wants the pool's aggregated demand accepts attestation as a condition of selling into the pool. Breach costs the demand floor plus ordinary contract remedies, no compulsion required.

The present hardware market crunch is not expected to end soon. (Most leading AI companies invest everything they can afford into compute, to not allow the competition to gain a decisive advantage, and pull all data of their clients to improve the training of their AIs. Many, if not the most of their customers are worried about misuse or leak of their data, and also invest everything they can into compute to build their own hardware infrastructure.)

Covenants signed in a sellers' market hold more weakly — though even a sellers' market prices multi-year demand certainty (take-or-pay contracts are signed in commodity booms for exactly this reason). There is also an adverse-selection dynamic: the covenant sorts the supplier population into three sorts:

- accepters, bound and less likely to offend
- refusers, facing the accepters' subsidized competition, more likely to offend
- toolless — reachable only physically (their output estimated into c_outside), and through their buyers (dark receipt).

A member that uses a dark receipt - that is, receives unledgered chips - can potentially be sanctioned. This can be discovered through watching for mass-balance anomalies and physical-side estimates (wafer starts, tool utilization, materials draw vs declared output).

Above-Council escalation carries a published materiality threshold: below it the leakage lives inside c; above it, it is by definition no longer its business.

3.1.1.d Recycled consumer compute.

An example could be a smartphone-scale fleet.

Can be cheaply obtained by the container from electronics recyclers. It is poor for dense training, but usable for the auxiliary tier. (See §3.2.7 for the danger this branch actually carries: a workload offload.)

Functional handsets by the container are costly. Discarded ones by the hundred containers are dirt-cheap. Many of these are still good. Most of the rest are repairable to farm grade (only CPU, GPU and RAM need to work). The remainder goes to a parts bin.

The sensor that detects this approach is a mass balance one tier down: bulk inflow of discarded compute-bearing electronics, that is not matched by a similar amount of refurbished devices redistributed through ordinary channels. The difference is the retained fleet.

So, a large-scale repair activity with device outflow below the expected redistribution is a fingerprint of a farm. (The realistic concealer redistributes some as cover. So the signal is the deficit against expectation, not a complete absence). Can be watched via recyclers, bulk marketplaces, and §3.1.1.e's telemetry.

3.1.1.e Fine-grained compute renting (thousands of small contracts).

These are individually invisible, but statistically visible. Marketplaces and payment rails can act as aggregate sensors, even if that is member-blind.

3.1.1.f Dark compute (botnets).

This is a criminal-channel. Its compute is heterogeneous, weak for frontier training, but nonzero for some workloads.

Detectable mostly through the law-enforcement interface, not by the Council's own sensors.

3.1.1.g Non-standard compute (quantum, optical, analog, fluidic).

This is a quiet paradigm development. It evades chip-centric sensors entirely. As a consequence, the sensor portfolio must watch the inputs of exotic compute (cryogenics, photonics, specialty materials supply). So, the dangerous-directions list must also contain the hardware-paradigm entry.

3.1.2 Power and thermal remnants.

Megawatt-months leave grid, cooling, and thermal traces. Some of these can be covered through masking by co-location — compute can be hidden inside legitimately high-consumption facilities (metallurgy, power stations themselves). This raises detection cost, but cannot erase some signatures. Duty cycles, power quality, and cooling type differ between a smelter and a GPU hall. Chips must still arrive (see §3.1). Anomalous consumption deltas are auditable. Masking fits inside the budget, but does not break it.

3.1.3 The instrumented substrate.

Hardware-level attestation for declared facilities is the strongest channel. It converts the deep question into "were all facilities declared?", which §§3.1 and 3.2 answer statistically.

There is a treaty precedent for protecting the data channel itself: the START telemetry-non-encryption clauses.

3.1.4 Financial flows.

Creative accounting is older than film royalties. It should be corroborative only as an information channel, not load-bearing. (Hollywood accounting is the standing proof that audited books can say what their keeper needs.)

Where extra detail is required, there will be business resistance. It can be decreased with confidentiality guarantees and verified destruction after audit (§7.2). If that is not possible, it might not be worth collecting at all.

3.1.5 Talent signals.

These can be aggregate only.

They are wider than just the hiring. Some potential sensors are:

- labor-market aggregates
- compensation surges in narrow skills
- mass talent-search and precursor-search patterns
- the standing-up of internal academies and training pipelines
- the educated-versus-deployed gap — a member training far more specialists than its declared projects can absorb is a signal in itself

The line, drawn in the mandate's ink, is: talent signals are field-level and aggregate. They must never come down to individual tracking — the cargo-not-person principle extends to sensing.

3.1.6 Deployed capability.

Hidden capability that is never used never pays. Its use exposes whether it is actually free, or not: rejected offers to rent compute that is declared as not used light up red lamps. That is why deployment determines each member's public capability floor.

3.1.7 Data flows.

Stocks copy silently (but their statistics determines the level of trust in them). The effects surface downstream in compute and capability, and should be priced accordingly.

3.1.8 Class-F observables.

These belong to three provability grades:

- (i) Measurable: market share, deployment reach — public, recomputable, algorithm-grade.
- (ii) Hard: influence exercised through intermediaries — estimable with error bars, tribunal-grade at best.

- (iii) Unprovable to an algorithm standard: covert political influence — a Developer can always dispute the estimate, and the dispute is unresolvable by mechanism.

The jurisdiction rule is: the mechanism's writ ends where member-recomputability ends.

Only grade-(i) inputs enter the algorithms. Grades (ii)–(iii) inform the adjudication lane.

Release-terms conditioning stays presence-proof-friendly regardless. Obligations of the form "must provide X" are verified by consuming X, and are bounty-funded.

3.1.9 Narrative-consistency auditing (extended version of Flyorov signal).

Raw detection of publication silence is defeated by any modern actor — decoy publication cadence, second-tier output maintained, ghost-authorship... The historical record teaches that concealment comes with active cover stories, not with silence. the Glomar Explorer's manganese-nodule cover (Project AZORIAN), Dimona's "textile plant," GCHQ's decades-long non-publication of public-key cryptography...

The matured detector therefore audits consistency:

- quality-adjusted output (top-venue share falling while volume holds)
- citation-network displacement
- topic drift away from flip-prone directions
- patent-filing shifts and strategic non-filing
- the AZORIAN lesson proper — declared purposes checked against observable economics (specialists doubted the nodule story on its own numbers).

Must be used as an advisory class — an input to H_f, but not a gate.

3.2 The concealer's playbook

Plainly: the known ways to cheat, each with its counter and its residue.

This list must be maintained and changed often, using adversariall approach. Concealers innovate — that is why this playbook is maintained by the Council's standing red team on the same schedule as the channels it attacks. Every new play is entered with its counter and its residue, and an aging playbook is itself a warning sign.

3.2.1 Under-declare.

Must cause conservation audits + supplier attestation.

3.2.2 Shells.

Must cause transitive aggregation + the 50-percent ownership pattern + AML/KYC practice.

3.2.3 Distribute below detection.

The interconnect counter holds today, but is time-degrading. The composition trend (models built of smaller models, connections mostly local) lowers the dense-interconnect requirement year by year.

Consequence: c is re-estimated as architectures evolve (§7.3's living portfolio), and composition advances join efficiency advances in §3.2.6.

3.2.4 Timing games.

Must cause remnant dating + accession pooling.

3.2.5 Sovereign black program / the vertical stack.

The "no state contains the whole chain" counter is time-limited. At least one state is investing at scale toward a complete AI vertical, and the counter weakens as verticals complete.

Meanwhile today's chokepoints are already state-adjacent. Frontier fab capacity is concentrated in Taiwan. The lithography monopoly in the Netherlands with its two critical component suppliers in Germany.

This cuts both ways. Few attestors make the sensor grid cheap while the map holds, and one policy decision moves the map (the Commons' rotation logic anticipated this (*The Commons*, in preparation)).

Design consequences:

- recruit the stack-holders early (the founding-component honesty of the membership companion (the membership companion, in preparation))
- an acceded vertical brings its internal chain into the sensor grid
- an unacceded one sits outside all sensors — carried by the margin, the outer instruments, and W1.

3.2.6 Attacking the sensors.

Efficiency and composition advances shrink or diffuse the physical footprint, degrading §§3.1–3.2 exactly when it matters. This must increase hazard-indexed audit intensity (§7.3), and the honest W3.

3.2.7 Workload offload — the legacy-fleet play.

Cheap recycled fleets (or dark compute) can absorb the auxiliary tier — data processing, evaluation farms, synthetic-data generation, inference serving. This frees declared compute for capability training without changing any observable trace.

This can be amplified by sparse-training advances, detonated if a log→flat flip lands. It is capable of pushing an affiliate-class actor or a dark horse toward decisive advantage, while the throttle brakes exactly the wrong actors.

Counters, in rising importance:

- sensors — the aggregation points of §§3.1.4–3.1.6, the e-waste mass balance (repair inflow with redistribution below expected), plus capability-per-declared-compute anomaly tracking for members and affiliates (the offload's fingerprint: deployed capability rising faster than declared training explains)
- hazard coupling — sparse-training and consumer-fleet-usability advances join the H_f signal set explicitly
- structure — the outside-aware throttle (§4.4).

For the true dark horse, the honest placement stands. Sensors bound this weakly. The endgame instruments and the convoy's combined velocity are the real carriers, as the divergence analysis says (ACoR §2.2) — this play is that residual's concrete mechanism.

3.2.8 Provoked easing.

A braked insider leader covertly passes know-how to an outside actor to manufacture the outside-approach weather, triggering the pre-committed easing and freeing itself — the easing clause turned into a lever, cui bono the insider.

No reliable block or detection exists: know-how copies silently (§3.2.7). Outward leakage is provability grade (ii)–(iii) at best, so the play cannot be gated by mechanism (the jurisdiction rule holds even against this).

What bounds the payoff instead:

- (i) The trigger's size — easing keys on the world-margin estimate at statistical tier with published bars, so the leak must be large enough to create a genuine near-peer outsider. This is an enormous, irreversible self-cost — the provoker births a real competitor it must then outrun, and hands to the endgame instruments a real target.
- (ii) The bundle — easing never arrives alone: the convoy-acceleration list (§4.4) shortens W_int in the same weather, so the provoker's remaining edge diffuses inward faster as the price of its freedom — having already burned it outward, its relative gain erodes from both sides.
- (iii) Cui bono auditing — an outside capability jump whose art resembles a braked member's, timed to that member's benefit, is a narrative-inconsistency input (§3.1.9). (This is advisory, adjudication-lane — must not be a gate.)
- (iv) The violation class — unauthorized outward transfer of frontier know-how enters the sanction table as a named class, adjudication-lane by construction.

The residual: the play is bounded and made expensive, not prevented — filed beside W2, whose soft spot (unobservable know-how) it exploits.

4 The central result: the gap's two purposes and the dial's bounds

Plainly: the safety margin the members keep already contains the allowance for what verification misses — and that allowance has a floor, a ceiling, and a pre-written response for the day several things go wrong at once.

The engagement–defended gap serves two purposes, previously conflated: the concealment allowance *and* the reaction buffer. The honest identity is:

$\mu_{eng} = \mu_{def} \cdot (1 + c) \cdot (1 + r)$

where c is the concealment budget, r is the reaction buffer.

The spec's 2.75/2.0 is an allocation of the 1.375 factor between them (e.g., c = 0.25, r = 0.10), published as such by the Council.

The dial's bounds and near-bound behavior are:

4.1 Lower bound: r > 0 always.

Even a perfect verification (c = 0) must not close the gap — a throttle engaging only at the floor has no reaction time. The buffer is load-bearing on its own.

4.2 Upper bound: c*, where margin collides with joinability

Raising c raises μ_eng multiplicatively. At some c* the required margins collide with ACoR's D* squeeze — margins so demanding that membership stops paying and the coalition thins, which itself weakens the margin.

Beyond c*, safety cannot be bought with margin: it must be bought with verification (lowering c), with the tightening instruments (cadence, audits) — or with the sharing instruments (field elevation, the window, joint-centre access), which shrink the gap itself rather than enlarging the allowance for it. (Capability shared back needs no margin to cover it.)

The design must know its c* and operate visibly below it.

4.3 Statistical honesty near the bounds

c is set at the upper confidence bound of its estimate, not at the point value. The dark-stock indicator couples to it, so the ledger raises c when the residual grows — trust doesn't have to.

4.4 The outside-aware throttle

The throttle's purpose is internal anti-divergence, and it is subordinate to collective survival. The margin monitor therefore carries a world-margin estimate — the best statistical-tier estimate of the leading outside actor's capability, with error bars published. When the estimated outside frontier closes within a published distance of the member frontier, throttle easing is pre-committed. (The convoy is never braked while a stranger accelerates past it.) Entry and exit are by published indicator rule, like the degraded-verification regime below. The two regimes may compose (degraded verification *and* outside approach is the design's worst weather, and its response is pre-written).

And easing is only the brake released — the same weather pre-commits the accelerator - the Council-maintained convoy-acceleration list. This list currently contains:

- the window interval W_int shortened (internal know-how diffuses faster, the whole body speeds up)
- throttle easing extended, even to internal leaders in violation of the internal margin — in this weather they face competition from outside the Council, which disciplines what the throttle otherwise would
- accelerated release and joint-centre surge (the sharing instruments run hot)
- the recruitment arm — pre-listed measures to attract outside developers, aimed with the weather's own asymmetry read correctly: the gaining that outside leaders will likely be less willing to join than before (why cap a lead that is growing?), while those behind them — threatened by the same leading outsider as the members are — become more willing. The arm therefore targets the outside field's second rank first, thinning the leader's future coalition and thickening the convoy (the membership companion's machinery, aimed at the followers, not at the front).

The list enters and exits with the easing rule. Each entry carries its own published trigger within it.

The bundle is also a defense. As easing never arrives without the W_int shortening, an insider tempted to provoke this weather by feeding an outsider pays for its freedom by diffusing its edge inward — see §3.2.8, the treatment of that attack.

4.5 The compounding case, pre-answered: the degraded-verification regime

The dangerous failure is correlated: dark stock rising *while* H_f rises *while* narrative inconsistencies accumulate. Each might be within tolerance, but the combination might not be.

The correct response to this would be a pre-committed regime shift. Entered on a published combination rule — μ_eng tightens, cadence shortens, audit sampling densifies, all automatically — so that no vote is needed in exactly the moment votes are hardest. Exit of the regime should likewise be by a published rule.

In essence, this is the redaction-class discipline applied to an operating mode.

5 The Reference Algorithm

The model in §§3-4 says what verification must buy: a bounded concealment budget and a reaction buffer, published as a composition of the gap. This section says what buys them. It presents a reference example — the algorithm the Council would run continuously to decide whether any member has drifted close enough to decisive advantage to threaten the margin, and if so, by how much, and through which remedies. The full normative specification, complete enough to reimplement from the text alone, is Supplement A (v3.19). A conformance implementation is Supplement B. Follows the argument.

Two kinds of statement run through this section, and the difference matters more than any number in the paper: what the design depends on — change it and the design is a different design — and what the Council sets and revises — the knobs. The prose keeps them apart by saying which is which; §5.9 collects both into one table, because every knob is a capture surface and the list of knobs should be short and public.

5.1 The spine: an individual rule that enforces a coalition rule

Plainly: cap what each member holds, and you have capped what any cabal can hold, without ever having to name the cabal.

Let **T** be the field's total capability score and **f_cap** the size of the covert coalition the design defends against. Define each member's **allowance**:

> A = T / ((1 + μ_eng) · f_cap)

Theorem. If every member's score is at most A, then any f_cap members together hold at most T/(1+μ_eng), so the compliant remainder outweighs them by at least μ_eng.

The proof's consequence is the design's foundation: a per-member rule enforces the coalition rule automatically. The Council doesn't have to detect a conspiracy to be protected against one. Coalitions larger than f_cap are the detection machinery's jurisdiction, not the throttle's. Declared blocs count as one member of bloc size, and undeclared coordination is a named violation.

Three load-bearing consequences follow.

No rank-based test. Throttling "the top k" fails, and the modelling shows how it fails: top-k identity churns from period to period, so a wealthy group rotates through the cap while the weakest stay non-protected. The instrument targets each member's own holdings, not its position in a league table.

The ratio form, for reading. The same rule is stated as a ratio anyone can compute: **d_i = C_i/S_i**, a member's capability against everyone else's together, with an engagement ceiling d* and a defended ceiling d_def. Members leave each other's denominators, which is why the pooling arithmetic works out exactly at both ceilings. The ratio form is parameter-free and publishable; the allowance form is the same statement for machines.

Feasibility and energy. Let **ρ = A/C**, the ceiling's height above the field mean. The mechanism requires **ρ > 1**: at ρ = 1 only a perfectly equal field complies, and the instrument degenerates into a leveller. Beyond that, an energy condition requires that the ceiling leave real room for a lead — parameter changes must publish the resulting ρ and satisfy it, or state, with a sunset date, why safety temporarily overrides. A design that forbids winning is a design no winner joins - here is where that promise is kept in arithmetic.

5.2 MEASURE and EVALUATE: what enters, and what it is trusted for

The Council maintains a set of parameters — compute, data, efficiency, capability, influence holdings, the sharing graph, timing, release compliance. Each parameter carries four things: how it is measured, an audit grade (countable and independently verifiable, or contestable), an intrusiveness score, and the remedies admissible against it. The aggregate intrusiveness is published, so the mandate cost of the current parameter set is explicit and arguable rather than hidden.

Scores are computed as weighted field shares, scaled so the field mean is one. The requirement on that aggregate is severe and deliberate: every member must be able to recompute its own score, and every other member's, from published inputs. Discretion in the measurement is where capture lives. Sophistication belongs in the remedy menu, never in the ruler.

5.2.1 Honesty tiers.

Every input is tagged as 'hard' (countable, independently verifiable), 'statistical' (estimated, with error bars published), or 'trust' (assertion-dependent). Two rules follow, and they are the algorithm's defence against its own instruments:

- The no-single-trust rule. No trust-tier input may alone gate an irreversible rung of the escalation ladder. Irreversible action requires hard- or statistical-tier corroboration.
- The jurisdiction rule. *The mechanism's writ ends where member-recomputability ends.* Only observables that every member can recompute enter these formulas. What cannot be recomputed can be used by the adjudication lane, but not by the model arithmetic.

5.2.2 Capability accountancy.

Capability leaves a member's ledger only by a verified exit — verified transfer into another ledger, or verified retirement within the published depreciation band,

disposal-attested and timing-uncorrelated (§2). An unexplained absence is recorded as a *discrepancy*, not as a lower stock.

There is an asymmetry that is the point:

- The holder's own test runs on accountable capability — verified-present plus the unverified-exit stock — so destruction gains its maker nothing.
- Every field statistic runs on verified-present capability only, so the safety arithmetic runs on physical truth and a genuinely weakened field correctly tightens allowances.

In short: the book binds its keeper; the field reads reality.

The obligation is not permanent: it expires with the band, or releases at any time on verified attestation, or through the estimation lane on documented catastrophic loss. *Release is by evidence or by time, but not by petition or payment.*

5.2.3 Anomaly tracking (advisory only).

The algorithm also maintains, per member, the ratio of demonstrated capability to declared training compute. Capability rising faster than declared compute explains is the fingerprint of workload offload. The signal doesn't gate the throttle — it causes audit sampling within published bounds.

Advisory signals may touch audit intensity, but nothing else — a member approaching a threshold is precisely the party motivated to suppress the signal. That is why nothing irreversible should hang on one.

5.3 TEST, OFFER, CHOOSE: what the member owes and how it pays

5.3.1 TEST.

Because reducing one member's score lowers the field total and therefore every allowance, the requirement is a fixed point. It has a closed form:

> $R_i = \max(0, C_i - A_i)$

A member within allowance owes nothing and is untouched — which, in the ordinary state of a healthy field, is every member.

5.3.2 OFFER.

This is an unit-partitioned menu.

A member reduces R_i by earning *compliance credit*. The excess in each parameter is partitioned into units; each unit may take at most one remedy — every credit must trace to a distinct unit of capability actually reduced, relocated, or defused, and different units may take different remedies.

Three remedies are available. All are under the conservative-credit rule: credit never exceeds capability actually removed or defused.

1. **Throttle** — slows further acquisition, but doesn't touch holdings.
2. **Open** — non-rival capability shared at a defusal level. It is credited by how much the field actually gains. An item the rest already holds lifts nobody and credits nearly nothing, while a unique item credits in full. A structural guard forbids routing the most dangerous capabilities to the bluntest opening levels.

3. **Lease** — relocates rival holdings, lowering the leader and raising the field at once, so it earns more than one-sided credit and strictly less than double.

Leasing may be charged for. It is done to rebalance capabilities, not to rob or punish the leader. Leadership is a product of hard work, and is commendable, not detrimental - it should be rewarded, not punished. Any losses that a leader can accrue due to the rebalancing are an undesired side effect that must be avoided at any cost that does not compromise the rebalancing of capabilities. So, compensating the leader is important.

5.3.3 CHOOSE.

The Council publishes the constraint and the credit ceilings, and every member computes its own least-inconvenient combination.

As with a leader who leases holdings, the goal is to balance capabilities, NOT to punish the members. That is why each member who is target of rebalancing must have the choice how to achieve it in the most painless way. The cost of each remedy to a given member is private — what is cheap for one is existential for another — and a Council that chose for members would be choosing with worse information and more discretion. That is why the assignment of the computation to the member is a part of the algorithm.

5.3.4 The no-choice default.

Silence throttles, proportionately, on two published axes — the uncovered remainder and the headroom already consumed — in three pre-committed zones:

- a cure clock with a throttle sized to the gap
- a full acquisition freeze with the case docketed
- immediate escalation with clocks compressed

Every input is on the public dashboard. The default has no discretionary knob, because the moment a member goes silent is exactly the moment discretion would be contested.

5.4 The clock, the ladder, and why a rulebook is the milder path

5.4.1 The clock.

Timeframes are computed, not negotiated. A base clock is used, by correction type and size — know-how is fast, leasing is medium-speed, divestment is slow — multiplied by an urgency factor that shortens as headroom is consumed.

Where the compressed clock falls below what a correction can physically achieve, a longer clock may be purchased by a bridge: interim measures that hold the member's ratio flat for the bridge's whole duration. *Time is granted against risk held flat; prolongation is purchased, not pleaded.*

5.4.2 The insufficiency state.

A member may be far enough ahead that no combination of remedies reaches R_i — because throttling acquisition cannot un-acquire holdings. This is a named state, not an embarrassment, and it has an ordered ladder. Each rung on it is used only if the prior does not suffice:

1. **Mandatory leasing** — where leasing can close the gap it may be required rather than offered, since it relocates holdings without confiscating them.
2. **Divestment** — the member sheds holdings. Stronger both as effect and as side effects.
3. **Field elevation** — raise the field instead of lowering the leader: accelerated release, joint-centre access, diffusion at zero window. The compute channel can be frozen but not un-diverged. Only knowledge, being non-rival, restores parity by being shared back.
4. **Referral beyond the Council** — where internal remedies cannot restore the margin, the matter goes to the standing security order that predates and outranks the Council. The Council is explicitly and intentionally not the top of the enforcement stack. Its terminal function is to hand that order a documented, quantified case while a case still exists to hand.

5.4.3 Why the decision makers of the leader should prefer the ladder.

The unpleasant rungs look like constraints imposed on the strong. They are the opposite, and the argument deserves stating plainly — it is the one that makes the ladder acceptable to the party most able to refuse it.

Without a rulebook, a member's escalation is detected anyway — by rivals, intelligence services, the market. What follows is a reaction that is unbounded, unscheduled, and shaped by whoever moves first: frightened parties acting on partial information, some of them frightened of you, some reporting to masters who are, and some holding a vested interest in your suppression and merely waiting for an occasion.

The alternative to accepting a graduated, published, bounded ladder is not the absence of consequences. It is exposure to drastic ones, chosen by others, at a moment of their choosing. The ladder is insurance against the panic of peers or / and parties with serious enforcement capabilities. It is cheapest to buy before it is needed.

5.5 Self-arming, and why nothing warns the watchers

The algorithm is self-arming. No separate alarm is required to switch the instrument on. Run continuously, it returns $R_i = 0$ for every member whenever the field is in the regime where returns to compute are gentle and nobody diverges. It returns $R_i > 0$ exactly when a member diverges, which is the dangerous regime manifesting. Regime detection is implicit in the divergence itself: there is no separate detector to suppress, spoof, or argue about.

An elasticity heuristic — how sharply capability responds to compute — may be kept as an *optional early warning*, but it is advisory and must never gate the throttle, for the reason given above: the member approaching a flip is the member most motivated to suppress the reading. There is also a credibility budget to respect: an advisory that cries wolf trains its readers to ignore it, so its threshold is tuned for signal, not for maximal caution. The bias toward early ringing belongs to *arming*, which is now automatic and free, not to *warning*, which is not.

5.6 The hidden flip: a named, unsolved attack

One attack defeats both the always-on algorithm and any early-warning heuristic, and the honest course is to name it rather than to reassure.

Imagine a member that:

- discovers privately that the conversion from compute to capability has changed — the flip — at small scale
- tests it below the tracking threshold, in the measurement noise, so its score stays unremarkable
- applies it at scale, whereupon the compute it *already holds*, previously safe, suddenly yields containment-breaking capability.

The capability of this member steps from within-allowance to decisive advantage between two evaluations. Both instruments watch for a rising signal. However, this attack produces a step from a member who looked safe an instant earlier.

The attack requires three conditions at once — private discovery, sub-threshold testing, and holding the advance unused until scaled — and removing any one defeats it. That yields three partial defences, none complete:

- **tracking nearer the scale at which a flip first becomes testable** (bounded, because the attacker tests in the noise deliberately)
- **Council-first discovery**, where common laboratories aimed at the flip-prone directions make such advances public, so that every member's compute re-prices simultaneously and a private step becomes a shared tide — the strongest of the three
- **one-cycle containment**, since activation at scale produces a step the next evaluation sees, bounding the attacker's window to a single cycle.

Whether one cycle suffices depends on how fast irreversible action can be taken. This is the endgame's problem, not this one's.

Two further measures tighten the bound rather than closing it:

- The evaluation cadence is hazard-indexed: it shortens as a published, member-blind, field-level hazard index rises, so the attacker's one-cycle window narrows exactly when a flip is near — at no intrusiveness cost, because the index estimates how close the *field* is to a flip, not which member.
- Audit sampling rides the same index, for the reason §4 gives: flips attack sensors first, so the audits thicken precisely when the sensors are most doubted.

Stated plainly: the hidden-flip attack collapses the reaction window to a single evaluation cycle, and this design does not prevent it. It bounds the payoff. The residual is flagged as a named task for the Council — *guarantee Council-first discovery of flip-class breakthroughs, and guarantee that nothing decisive can be done irreversibly within one cycle* — and it is stated here as an open problem because a sharp flag is worth more than a vague reassurance. It is also the first of this paper's falsifiers (§8).

5.7 Two operating regimes, pre-written

Two situations would otherwise demand votes at exactly the moment votes are hardest to obtain, so both are pre-committed and the algorithm inherits them as written:

- the outside-aware throttle (§4.4), which releases the brake and engages the acceleration list by indicator rule when a non-member closes on the frontier

- the degraded-verification regime (§4.5), which tightens the margin, shortens the cadence, and densifies sampling when the three warning indicators breach together.

The two compose. Their combination is the design's worst weather, and the response to them is pre-written.

5.8 What this algorithm is not

It is a reference example, not a constitution. Its numbers are first calibrations and are expected to move. The Council's opening task is to locate the horizon past which the remedy menu can no longer reach, and to set the engagement margin below it.

A control mechanism that requires calibration is not weak — it is what every working governor is. What is not expected to move is the short list of structural results:

- individual allowance is coalition safety
- target holdings, not rank
- act on acquisition, not on holdings, save by the member's own election
- credit conservatively
- one remedy per unit
- let the member choose on its private costs
- engage early enough that the menu can still reach
- end the ladder outside the Council.

5.9 What is fixed and what is tuned

A reader deciding whether this is a discretionary instrument needs one page of information: the list of what the design depends on, and the list of knobs the Council holds. Every knob is a capture surface, which is why the second column is short and public.

Fixed by the design (change it and the design changes)	**Set and revised by the Council (the knobs)**
Individual allowance implies coalition safety (the spine theorem)	The defended coalition size f_cap, via β, and its rounding rule
Target each member's holdings, never its rank	The engagement margin μ_eng, and the published c/r split inside it
Strict feasibility: the ceiling above the field mean (ρ > 1); an energy condition for lead room	The energy margin ρ_min; the floor fraction φ; the sunset coefficient ε; the count alarm line n_min
Every score recomputable by every member from published inputs	The parameter set, weights, audit grades, and the published aggregate intrusiveness
Honesty tiers; no single trust-tier input gates an irreversible step; the writ ends where recomputability ends	Which observables enter each tier
Capability leaves a ledger only by verified exit; the book binds its keeper, the field reads reality	The depreciation band schedule
One remedy per unit; conservative credit; the member computes its own combination on private costs	The unit granularity; defusal levels and their credit factors; the lease credit
The no-choice default's three zones; silence throttles proportionately	The zone thresholds
Clocks computed, never negotiated; prolongation purchased by a bridge	The base-clock table and the urgency multiplier
The escalation ladder's order, ending outside the Council	—

Fixed by the design (change it and the design changes)	Set and revised by the Council (the knobs)
Self-arming; advisory signals touch audit intensity and nothing else	The advisory heuristic's threshold
Hazard-indexed cadence and audit intensity, field-level and member-blind	The baseline cadence and the hazard mapping
The two operating regimes enter and exit by rule, not by vote	The indicator thresholds and the combination rule
Changes by Byzantine-or-higher majority, never unanimity; listing guarantees the lister's advantage	The current values of everything in this column

6 Hardware supply neutrality

Plainly: a member that cannot buy machines does not need to be out-competed. It can be starved.

Verification bounds what a member can hide. It does nothing about what a member can be denied, and denial is the cheaper attack.

The chokepoints of AI hardware — chip design, leading-edge fabrication, lithography, high-bandwidth memory, advanced packaging, datacentre capacity, power, cloud allocation, raw materials — are held in few hands, and most of those hands might not be Council's members. Two structural facts organise everything that follows:

- The binding link rotates: the bottleneck migrates as technology and demand move, so neutrality obligations must attach to whichever link currently binds rather than to a link named once in a treaty.
- Most chokepoint holders might be non-members. Members control some links, member *states* control others through export law, and pure suppliers hold the rest.

The instruments divide accordingly.

Starvation is done in recognisable ways, and each has an instrument.

Allocation bias — queue position, quotas, delivery schedules — is subtle and deniable. Pricing discrimination is blunter. Information exploitation is the insider-trading analogue: openness reveals a rival's roadmap and a supplier's affiliated laboratory front-runs the demand. Export-control capture weaponises a member state's jurisdiction against co-members. Beyond these sit standards capture, toolchain lock-in, withdrawal of maintenance and spare parts, firmware implants, cloud deplatforming, discrimination in power contracts, financial denial, and administrative friction — customs and licensing delays that show up in the statistics as the asymmetric aging of like requests.

The taxonomy is expected to grow. Maintaining it is part of the living chokepoint map, which is a founding deliverable of the Council and a concrete early task of obvious value.

The architecture has five layers.

6.1 Obligations

For member-controlled links:

There must be published, non-discriminatory allocation terms, hardened by the essential-facilities logic of competition law and the common-carrier model from telecommunications.

Two admissions keep this from being wishful. Controllers of unique facilities understand their value precisely and will not part with control, so the realistic instrument is not doctrine imposed but negotiated assurance agreements — incomplete in some cases, better than the present in all.

The borrowed doctrines need adaptation, not direct adoption: the Council's purposes will make "fair and non-discriminatory" mean something subtly different from the standards-body original, and saying so now is both honesty and evidence of planning.

Against information exploitation, there must be a ladder rather than a demand, because the strong form will be refused by vertically integrated members. This ladder might include:

- structural separation at the top
- functional barriers below it
- escrowed allocation algorithms below that
- and (the rung that lowers resistance most) a performance standard: demonstrate statistical neutrality in outcomes and your internal structure is your own business.

Regulating outcomes rather than organisation charts converts an existential demand into an auditing one. For state-controlled links, a most-favoured-member undertaking on export licensing for Council-verified uses can be used, where the design holds a good card.

Transfers on instrumented substrate are precisely the verified-end-use case that export-control regimes claim to exist for.

6.2 Detection

Discrimination hides in noise, and statistics finds it. Standardised procurement and allocation reporting makes the comparisons computable. The telltale is the differential remnant — a member's orders aging unfilled while like orders fill.

Burden-shifting after a statistical deviation spares the victim the impossible task of proving intent.

One precedent must be checked rather than assumed: the fair-lending and employment models rest on millions of decisions, while allocation among dozens of members is a small-sample problem where those methods weaken. What transfers is the burden-shifting philosophy, not the econometrics — and the small sample cuts the other way too. Decisions too few for statistics are few enough for exhaustive review: the audit organ can read every allocation above a size floor, which no fair-lending regulator could dream of.

6.3 Relief

An earlier version of this design specified a standing compute reserve. It was retired, for a structural reason worth knowing: a reserve is insurance whose fund evaporates in the correlated shortage it insures, since compute is rival and scarce exactly when needed.

What replaced it works on the attack rather than the shortage — notice-and-standstill obligations, a credit facility that buys on the spot market and bills the denier, adjudication-indexed forgiveness or penalty, and an option book of pre-priced calls on capacity with a bound reclaim latency, where the service level is the teeth and an unhonoured exercise is a named violation.

6.4 Engagement

Non-member chokepoints are reached by bargain, not by rule. The observer-supplier class offers neutrality covenants in exchange for demand aggregation — the body procures as a bloc, giving suppliers a counterparty too valuable to discriminate against, and small members the bloc's deal terms.

Diversification co-funding invests in second sources, shrinking the chokepoint itself. The spillover reaches every buyer downstream, down to the ordinary purchaser of memory. (This has a stated limit: a true monopolist does not need stable demand and may decline the bargain.)

6.5 Deterrence of force

The record of stopping rival programmes by force is long and documented: strikes on facilities, sabotage of equipment, cyber- or physical attacks on infrastructure, assassination of programme scientists. Corporate participants tend to file this under the state's business. The sentence they need before they experience it is: *as AI capability approaches decisive relevance, frontier laboratories inherit the threat model of nuclear facilities*. Their buildings, their datacentres, their people, and their supply chains - all these do.

The Council's contribution is threefold and already in the design:

- prohibition among members as a named violation at the top of the ladder
- devaluation, since where advances diffuse on a bounded window, destroying a rival's copy buys little
- collective attribution and response, which converts a lone laboratory's exposure into the hardest currency in the membership dividend.

One economic note, because it answers the objection that neutrality cannot conjure supply:

Neutrality makes the queue fair — it does not lengthen the queue. Three instruments ease the shortage itself — demand aggregation, diversification co-funding, and, cheapest of all, utilisation. Pooled procurement, preemptible spot markets, and demand smoothing raise the yield of compute that already exists. And scarcity quietly argues for the rest of the design: members burning scarce compute to re-derive what a peer already holds is pure waste, so shortage makes openness economical, and the sharing that scarcity purchases compounds into the trust the body runs on.

7 The sensor system's own constitution

Plainly: the watchers are designed to be useless to spies, and priced and proceduralized where uselessness cannot be engineered — auditable by the watched, and upgraded forever.

7.1 The mass-balance audit and the dark stock

> production − attrition − declared = the published residual the field cannot account for

It must be capability-weighted per §3.1.1. The provenance for it is the IAEA material-balance lineage.

It is the empirical floor under c.

7.2 Minimal-emission sensor design

This is the anti-spying guarantee. The verification system must be verifiably useless for espionage, or members will refuse it and will be right to.

The precedent is nuclear disarmament's information barriers: instruments engineered to emit pass/fail and nothing else. The rules to design them are:

- every channel emits the minimum sufficient statistic (attestation, aggregate threshold-crossing — never raw feeds)
- raw data never leaves the instrument
- sensor designs are member-auditable (open hardware for the watchers of the watchers)
- audit-collected detail is destroyed on schedule with verified destruction
- everything inherits the Commons' partitioned transparency and its outcomes-not-structure principle (*The Commons*, in preparation).

What the sensor cannot reveal, its theft cannot either.

7.3 The living portfolio

Channels decay — architectures evolve, chains diversify, concealers learn. The portfolio must be red-teamed and upgraded on a standing schedule. c and the c* estimate must be reviewed on the same cadence. Hazard-indexed audit intensity must ride beside hazard-indexed cadence (H_f up → sampling up), because flips attack sensors first.

7.4 The tiers

Every MEASURE input carries its tag —

- 'hard' (frontier-scale acquisitions, deployed floors, grade-(i) class F, conditioning compliance)
- statistical (stocks via mass balance, data acquisitions, timing remnants, corporate capex, talent aggregates)
- 'trust' (sovereign budgets, grades (ii)–(iii) influence, silent data, *undeployed efficiency* (the last is the soft spot of the tiers).

No single trust-tier input may gate an irreversible action. The mechanism's writ ends where member-recomputability ends.

7.5 Where minimality fails — the residual-intrusiveness toolkit

Some collection is irreducibly revealing: deep research assessments, audit-collected raw detail, anything whose sufficient statistic is itself sensitive. For these cases §7.2's guarantee cannot be engineered into the instrument.

The answer is layered procedure instead, each layer precedented:

- (a) managed access — the inspected party controls the path and may shroud what is out of scope (the CWC's managed-access regime: verification proceeds, secrets stay covered)
- (b) cleared mixed teams with split knowledge — small, vetted, mixed-nationality inspection teams, no single inspector or nationality seeing the whole, personnel rotated on schedule (IAEA/OPCW practice)
- (c) instrument-side hygiene — air-gapped analysis, one-way data flow into the instrument, and §7.2's verified destruction on schedule for everything collected
- (d) pre-agreed breach liability — espionage through the verification channel is a named violation class with compensation terms agreed at accession, so the fear has a price tag and the victim a remedy, before anything happens
- (e) reciprocity — the audit chain is itself auditable: members inspect the inspectorate's handling, and watchers of the watchers are a standing right, not a concession
- (f) the elective trade [the design's own move] — intrusive assessment is never compelled: a member may decline an irreducibly revealing assessment and accept conservative scoring in its place — the declined parameter enters at its upper bound, paying in margin what the member refuses in access.

Espionage fear is thereby priced, and not forced.

Honest residual: the toolkit reduces the exposure and prices what remains. It does not make the irreducible cases minimal — W5 carries that, and the record of years remains the only full cure.

7.6 The countermeasure literature, anchored

The system's instruments stand on published lines of work:

- zero-knowledge proof of compliance (the Goldwasser–Micali–Rackoff line)
- hardware-anchored verification of large training runs (the "what does it take to catch a Chinchilla" line)
- the compute-governance survey literature
- the accumulated safeguards experience of the nuclear regime — its compromises acknowledged instead of hidden: invitees will know them anyway.

8 Weaknesses, stated — and what would refute our paper

[FLAG for the final pass: W6 and W7 partially repeat §§3-4; compress once the whole paper exists.]

8.1 Weaknesses

W1 — The sovereign residual, now time-qualified.

The black-program case is bounded by chain internationality. That however is eroding as vertical stacks complete. The bound weakens on a visible schedule.

The design's answer is recruitment of the stack-holders and the honest carry of what stays outside. However, this decreases the danger, but cannot eliminate it completely.

W2 — The efficiency soft spot.

Undeployed capability-per-unit-resource remains the least observable, most dangerous quantity. Consequences are bounded by this verification model (by hosted evals, bounties, one-cycle, hazard indexing), but complete observation cannot be achieved.

W3 — Sensor decay under flips (now including composition flips).

The meter confesses (dark-stock error bars widen visibly), but does not cure.

W4 — c's own estimability.

It is partly trust-laden. Must be set conservatively, published with derivation and stress-tested like D*. The ledger-coupling of §5.3 is its counterweight.

W5 — The sensor system as espionage fear.

Priced and answered by §§7.2 and 6.5. The residual is that guarantees reduce the fear, but only the record of years retires it.

W6 — Single control over critical supplies.

Most frontier compute is manufactured in a single country by a single company. All frontier-class production machines are manufactured in a single country by a single company. These machines' two key components are manufactured each in a single country — the same one — by a single manufacturer.

Any of these de facto monopolies could use its position to influence AI development toward permitting a decisive advantage — or toward other goals whose pursuit produces that as a side effect.

This weakness is strongest, and therefore most attractive as a tool, in a supply crunch. We are in one now, and it is not expected to resolve soon. §4.5 priced this map's sensing side (few attestors, cheap grid, while the map holds). This entry names its influence side.

Partial answers, all standing elsewhere, can be:

- recruit the stack-holders early (the membership companion's founding-component rule, in preparation)
- the Commons' rotation logic
- peacetime and accession contracting
- supply diversification as a standing Council aim.

This danger cannot be easily eliminated. One policy decision — or one boardroom — can move the map.

W7 — Supplier obligations in a crunch, and the exit-incentive bound on member sanctions.

In a crunch, suppliers are hard to attract into obligations and hard to punish — especially where their production carries strategic importance beyond AI. Compulsion therefore reaches the buyers — members, through dark receipt — but member-side sanctioning has its own ceiling. If overdone, it creates an incentive to leave the Council, and a Developer chasing decisive advantage may be seeking exactly that excuse.

The liability calibration is thus a named design constraint, not an implementation detail: proportionality and the materiality threshold (§3.1.3) keep member sanctions below the exit-rational line, and the sanction table inherits this bound explicitly. The attraction side rests on the adverse-selection covenant (§3.1.3), which holds only as well as accession-time contracting made it — it is stated, not assumed.

8.2 What would refute this model

The design makes four claims that evidence could break, and it should be held to them.

1. *A channel with a residue the model cannot price.* If a concealment route is found whose hidden fraction exceeds the tolerated residual and for which no counter exists at any intrusiveness the Council can bear, the budget framing fails for that route and the margin must be set as if the route were fully dark.
2. *The hidden flip realised.* If a member steps from within-allowance to decisive capability between two evaluations, the one-cycle bound of §5.6 was too loose, and the design's containment claim is false as stated.
3. *Declared efficiency failing to bound.* The consistency test of §2 assumes that declared compute, declared efficiency, and demonstrated capability can be reconciled within published tolerances. If real declarations cannot be reconciled without discretion, the recomputability rule of §5.2 cannot be met and the arithmetic loses its member-blind character.
4. *The throttle's numbers failing to transfer.* The margin throttle holds in the models of §9. If it fails to hold under a materially different convertibility model — one calibrated on data this design did not have — the structural claim survives but the calibration does not, and the Council's opening task in §5.8 becomes the whole task.

9 The models behind the numbers

The numbers in this paper come from models that can be run, and the models were built by failing. Supplement C carries the eight models behind this paper, with their development trail. This page states what the trail found.

9.1 The master variable is the convertibility regime

The first model contained no instruments at all — a bare market for compute among developers of different means — and it was enough to show that scarcity is not what decides between a convoy and a singleton. What decides is how capability responds to compute. Where returns are logarithmic, the field stays plural whatever the distribution of means; where returns turn linear, the leader diverges from any starting position. An efficiency breakthrough is the event that can flip the regime from the first to the second. Everything the Council builds is a defence against that flip.

9.2 What failed, and why the failures are worth keeping

A compute floor — guaranteed shares to the weakest — fails at every level tested. It prevents starvation - but only that. It bounds nobody's divergence, because a floor acts on the bottom of the field while the danger grows from the top.

Anti-starvation and anti-divergence are different problems, and the reserve that an earlier version of this design specified was an answer to the wrong one. A rank-based

throttle — cap the top k — fails more instructively: the identity of the top k churns from period to period, so a wealthy group rotates through the cap while the members it was meant to protect are never protected. The failure is the reason §5.1 targets holdings and not rank.

9.3 What held

A throttle that is:

- *individual* (each member against its own allowance)
- *continuous* (no thresholds to game)
- *margin-anchored* (the allowance derived from the margin the design requires rather than set by hand)
- *engaged early* (before the remedy menu can no longer reach)

Such a throttle holds the compliant remainder's margin at or above its target in the worst case tested, and is dormant — returns nothing to do — in the logarithmic regime where nothing is diverging. That dormancy is the self-arming property of §5.5, observed first experimentally.

9.4 Two adjacent results

A window on know-how diffusion holds the logarithmic regime at any width and does nothing against a linear compute spiral — the two channels are different, and an instrument on the wrong one is inert; the window models accompany *The Bounded Window* (No. 5). And in a search-race model of common laboratories, plural internal centres reach a discovery first more often than a single monolithic one, which is the modelling basis for the Council-first-discovery defence in §5.6; that model accompanies *The Commons* (No. 6).

9.5 Conformance

The reference implementation in Supplement B executes the algorithm of §5 against specification v3.19. It passes its full assertion suite — 2,360 checks covering the allowance arithmetic, the credit rules, the no-choice default, the clock, the ladder, and the two operating regimes. A reader who doubts a rule in §5 can run it.

10 Related work

Verification against a capable adversary is not new; what is new is the vanishing of the remnants it relied on. The lineage this paper stands on is therefore mostly outside AI.

10.1 Nuclear and chemical verification

The IAEA's comprehensive safeguards (INFCIRC/153) and the Additional Protocol (INFCIRC/540) built material-balance accounting, declared-facility inspection, and environmental sampling into a regime. Its logic is the budget logic of §1: bound the undeclared and never assume that the declared is complete.

The Chemical Weapons Convention's challenge inspections and the CTBT's on-site inspection regime are the precedents for §7's intrusiveness schedule — inspection as a scheduled, bounded, published intrusion rather than an ambush.

Baker (2023) surveys what nuclear arms-control verification can and cannot teach an AI treaty. This paper takes his taxonomy as the starting inventory and asks what survives when the observables shrink.

10.2 Compute governance

Shavit (2023) posed the question that §5 answers at the council scale — what it would take to verify rules on advanced AI development, chip by chip — and proposed hardware logging and sampling as the instrument.

Sastry et al. (2024) map compute as a lever for governance in general.

Heim and Koessler (2024) analyse training-compute thresholds as regulatory triggers, the classical framing that §1 argues is being outrun.

Aarne, Fist and Withers (2024) and the hardware-enabled-mechanisms line describe the instrumented substrate that §3.1.3 treats as one channel among nine rather than as the regime.

Wasil et al. (2024) collect verification methods for international AI agreements.

This paper's contribution to that list is the pricing of each method's blindness and the rule for combining them.

10.3 Verification of restrictions on research

Scher (2026) asks whether restrictions on frontier AI *research* — not compute — can be verified at all, and catalogues twenty-eight candidate mechanisms, from whistleblower programmes and code review to search warrants and ordinary intelligence collection, while noting that AI systems conducting research are copyable and far harder to detect than the humans they replace. Wasil et al. (2025) organise verification into six layers of redundancy and ask how mechanisms combine. Peigné et al. (2026) argue that zero-knowledge verification of frontier training is possible in principle, with an architecture that would strengthen the instrumented-substrate channel of §3.1.3 if it matures. This paper's contribution to that line is not another mechanism but the rule for combining them: price each channel's blindness, combine until the residual is below what the margin tolerates, and state the residual as a number.

10.4 Patterns borrowed from other regimes

The athlete biological passport (WADA, 2009) is the model for §3.1.9's narrative-consistency auditing and §2's consistency test: no single reading convicts — a longitudinal profile that fails to cohere does.

The wash-sale rule of tax law is the model for the capability-accountancy asymmetry of §5.2 — a disposal that is timing-correlated with a benefit is presumed round-trip.

Fair-lending and employment-discrimination statistics supply §6's burden-shifting philosophy, with the small-sample caveat and the exhaustive-review answer stated there.

Common-carrier and essential-facilities doctrine supply §6's obligation layer, adapted here rather than adopted directly.

10.5 The head paper and its companions

A Council of Rivals (The Convoy Papers No. 3) states the design this paper serves and the sanction table its ladder references.

Physical Prices of Knowledge (The Convoy Papers No. 1) and *Why the Winner Starves* (The Convoy Papers No. 2) supply the two results the margin arithmetic assumes.

The Commons, the Bounded Window, the membership companion, and the Council's organs are in preparation. Where this paper refers to them by name, the reference is to a forthcoming paper of the same series.

All citations in this section were selected by the author and verified against the sources before publication.

The consolidation, in one paragraph

Verification is run as a published budget: a Council-maintained portfolio of channels watches stocks, flows, timing, and terms. Every input carries its honesty tier, and nothing trust-graded gates an irreversible act alone. The margin the members hold is sized to cover both what verification misses (c, ledger-coupled, geographically structured) and the time reaction needs (r). The ledger's unexplained residual — the dark stock — is published as a standing meter. Sensors emit only what verification needs and nothing espionage could use. Audits and cadence tighten automatically as flip hazard rises. The whole apparatus eases by pre-commitment if the outside world's frontier closes in. A correlated degradation of the indicators triggers a pre-written tightening with no vote required. Where the budget cannot stretch — the completed foreign vertical, the true dark horse — the model says so, and points to the instruments that do carry those cases: recruitment, the margin, referral beyond the Council, the endgame.

References

- Aarne, O., Fist, T. & Withers, C. (2024). *Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing.* Center for a New American Security.
- Baker, M. (2023). Nuclear Arms Control Verification and Lessons for AI Treaties. arXiv:2304.04123.
- CTBTO Preparatory Commission (2026). *On-Site Inspection.* https://www.ctbto.org/our-work/on-site-inspection (accessed September 2026).
- Heim, L. & Koessler, L. (2024). Training Compute Thresholds: Features and Functions in AI Regulation. arXiv:2405.10799.
- IAEA (1972). *The Structure and Content of Agreements Between the Agency and States Required in Connection with the Treaty on the Non-Proliferation of Nuclear Weapons.* INFCIRC/153.
- IAEA (1997). *Model Protocol Additional to the Agreement(s) Between State(s) and the IAEA for the Application of Safeguards.* INFCIRC/540.
- OPCW (2026). *OPCW Basics — monitoring compliance* (page and embedded video). https://www.opcw.org/about-us/opcw-basics (accessed September 2026).
- Peigné, P. et al. (2026). Zero knowledge verification for frontier AI training is possible. arXiv:2606.05433.

- Sastry, G. et al. (2024). Computing Power and the Governance of Artificial Intelligence. arXiv:2402.08797.
- Scher, A. (2026). Verifying Restrictions on Frontier AI Research. Second Workshop on Technical AI Governance Research (TAIGR) at ICML 2026. arXiv:2606.28694.
- Shavit, Y. (2023). What does it take to catch a Chinchilla? Verifying Rules on Advanced AI Development using Compute Monitoring. arXiv:2303.11341.
- Stremsky, A. (2026a). *Physical Prices of Knowledge.* The Convoy Papers, No. 1. Zenodo. 10.5281/zenodo.21710817.
- Stremsky, A. (2026b). *Why the Winner Starves.* The Convoy Papers, No. 2. Zenodo. 10.5281/zenodo.21710084.
- Stremsky, A. (2026c). *Preventing a Singleton in AI Development: A Council of Rivals.* The Convoy Papers, No. 3. Zenodo. 10.5281/zenodo.22019542.
- World Anti-Doping Agency (2009). *Athlete Biological Passport Operating Guidelines.*
- Wasil, A. et al. (2024). Verification Methods for International AI Agreements. arXiv:2408.16074.
- Wasil, A. et al. (2025). Verifying International Agreements on AI: Six Layers of Verification. arXiv:2507.15916.