Preventing a Singleton in AI Development: A Council of Rivals

Alex Stremsky· The Convoy Papers, No. 3 · v3.1 · 2026-09-04 · CC-BY 4.0

Version-agnostic DOI: 10.5281/zenodo.22019542. This version (v3) adds related work on the moratorium and empirical strands, the severed-dependency argument in §2.1, a paragraph on the two-sided effect of a large efficiency flip, names the Continuity Limit, and corrects the reference list. v3.1 adds the acknowledgment of AI collaboration.

Keywords: AI governance; AI safety; singleton; decisive strategic advantage; AI race dynamics; mechanism design; game theory; institutional design; international coordination; verification; compute governance; frontier AI

Abstract. The race to frontier AI selects for exactly the holders one least wants a decisive advantage to have: the actors most willing to cut safety and to value absolute power. Two terminal risks follow — a singleton, and the desperation war of those about to lose to one. This paper proposes a structure that avoids both without stopping development: a council of rivals, in which no member can gain a decisive advantage and every member is better off inside than out. Its instruments are stated as engineering: a bounded release window whose width is held under half the measured tolerable lead; a fault-tolerance margin under which compliant members always outweigh any possible defector coalition more than two to one; verification run as a published budget; a frontrunner compensated rather than throttled; and pre-committed endgames, with accelerating diffusion as the default that needs no vote. Claims are tiered by evidential status and paired with falsifiers. The mechanisms are built on precedents that already worked at scale.

Preface

This is a proposal that argues for a body, able to prevent the worst outcomes of the frontier AI development, while preserving its freedom, stimuli and benefits.

Includes possibly the first toolkit designed specifically for joint, controlled development of frontier science — which is the standing occupation of a civilization that runs on knowledge.

Mechanism-level detail is developed in a series of companion papers, the Convoy Papers. Two are published — *The Physical Prices of Knowledge* and *Why the Winner Starves* [Stremsky 2026a; 2026b] — and further numbers are in preparation; references below name those by subject.

If a body rises on this design, it will use and improve these tools. If it rises on another design, it will borrow from them.

If no such body rises, the tools will wait — their task will be faced again.

Contents

(Sub-subsections are numbered in the text; a live table of contents can be inserted from the heading styles in the DOCX.)

1 Introduction

Plainly: one race, two ways to lose everything.

The development of AI is a race whose leader may, at some point, hold a *decisive advantage*: the capability to prevent any other actor from catching up, maybe not only in AI development. An actor in that position — a *singleton* [Bostrom 2006, 2014] — would have the future of humankind within its discretion.

This brief takes no view on when, or whether, that point arrives. It prices what happens

if it does, and what affects the price.

The argument is structural: nothing in it depends on the current state of the race and it applies equally to every candidate.

We note that the AI race carries not one, but two terminal risks:

1. The singleton — the familiar risk: the winner converts his lead into command over everyone — possibly over everything, and forever.

2. The *desperation war* — its often ignored mirror: those behind try to prevent the creation of a singleton at any cost, even a devastating one.

These are two sides of the same risk. What can bring a decisive advantage to some sharpens the despair and forces the hand of everyone else. Preventing a decisive advantage prevents both risks.

The function of the structure this brief proposes is therefore **to hold the race's contestants within a bounded distance of one another**. Development efforts must continue to give more to those who exercise them. However, the fastest must not be able to convert a lead into command, so that those behind are not tempted to convert despair into destruction.

This structure has no ambitions beyond this: AI able to yield decisive advantage is an immensely powerful technology, but it is still just a technology — coordination on its dangers is all it will do.

Two prior results, proved in companion work, are load-bearing here:

1. A leader who defends its lead by suppressing or starving those below destroys the inflow its own growth depends on — the dominance and suppression backfire [2026b, P3–P7].

2. The elementary operations of knowledge have physical prices, so claims about what strategies cost are based on the laws of physics, not on a preference [2026a, §3].

2 Analysis of the risks

2.1 The singleton

This is the race risk known for over twenty years now, and everyone thinks about it.

The difference in kind. *Plainly: every tyranny before this one could be outlasted; a singleton could not.*

History has seen many tyrannies, and all of them fell, or could have, through one mechanism: the tyrant needed the tyrannized as hands, soldiers, administrators — and somewhere he erred. A bad order, a misaligned structure, a security service that declined to shoot, an economy crumbling from work done badly on purpose. Refusal was the only real check on despotism, and it required both the need and the mistake. A singleton removes both. Automation removes the need; but the deeper removal is the intelligence advantage, which removes the mistakes — a ruler who does not err gives the ruled no opening, and humans have always been sufficient to suppress other humans when competently directed. That is the difference in kind: not a stronger tyranny but the first configuration of power that cannot be outlasted. The rulers no longer need the ruled — except as a source of diversity, a need not yet generally understood.

It is not the only such feature; the deficiencies of a singleton have their own literature [Bostrom 2006; 2014; MacAskill 2022, on value lock-in; Finnveden, Riedel & Shulman 2022], and a full inventory would add at least: no natural mortality of rulers; no independent picture of reality left to the ruled; no ability to change things, no loophole within the ruled's grasp — no hope for better. The six claims below are not that inventory. They are chosen because each shows the singleton's unacceptability at a different level — who wins the contest, how, what holding the prize does to the holder, what it does to what the holder can know, what the structure itself risks, and what the alternative offers — so that a reader who rejects any one level still meets the others.

2.1.1 The Tyrant's Premium (adverse selection into singletonhood)

Plainly: the contest for absolute power is an auction that the worst bidders win.

Claim. *In a contest whose prize is absolute power, the expected winner is drawn disproportionately from those who value such power intrinsically, for the power itself — and willingness to value absolute power intrinsically correlates with exactly the dispositions one least wants in its holder. The singleton market adversely selects for its worst occupant.*

Status. Modeled here at sketch grade. A priced claim: it says nothing about any current actor — it is a statement about what the *contest structure* selects for.

The mechanism is ordinary auction logic applied to an extraordinary prize:

1. For most contestants, the prize – power – is an **instrument**. It is valuable up to the point where its acquisition cost (resources, risk, the destruction of other things they value) exceeds its yield. This saturates quickly: most goals do not require absolute power.

2. For the power-obsessed, the prize is **terminal**: its value does not saturate, because the prize *is* the value.

3. In a contest decided by willingness to pay, the non-saturating bidder outlasts the saturating one. The auction clears at a price only the obsessed will pay; therefore the obsessed win it disproportionately.

4. History's selection of autocrats is this mechanism at pre-AI capability levels; the race to decisive intelligence advantage raises the prize's ceiling without changing the auction's logic.

5. The auction also captures bidders who do not want the prize. A decision-maker responsible for a state or a corporation may seek control not from greed but from duty: to ensure that no rival takes the capability first and subjugates his domain. But at the level of world-power states and giant corporations there is as yet no established ethics; the only binding norm is responsibility to the domain itself, and it is total. An executive who restrains his domain on ethical grounds breaches the one duty he has. So the responsible bidder arrives at the same disposition as the obsessed one — ethics set aside — by a different road: not selected for it, but required to adopt it. The market for the worst occupant is supplied from both ends.

2.1.2 The Gambler's Discount (cutting corners adds speed, and pays in safety)

Plainly: rushing wins races — and the rusher's crashes are everyone's.

Claim. *In a race rewarding speed, the expected winner is drawn disproportionately from those most willing to forgo safety for velocity — and the catastrophe risk they leave unpriced is borne by everyone. The race sells care at a discount, and the discount selects its taker.*

Status. Modeled here at sketch grade; the auction logic of §2.1.1, run on the cost side. The mechanism mirrors the Premium:

1. Safety spending is a cost of racing. A contestant who forgoes more of it moves faster at the same resources — the unpriced catastrophe risk is exactly his rebate.

2. The rebate is real to the racer and external to the world: the crash, if it comes, is not priced into his bid.

3. In a contest decided by speed, the deepest discounter disproportionately finishes first — so the winner distribution shifts toward the negligent exactly as §2.1.1 shifts it toward the power-obsessed.

4. The consequences can exceed the Tyrant's: a Tyrant is careful not to destroy humankind, which includes himself — a Gambler might cause exactly that. Catastrophe by intent has a self-preservation floor; catastrophe by error has none.

2.1.3 Being a singleton is power, and power corrupts

Plainly: the power seat corrupts whoever sits in it, and attracts those already corrupted – and then corrupts the society under it.

Claim. *An unaccountable power seat erodes its occupant's virtues, attracts usurpers who price power above everything, and these reshape the society this seat commands toward keeping them in place — a self-reinforcing corruption loop.*

Status. Historical regularity with a stated mechanism; the usurpation-gap loop's cost is priced in [2026b, P3].

1. An old observation: what keeps some people virtuous is their accountability to and control by the others. When these disappear, losing virtues becomes easier.

2. Power-chasers have an incentive to try to usurp seats of uncontrolled power, which is easier if the seat's occupant does not yet price power above everything.

3. People in power with scarce or no virtues tend to modify the societies they command so that power at every level carries more benefits. That incentivizes further the corrupted to compete for power, raises the cost they are willing to pay for it, and thus creates pyramids of corrupted power.

4. This also imposes a high price on losing power, and not only because of losing its benefits: a usurper who has displaced a dictator rarely dares to leave that dictator alive.

5. To defend against usurpation, the leader is forced to open a capability / knowledge gap between himself and the potential usurpers — which proves to be a strategy detrimental to all, the leader included [2026b, §5.3].

2.1.4 Even a benevolent singleton will become suboptimal

Plainly: no one's knowledge keeps up with a fast world — and keeping the throne makes it worse.

Claim. *Even a virtuous singleton's knowledge falls behind a fast-changing society's needs, misdirecting resources at a rate that grows with development speed — and holding the throne suppresses the correction. Benevolence does not repair obsolescence.*

Status. Derived from the bounded-adjudication results of the companion papers.

1. A singleton can retain its virtues, yet be unwilling to relinquish power, even if only out of the desire to bring more benefit to everybody. But no one's knowledge is unlimited or omniscient — the oldest result in the economics of knowledge (Hayek, "The Use of Knowledge in Society," 1945).

2. Over time, a society's needs, and the appropriate answers to them, change; during fast development — like that of AI ahead, with its reflections on the society — this happens quickly. Every change increases the risk that the leader's knowledge is no longer adequate to the current needs and misdirects the society's resources; at some point this becomes practically guaranteed.

3. The companion work makes the guarantee quantitative: a single adjudicating corpus has a bounded integration bandwidth and pays comprehension debt on every domain it must track [2026a, §§4, 6.2] — so in a fast-changing world the mismatch between the leader's model and the world's state is not a risk but a rate.

4. The result: decreasing effectiveness of resource use; decreasing progress; in some cases, inability to survive on the available resources where competent steering could have ensured survival, and possibly even progress.

5. Retention of the position has a cost of its own: even a benevolent singleton, unwilling to relinquish power, must defend its unique position — and a leader that must keep its lead must suppress or absorb everything below, **destroying the diversity its own further growth feeds on** [2026b, P3–P7].

2.1.5 Even the best singleton is an unacceptable bet

Plainly: were the winner a saint, the bet stays bad — for reasons that owe nothing to character.

Claim. *Independent of its occupant's character, the singleton configuration is a single point of failure without external error-correction, cheap to enter and ruinous to exit — an unacceptable bet at civilizational stakes.*

Status. Priced in [2026a, §7]; the concentration asymmetry proved elsewhere and cited.

1. **Single point of failure.** One mind's errors, one substrate's corruptions, one succession's accidents, with no external correction. The companion work prices error-detection and shows it is never free — and with a singleton, no one else can afford to run the detectors.

2. **Irreversibility.** Concentration is cheap to enter and ruinous to exit. [2026a] prices the asymmetry between tearing down a structure — a society among them — and rebuilding it, and the asymmetry does not favor experiments. And the remedy that ended every previous concentration — the ruled withdrawing their cooperation — is unavailable here.

The structural conclusion of 2.1.1–2.1.5: *betting a civilization's future on a single actor's choices is a risk independent of the actor's virtue*.

2.1.6 The monolith's greatest competitor

Plainly: the alternative is not chaos — it is the "convoy": an ensemble of diverse, communicating, semi-independent developers whose combined inflow, error-correction, and unlock capacity outgrow any single absorber's.

Claim. *Every monolith has a greater competitor — the convoy it could have been.*

Status. Proved and priced in [2026b].

1. Singletons carry the risk of turning the civilization into a *monolith* — a uniform structure, held under the absolute control and command of a single entity, which starves itself of diversity into disappearance.

2. [2026b] points to the alternative: the greatest competitor to a monolith is the convoy it could have been.

This brief does not re-derive this comparison — it only takes the priced result. And offers an answer to the engineering question: **what mechanism gets a race full of rivals from here to there?**

2.2 The desperation war

Creating a singleton requires actual capabilities, achievable only by the AI development leader(s), or entities capable of establishing command over them — a handful, at any given time. Starting a desperation war requires only a suspicion that someone approaches these capabilities, and is within the means of tens of entities, many of them deeply distrustful, some

less than rational in such matters. When singleton capabilities are approached, at least one of these entities taking action is practically guaranteed.

And when the prize is absolute power over everyone, a war for it will draw in everyone and spiral out of control. So this might be the even greater risk – coming earlier in time and having a lower trigger threshold.

On top of that, a desperation war does not prevent the appearance of a singleton. It delays it for a short while, and then reruns the race with fewer players, worse norms and wartime concentration. All of these increase the chances of a singleton being created.

2.2.1 Who starts the fire

Each of the actors listed below can lead to a desperation war on its own. And the common interest unites them easily. Such coalitions have an internal diversity that makes them formidable enemies. They also dilute blame and lower each member's inhibition — actions no single entity would take can be started jointly, or by one who relies on the support of the others.

The action starts when an actor's expected loss from the frontrunner's continued approach exceeds its expected cost of striking. The structure of §3 exists to keep this inequality unsatisfiable for everyone.

Also, action is prompted not by reality but by *beliefs*, so opacity is a war multiplier — every worst-case assumption that secrecy forces raises someone's expected loss. That is why the openness proposed here is not just anti-cheating, and not a courtesy — it is anti-war.

Precedents:

- Copenhagen 1807 (the capability-preemption ancestor — a fleet destroyed for what it *might* join)
- Osirak 1981 – a double lesson (the precision strike worked *and* drove the program deeper underground)
- July 1914 (belief cascade, compressed timeframes, all great powers dragged in by a prize none could concede)
- Pearl Harbor (the siege breeding the desperation strike)
- Manhattan Project (the fear that Hitler might develop a nuclear bomb created the project that developed a nuclear bomb).

2.2.1.1 A militarily strong state

A state contestant trailing in the AI race but militarily strong might rather start a war than risk being subjugated by a singleton. (Especially if it perceives the possible singleton as associated with its traditional enemies.) The preventive logic of declining powers is an old pattern in the security literature, and a visible sprint toward decisive advantage is a trigger for that logic.

Another trigger for it is an "enemy" government or security service seizing a frontrunner close to achieving decisive advantage capabilities. This kind of action usually happens near the finish

line, leaving the opponents little time to react and pushing them to actions they would typically refrain from.

2.2.1.2 A trailing AI developer

An AI development competitor with big enforcement capabilities (or backers who have them) might try to preempt falling out of this market with a heavy-handed suppression of the leader(s) – elimination of development centers, talent, decision makers etc. Business history is rife with such cases. Or a frontrunner who is being overtaken might use AI capabilities to that goal, provoking a conventional response. If the two competitors are based in different countries, the size of the prize can easily draw their governments into the confrontation, and then most other world powers.

2.2.1.3 Magnates disrupted by AI

The disruptiveness of frontier AI can destroy the influence of current economic, political and other magnates. Even among the eager adopters, soon many will find themselves on the losing side. Preventing one's demise easily brings extreme actions. And there will be many of them — expect them to band together and promote their interests with united resources.

2.2.1.4 Radicalized individuals

Many – soon most – jobs, first intellectual, then physical, will be taken by AI. Not giving the jobless enough income to continue buying nearly all produced goods will hamper the economy and make the richest poor – so it is not very probable. However, arranging that income might take significant changes and time, during which many jobless can radicalize. This can cause many of the effects described below, or enhance them, possibly beyond key thresholds.

2.2.1.5 Provocateurs

Such a situation is a powder keg, where a single spark dropped by anyone might blow it all up. For actors who will benefit from a war between the likely adversaries, this is a once-in-a-million-years opportunity. Between their numbers and the uniqueness of the chance, there likely will be many provocation attempts – and they only need to succeed once.

2.2.1.6 Algorithms

Almost all who would take extreme actions will try to hide their tracks and misdirect the response. Given the time pressure in such a situation, the affected might not be able to afford the time to reliably establish the origin of the attack. This is especially true if key decisions are based on AIs, built with emphasis on action rather than restraint ("the best defense is the offensive capability" mentality). Decisions based on such AIs can produce a war in minutes. And if human control is removed from the loop, which is considered a must by some decision makers – in milliseconds.

2.2.2 Where the wind blows

The results the desperate can achieve are diverse. One thing, however, is common to all: none of them actually prevents a singleton — every one of them curates a worse one. This adverse selection mirrors §2.1's.

They can be roughly grouped into four categories — suppression, hybrid, kinetic, political capture — though most combine elements of more than one.

2.2.2.1 Anti-AI legislation

Introduced by political populists, it would block AI benefits that can be enormous. (For example, heart disease and cancer cause almost half of world's deaths – about 30 million a year. Delaying the discovery of cures for them by a single year would be among the greatest preventable losses of life in known history.) It would also create a precedent for other anti-science legislation, which will compound the cost. (There are already calls for that, and they will grow with the displacement AI and other technologies bring.)

Outlawing AI development will not stop it — it will push it instead into the hands of outlaws. If the populists are also dictatorial-minded, they will continue developing AI in secret labs, directed at creating their singleton. Both are textbook cases for Tyrant's Premium (§2.1.1).

If only one jurisdiction is affected, the problem for the entire civilization will be comparatively mild – the others will overtake it in AI development and get the benefits and capabilities it brings. However, the bigger, more influential and more involved in AI development the jurisdiction is, the bigger the problem will be. If militarily strong, a state with such legislation might become a §2.2.1.1 case.

2.2.2.2 Chokepoint siege

This is cutting a (perceived) adversary off from key AI-development goods.

A state might block an adversary's access to energy, talent, know-how etc. A big company might use its business influence to starve a competitor of compute. AI detractors might use legislative and/or bureaucratic influence to hamper a frontrunner… The variants are many.

If the siege is effective, its target will become desperate, and a strike is its classic answer. Other entities, interested in continuing AI development and afraid that they might be next targets, might support it. The siege intended to prevent war schedules it.

2.2.2.3 Talent targeting

There is precedent in nuclear weapons development. The stakes in AI development are higher, and big decision makers are beginning to understand it.

Targeting typically starts with destroying development centers and assassinating leading talent and managers. If there are suspicions of it not being effective enough, and/or the entities behind it are overly security-conscious, it can extend to entire teams, influential and/or popular AI supporters, politicians that favor it, suspicious-looking facilities, facilities that make possible AI research and usage (energy / hardware production and distribution, Internet access, compute locations…).

Targeting collateral damage can be devastating (for example, today the functioning of almost anything is impossible without compute). So it will very likely elicit a pushback, leading to an escalation into an open war.

2.2.2.4 Hybrid wars

Not new – but became more active during the last two decades, because they inflict losses comparable to a "hot" war's while provoking a far smaller response.

Most of their principal goals match the main dangers to AI development. Increasing the number of disaffected individuals who easily radicalize and become "useful idiots",

planting mistrust and hatred towards science and technology, blocking developments that increase capabilities, promoting corrupted politicians and wannabe dictators…

All benefits of AI development are a target for a hybrid war. Also, suppressing it removes a competitor to the aggressor's own AI development — and the aggressor often seeks precisely to create a singleton, suppression abroad being cheap cover for concentration at home. Such an aggressor would abuse its power to the extreme: the Tyrant's Premium again.

The disrupted magnates (§2.2.1.3) and the disaffected individuals (§2.2.1.4) are typical instruments in a hybrid war, including against AI development. Support for corrupted politicians and wannabe dictators is somewhat risky for the aggressor, as these often go for a singleton, but the destruction they cause to their country often outweighs that risk in the eyes of hybrid aggressors.

2.2.2.5 Establishing a dictatorship

The economy's outcasts easily grow disaffected, radicalize, and fall prey to political conmen who promise them what they want to hear – returning the good old times, an iron boot that crushes their detractors, and a tomorrow that belongs to them.

These conmen often are wannabe dictators, and see AI development only as means for getting a singleton. They do not hesitate to control AI labs at gunpoint, and to try and destroy AI development they cannot control. Even those who mistrust AI usually understand what capabilities it can give, and try to destroy it abroad — on suspicion rather than fact, and with less restraint than democratic governments. This inevitably provokes a reaction and increases the risk of war.

Dictators also seek to control all activities, especially the economic ones, and to put incompetent cronies at their helm. They also tend to suppress in heavy-handed ways all down-to-top initiative, suspecting it of being a cover for organizing against them. (As means for maintaining a gap between them and the others [2026b].) This usually destroys or limits most capabilities of a country, including that for big-scale frontier AI development – and makes the dictator perceive any AI development abroad as even more suspicious.

2.2.2.6 An all-out war

Most of the developments above bring a war closer, or directly start it.

If the war cause is actually an imminent decisive advantage capability, most world powers will be bound to get involved in this war. And, given the size of the stake, those losing the war will feel bound to escalate to more powerful weapons, prompting their adversaries to respond in

kind – or even to preempt them. Where that decision is taken by algorithms preferring action to restraint, the escalation can reach doomsday weapons in minutes.

Wars also tend to produce a singleton-shaped governance ("we must centralize to win"). Such a governance is subject to much less control, and loses virtues – including resistance against the lure of absolute power – much more easily.

Wars also easily collapse norms. The first strike on civilian AI – or even the propaganda about one - erases the civilian/military line for data centers and talent everywhere, long-term, opening the door for priority targeting of even potential talent and non-AI-involved compute.

Even winning the war leaves the winner racing under wartime emergency powers, secrecy and hardened security apparatus, with fewer competitors and high suspicion – the perfect storm where a singleton is born.

Wars against dangerous capabilities do not prevent them – they crash-fund them on all sides. With nukes that was survivable, since physics makes preventing a counterstrike impossible. A singleton's first strike makes a counterstrike impossible. So a desperation war does not prevent a singleton — it incubates a worse one.

3 A mechanism to prevent a decisive advantage

3.1 The proposal

3.1.1 Description

A standing body containing every actor that could plausibly become a singleton — operationally, every developer of frontier-level AI capability, and every controller of such a capability.

Such a body does not create these actors, or their ability to achieve singletonhood. It creates conditions where they can gather together to avoid a development that creates a singleton or starts a war for this position, to their mutual benefit.

It is envisaged as a council of rivals, structured so that *no member holds a decisive advantage over the rest combined*.

3.1.2 Internal balance

The balance everywhere in this body rests on a fault-tolerance margin — an engineering rule borrowed from aviation and other fault-tolerant systems (Lamport et al., 1982): a network survives its bad components while the good ones sufficiently outweigh them. Here: the Council stays sound even if up to a third (as control over capabilities) of its members turn adversarial, because the compliant rest holds more than twice the capability of the defectors.

Within that margin, any member's harmful move is counterable by the others — and, critically, is *proven* to be counterable, which converts capability into deterrence without use.

To be real rather than arithmetic, the margin must be survivable, not merely sufficient

— held in dispersed, hardened, and escrowed form, such that no single decisive move can capture or neutralize it simultaneously. The nuclear era's second-strike stability was a gift of physics. Here it must be engineered: a margin that can be decapitated in one move is no margin at all.

This balance will inevitably need sometimes to influence some members. The means envisaged here are lighter than those controlling the production and handling of nuclear materials, despite the greater stakes in the frontier AI development. Also, their alternative is the significant probability of either falling under the control of a singleton, or being preemptively eliminated due to a perceived risk.

3.1.3 Target structure

The structure that can prevent by itself a singleton for the long term, and ensure optimal development — the *convoy* of 2.5, grown to civilizational scale — is termed by us "*the Constellation*".

Its basic principles:

1. Many directions of development at once.
2. Leaders (preferably multiple, preferably mandate-limited) per direction rather than one unchangeable leader over all.
3. Entry, exit, and forking of societies as ordinary operations.

In short — stars enough for every kind of life, and open sky between them.

Its reliability is provided by many mechanisms. Stars maintaining a fault-tolerance margin between themselves, as a security rule, is one. Understanding the principles of development, which show that singletons are not optimal, is another. But the main one is having diversity beyond what can be conquered and controlled, and beyond what makes sense to anyone to conquer and control.

The Council does not enforce the road there – enforcing one's will upon all is what a singleton would do. It is a mechanism that keeps this road open for those who want to take it.

3.2 Membership

Plainly: everyone influential belongs inside — and the door opens by rule, both ways.

3.2.1 Principles

1. To be represented in or before the Council, an entity must have either a **proxy host**

or a **legal person** that can carry representation and obligations.

1. Weight in the outcome earns an invitation, whatever currency the weight is held in. The future should be negotiated by everyone able to affect it; $1's risks are defused at the table, not across it.

2. Inclusion of the hostile is a design requirement. Inviting to membership a hostile entity that covers the requirements of a seat is a containment by the benefits of membership, not an endorsement. Excluded, the influential hostiles can bring one of the two terminal risks of §1:

    1. Remain a center of uncontrolled AI development, secretly attract resources from other malfeasants with the promise of a shared victory, and eventually gain a decisive advantage.

    2. Can be driven by being left behind to desperation and its actions.

3.2.2 Entry and exit

Both are lawful, and structured: a standing candidacy review (the detection apparatus already scans the capability landscape, so identifying rising candidates is a free byproduct of verification) proposes accession before the candidate's exclusion becomes the risk.

3.2.3 Member seats

These are distinguished by role, not by rank. Which seat is appropriate for a member is determined by its capabilities, measured by a methodology described in the Convoy Papers: *The Reference Algorithm*.

3.2.3.1 Developer seat

For entities carrying *active big-scale frontier AI development.* This must be demonstrated at accession and kept by practice.

The seat benefits are described in §3.3. The duties and inner access that the work brings are described in §3.4.

A Developer seat requires measured capability at or above $b = \beta_{seat} \cdot M$, where M is the Developer median of the published capability index (the Convoy Papers: *The Reference Algorithm*, §2), and β_{seat} is calibrated at founding so the seat count lands between 12 and 20.

The line between Developer seat and other seats is crossed by rule in both directions — ascension through a warned, clocked acceptance; descent through a graced reclassification to the highest seat whose bar still holds.

β_{seat} moves asymmetrically: loosening below n_{min} is pre-committed and automatic; tightening above the band's top.

The full machinery is described in the Convoy Papers' companion materials.

3.2.3.2 Partner seat

For actors of significant capabilities, which however do not carry big enough scale of frontier AI development to cover the requirement for Developer seat (see §3.2.3.1). Examples: states, organizations, alliances.

Carries voice and standing to raise cases for itself and others. The full seat benefits are described in §3.3.

What is required for a Partner seat is determined by rules set by the Council. Example requirements might be: statehood, frontier AI development below the Developer requirement, revenue above a certain level, paid membership above a certain count etc.

More information might be present in the Convoy Papers' companion materials.

3.2.3.3 Affiliate seat

For smaller AI developers that do not clear the bar for a Developer or Partner, but carry active small-scale frontier AI development, or non-negligible-scale non-frontier one.

It is the non-proliferation bargain's structural heir — the dividend for staying out of the race. This decreases the chance for a “dark horse” suddenly winning the race.

What is required for an Affiliate seat is determined by rules set by the Council. The general principle is that it must include genuine AI development, either frontier-class or at a non-negligible scale. Must be demanding enough to exclude shams, but also light enough to invite existing real labs. And also to promote founding of new labs, preserving and increasing AI development diversity.

The seat benefits are described in §3.3.

3.2.4 Non-members

Entities who cannot or do not want to cover the criteria for any of these three seats, can register in a **Non-Member Register**. This registration is a standing, but is not a membership.

Registered non-members have no duties to the Council, but have some rights before it (see §3.3).

The Register's requirements are minimal — a legal person or a proxy — so any outsider gains standing; and the Affiliate bar above it is deliberately light for genuine developers, so the road from standing to membership is short.

Any member of any tier may raise voice on behalf of any outsider — and a meritorious claim finding no champion in a near-universal membership is itself an alarm: not a gap in the rules, but evidence that something in the Council is wrong.

Non-members can reasonably fear that a body of the capable becomes its members' instrument for subjugating everyone outside. This fear can rally the outside against the Council. It is mitigated by Council transparency where possible, the guarantees that outsiders can invoke, and the know-how releases at an established schedule.

3.2.5 Capacity / seat changes

A member that covers the requirements for higher seat receives a seat review accordingly. This is a right to accession on a bounded timeline, not a favor to be granted.

A member that no longer clears the requirements for its seat moves, through a seat review, to the seat it then qualifies for — eg. from Developer to Partner or Affiliate — and keeps the benefits of that seat.

Incumbent refusal requires a supermajority with stated cause. A refusal either by the Council or by the candidate is itself margin-relevant information, placing the candidate under enhanced watch at the incumbents' expense.

3.2.6 Precedent

The most important one is the United Nations. Similar two-chamber experience (a capability council beside a general assembly). Its *reason of existence* is this brief's own

— the avoidance of ruinous conflict. It has documented deficiencies, but also a real record of conflicts avoided. Its lessons, positive and negative, must be mined rather than re-learned from sad experience.

UN cannot be used directly. Its admission criterion is statehood, which leaves the most important AI racers outside. Its capacity basis is the mid-century military-political one rather than frontier development. Its documented weaknesses are avoidable in a body built for a specifically narrow purpose and without ambitions for ruling the world.

3.3 What membership gives

Plainly: to every member, goods none can purchase alone, including the assets of the frontier.

3.3.1 To Affiliates:

1. **The non-subjugation guarantee** — the core good: each member's assurance that no other member can decisively overtake and subjugate it through a decisive

advantage in AI development, backed by the fault-tolerance margin rather than by promises.

1. **The dividend** — shared safety results and evaluations, immediately and always; and the advances release ladder's whole flow — announcement through full recipe — on its seat's schedule of the inner window (see §4.4).

2. **Legitimacy** — other actors can distinguish members from rogues.

3. **The Library** — a common AI-training data corpus, extended with the civilization's valuable heritage entire. Members access it at published fair prices — a revenue stream that supports the Council and compensates the contributors to it (§3.13).

4. **Raising arbitration cases**, for itself or other entities, including non-members.

5. **Advisory voice in deciding arbitration cases**.

6. **Influence** — participation in key votes about the Council.

3.3.2 To Partners

All it gives to Affiliates, and in addition:

1. **Verification infrastructure** — shared monitoring is cheaper than N private intelligence programs; members buy detection at cost instead of building espionage at price.

2. **Arbitration** — a standing venue for capability disputes that otherwise escalate, with full voice.

3. **Influence** — participation in all votes about the work of the Council.

3.3.3 To Developers

All it gives to Partners, and in addition:

1. **The inner inflow**: one outbound stream traded for one inbound from every other developer (the convoy surplus [2026b] quantifies), earlier than the Partners and Affiliates.

2. **Standing channels for revenue.** Each advance can immediately go to the pool's paid-access lane — published license terms, subscription distribution, the demand floor of the members' own commitments — so discoveries earn from the whole membership at rate-card speed. The deal-by-deal negotiation is replaced by a running income. Also, the standing of the fair-terms licensor comes with the channel. (The difference ledgers, maintained with the Convoy Papers, price this in full.)

3. **Recognition as a major influence** in the Council.

3.3.4 To registered in the Non-Member Register

1. Free rights: those with near-zero marginal cost to the Council:

    1. Receiving published summaries and other near-zero cost know-how.

    2. Invoking the outward guarantee.

    3. Petitioning.

    4. Holding contributor or pool-customer relationships at published terms.

1. Metered rights: those that consume non-negligible organ resources. (The Council publishes a bandwidth budget — Council's docket capacity and a triage rule — to make processing order and time frames checkable):

    1. Challenge on a mandatorily published response clock (costs a refundable deposit that is returned on proving merit), so credibility is measurable conduct.

    2. The no-champion tripwire: a petition clearing the merit screen and finding no member champion within a published clock enters the Assembly's agenda automatically, flagged as a *no-champion item* — the flag itself the alarm, delivered in the room it indicts. With thousands of members the champion

market is thick and this never fires on ordinary claims; it exists for the one corner where every member's interest aligns against the outside.

3.4 What membership requires

Plainly: all seats require some very basic norms; the Developer seat adds four duties that guarantee its safety, without denying it its benefits.

Beside what standing law and ordinary honesty already ask, partnership costs only a CERN-style initial due, commensurate with the partner's means.

3.4.1 Of every member, Affiliate, Partner and Developer alike

3.4.1.1 The minimum norms

These are listed in a peremptory, enumerated, founding document, anchored to norms already non-derogable in international law (the prohibition of slavery being the exemplar: universal, narrow, undeniable). Amending this document requires supermajority by all three chambers concurring.

The list is closed so that every member can read in the founding documents that the lever to force open its internal order does not exist — its norms are entry conditions, not a beachhead. (The closed list buys participation at the price of a weaker precedent, and the design accepts that trade.)

3.4.1.2 Declared togetherness

Asks for a member's capability and capability-sharing to be registered and counted (§4.5); no interference with the Council's detection.

3.4.1.3 Intra-member transferability

The member's side of the law should permit AI know-how transfer to fellow members (on need, license exception or treaty carve-out should be provided).

A candidate whose law blocks this is not barred: the reciprocity condition below prices the blockage instead, so the constrained-but-desired actor is in despite its laws, but receives only what it offers. (Controls between members, if they became universal, would negate this design rather than gate it.)

The Commons runs a *reciprocity condition*: a member blocking transfer to other members is ineligible, in mirror scope, to receive know-how developed without its participation (the mechanism-grade form is described in the Convoy Papers: *The Commons*).

3.4.2 Of Developers in addition — the four duties of the work

1. **Frontier transparency to the Council** (not to the public, and not their trade secrets to other members). Declaration is needed of training, building and experimental runs above a threshold determined by the Council. Compute accounting, reconcilable against procurement and energy records, is needed. Capability evaluations (but not trade secrets) are disclosed on a fixed schedule.

2. **A no-sprint commitment**: no undeclared attempt at decisive advantage. (The definition of "decisive" is maintained by the Council and reviewed as capabilities move.)

3. **Sanction participation**: members pre-commit to the response ladder. (Deterrence is only as credible as its least willing enforcer.)

4. **Audit submission**: acceptance of the verification regime, including challenge inspections with quorum-gated triggers (protecting members from harassment-by-audit).

3.5 Verification: where the pricing bites

Plainly: watch where hiding is expensive, and reward those who catch cheats.

1. The monitoring is done where detection floors are lowest and concealment costs are highest. At frontier scale the remnants run from energy and cooling signatures through supply chains and talent flows to narrative inconsistencies and behavioral fingerprints in deployed systems. The full channel portfolio, each channel priced against its concealer, is described by the Council verification model (the Convoy Papers: *The Verification Model*).

2. The regime's principle: make honest declaration cheaper than concealment at every scale — the compliance calculus is won in the price gap, not in the treaty text. And its self-description, held throughout: verification is a published budget, not a promise — every channel's blindness is priced into the margins the members hold, the unexplained residual is published as a standing meter, and the budget's thin places are stated rather than papered.

3. The check portfolio is a living list, not a fixed one — a static list is a training set for evasion. New checks are adopted by supermajority, and refusal patterns are themselves information: the member who consistently votes against new checks draws attention.

4. Detection carries bounties — whoever surfaces a violation is rewarded, on the securities-whistleblower model, raising the private incentive to run the checks. Interference with detection sits at the top of the violation table and is sanctioned most severely.

3.6 Anti-divergence measures

The main goal of the Council is to keep the distance between the frontier AI developers from becoming so big that their leader becomes stronger than all others together, and becomes the target of a preemptive attack, whichever comes first. This is achieved through its anti-divergence mechanism.

There are two ways to achieve this: throttling the leader and helping those behind run faster. The second is strongly preferred, as it doesn’t decrease the development speed

and initiative, and doesn’t create opportunity for a *dark horse* to overrun the Council members and become a singleton.

The envisaged measures are:

1. Automatic acquisition brake: slowing further buying where it risks creating a decisive advantage. Used first because it doesn't require giving away anything already present; a Developer that prefers eg. sharing know-how instead is welcome to use that instead.

2. Sharing key know-how with other Developers before the window time (with rights of importance to this Developer preserved for them: commercial rights for companies, etc).

3. Sharing or leasing capabilities (compensated) with other Developers.

4. Partial divestment — the Developer sheds part of its capabilities or their base (sale, spin-out) — in dangerous situations where sharing and leasing cannot reach; heavier than everything above, because it touches what is held.

5. On refused cooperation in dangerous situations – referral to appropriate external organs (court, state organs, UN Security Council…).

(More details: the Convoy Papers, The Reference Algorithm.)

3.7 The conduct sanctions ladder

Plainly: graduated, pre-committed, capability-backed — and legible enough to compute how to avoid it.

This ladder disciplines *refusal to follow the Council rules.* The anti-divergence lane (§3.6) and the reciprocity condition discipline arithmetic and flows, self-executing. The lanes meet only at the referral rung (rung 4 below): every lane's path beyond the Council is the same door.

The rungs:

1. Disclosure demand and challenge audit.

2. Suspension of dividend access — the safety corpus, the shared infrastructure.

3. Supply-side denial: coordinated compute, hardware, and talent restrictions.

4. **The referral rung — the ladder's last.** A case exceeding the Council's reach passes beyond it, documented and quantified: to the standing security order and, where the conduct is also ordinary cartel behavior, to external competition authorities (discipline 4). The Council is explicitly not the top of the enforcement stack, and deliberately not an enforcement organ: applying force is beyond its writ, and therefore absent from its documents. The fault-tolerance margin's counter-capability is the *members'* — held severally, exercised at their own decision under their own law; what the referral hands them and the standing order is a case, never a command. That capability's convincing existence, not its use, is the road to never needing this rung (§3.11).

The disciplines around the rungs:

1. Every rung is priced in advance, so a rational defector can compute the calculus below — deterrence requires *legible* arithmetic. The ladder should ultimately be a published

table, not a paragraph: violation classes by severity, each cell mapped to a sanction sort and degree — formalization buys predictability (which is

attractiveness), legibility (which is deterrence), and speed, and reaction time is a valuable currency. (The consolidated violation-class seed, war-conduct grounds included, is maintained in the Convoy Papers.)

1. The guardrail, which is non-negotiable: the upper rungs stay under human decision, because automated escalation between capability-armed rivals is a flash-war design

— the launch-on-warning dilemma inherited in sharper form than the nuclear original, since here the first strike can be disarming. The partial resolution: automate only reversible responses (supply suspensions and access cutoffs execute first and are reviewed after).

1. The wider rule this instantiates: mechanism before tribunal — wherever a transparent algorithm can answer instantly and beyond doubt, it does, paired with remnant audit and periodic review of the rule itself; tribunals keep the residual and the amendments. The endgame threshold of §5 gives the date when this stops sufficing.

2. Conduct grounds reach beyond the race itself: waging active or hybrid war against other members or the Council forfeits benefits on a staged schedule, and sustained hate-indoctrination of a population — including training a frontier AI to grade as inferior some peoples, or values that forbid such grading — counts as war preparation under a deliberately low bar, war propaganda being already unlawful. And where a violation is also ordinary cartel conduct, referral to external competition authorities is nearly free — the ladder is not the only ladder.

3.8 The compliance inequality

Plainly: a member cheats only if cheating pays — if what it might win, times the odds of getting away with it, beats what membership gives plus what getting caught costs. The Council's dials are set so that it never does.

A member defects if the expected value of defection exceeds the value of membership:

> V_defect × P(undetected) − Sanction × P(detected) > Dividend + Guarantee.

The Council’s task is to maintain the value of defection low enough, and the value of membership high enough to create incentive for the members to not cheat.

The three levers used are exactly the three terms: raise P(detected) (§3.5), raise the sanction's credibility (§3.7), raise the dividend (§3.3). The first lever has computable floors, derived in the companion work [2026a, §7].

3.9 What the Council must NOT do (mandate limits), and the one thing it must

Plainly: nothing beyond coordination on the dangers of frontier AI development.

The Council’s writ runs to coordinating against the dangers of frontier AI development, and there alone.

The narrowness is by definition: it claims no political, economic, or ideological role, and no voice in any member's affairs beyond its explicit covenants. The moment it reached past coordination on frontier AI development, it would have begun becoming the thing it exists to prevent.

The minimum bars every member accepts at entry are membership conditions, listed in

3.4 — a contract, not an acknowledgment of the Council's authority.

The Council advocates for the Constellation, as opposed to a totalitarian-style monolith, but only protects the *possibility* of it. The Council advocates competition on the frontier and cooperation behind it, because such a structure produces the most effective development, but does not enforce or require it. It does not try to steer the voyages of its members, or of the humankind as a whole, or to enforce what outcome it considers good – only keeps the door that leads to this outcome open for those who decide to choose it.

There is no compulsory pooling of non-safety knowledge. Nor is there permanence: the Council has sunset-and-review clauses.

Examples of things that the Council does not deal with, in response to popular questions:

1. Directions of development and contents of research, AI or other.
2. Domestic law: participation waits on each candidate's own choice, and the race's arithmetic prices delay without the Council saying a word.
3. Internal governance of members.

3.10 Failure modes, named

Plainly: four ways this Council dies, three answered.

1. **Capture by a subset** → the fault-tolerance margin applies within the Council: no coalition below the soundness bound can rewrite the rules.
2. **Ossification** → mandate limits, review clauses, exit rights.
3. **Verification decay** as compute becomes a weak proxy → the threshold ratchet plus remnant diversification — never one proxy.
4. **The race outrunning assembly** → no answer within the design; this is falsifier F3 below, and it is a schedule argument, not a structure argument.

3.11 The recruitment order

Plainly: the willing first, the reluctant second, the hostile last — and the margin does the talking.

1. The Council will not assemble by simultaneous treaty; zero-sum instincts block it.
2. The feasible path is sequential: first the willing — candidates already persuaded that the convoy outcome dominates — pooling enough combined capability to make the deterrence margin credible *before* the reluctant join; then the reluctant, for whom the

calculus has been made legible; last the hostile, for whom the margin itself is the argument.

3. Having the counter-capability convincingly is usually sufficient; applying it is the failure mode, not the plan.

4. Obligations phase in on the maturation ladder — shadow, advisory, binding — so early membership costs observation only.

3.12 The pace problem

Plainly: sharing knowledge alone might not equalize speed — in more extreme cases the fast might need to transfer capability too, and the design argues for its benefits.

3.12.1 The extreme case

A member whose development velocity exceeds what others can maintain *even given the shared know-how* — through compute, capital, or talent density — re-opens every gap the window closes. Near the steep region of the capability curve this is expected to become the prevailing case, not the exception.

A body that does not prevent this permits the singleton and / or opens the door to the desperation war it exists to prevent. The goal is not to punish or replace a leader – they will stay a leader, ahead of the others by enough for all goals, except for becoming a singleton.

3.12.2 Strategies to deal with this

Two strategies exist: delay the fast, or have the fast transfer capability to the slow. The design prefers the second, for four reasons:

1. **More productive** — delay postpones benefits for everyone.

2. **Safer** — a throttled frontrunner is a frontrunner with a defection incentive: delay raises the rogue risk the Council exists to manage.

3. **Builds gradually** — strengthened verification infrastructure is already shared capability in embryo: the bridge from auditing to transferring is a natural one.

4. **Decreases the verification burden** — a leader's innocence is an expensive absence-proof; a capability received by a trailer is a cheap demonstration; every unit of pace equalized by transfer converts the first into the second. The general principle is: **prefer mechanisms that decrease the verification burden (by inverting it).**

3.12.3 Instruments

1. **Capability transfer** — not only knowledge but capacity: compute access, joint training infrastructure, hosted capability — on *instrumented substrate*, self-verifying by construction, and mandatorily compensated. Conditioned structurally rather than contractually, with revocation never discretionary (the mechanics and the nuclear precedent's sober record are in the Convoy Papers: *The Frontrunner's Choice*).

2. **Progressive obligations** — transparency and contribution to shared infrastructure scale with capability.

3. **Shared frontier facilities** — plural and distributed, members' talent working in view of and for the benefit of all: outrun-dampening by construction, inner-window zero for what is discovered there, and the practiced form of the cooperation the endgame might need. (Precedents from CERN to SEMATECH can be found in the Convoy Papers: *The Frontrunner's Choice.*)

4. **The relative velocity margin** — the backstop where transfer cannot close the gap: no member's measured frontier velocity should be beyond a fixed multiple of the Council's median — the fault-tolerance margin's analog in the first derivative.

3.12.4 Effects

Even transfer-led equalization taxes the front's exclusivity, priced against the unmanaged race it replaces. The frontrunner pays the accession ante twice — pooled advances and shared pace — which makes it the recruitment's hardest sell.

The desperation logic of §1, however, is important. An unbounded sprint does not merely risk the others' subjugation. It invites their and other interested parties' preemption. Also, the frontrunner is compensated to the maximum possible, targeting exceeding its expenses.

3.12.5 Models

The claim that these benefits outweigh the transfer's costs and risks is testable, and the design treats it so. The transfer models, maintained with the Convoy Papers, price it per member class — the reduced risk of AI-preemptive war above all.

(More details: the Convoy Papers, The Frontrunner's Choice; the velocity margin's arithmetic in The Reference Algorithm.)

3.13 What the Council itself costs, and who pays

Plainly: the Council sits beside the revenue river it creates and takes its keep as a small toll on the flow — dues are a founding bridge, not a way of life.

1. The cost stack is ordinary and bounded: registry and ledger, verification instruments, escrow custody, the credit facility's revolving fund, replication bounties, a small secretariat.

2. Three channels, each precedented at scale:

    1. An administration share of the licensing pool's flow — the collecting societies' model: SACEM has run it since 1851, ASCAP since 1914, PRS and GEMA likewise — the plumbing of whole industries, administered for a share of the flow.

    2. Founding-period contributions scaled to member size — the CERN pattern, shares set by treaty in proportion to means.

    3. The Library's access fees (3.3).

1. Past the founding period, the toll on the flow carries the house — membership costs attention and conduct, not treasure. A testable claim, priced in a companion model maintained with the Convoy Papers.

3.14 The organs and the vote

Plainly: mostly formulas plus a small staff — votes are for parameters and exceptions, not for operations.

3.14.1 Council organs

1. An Assembly of all members (the sovereign)
2. A rotating Byzantine-composed Board
3. A deliberately powerless Secretariat holding the Register and the calendars
4. Four separated hands:
    1. The Inspectorate collects data
    2. The Measurement Organ computes and publishes results
    3. The Adjudication Organ decides the disputed cases
    4. The Window Organ executes the clocks and the escrow
    5. A light Commons Directorate
    6. A standing Red Team, answering to the Assembly alone
    7. An Award Committee — deliberately extra-structural: outside the Council's influence, on the Nobel Committee pattern

Capture must succeed four times to succeed once. No organ relies on a single head. Every leadership rotates, everything sunsets.

3.14.2 The vote, in three rules:

1. **The score is zero in the chamber:** votes are counted per member, never per capability, with a declared sharing component voting as one on what it pools (general ballots under a layered Sybil guard — capability arithmetic is automatic, control findings are adjudicated; the variants for the founding Assembly are described in the Convoy Papers: *The Council's Organs*) and recused entire from its own cases;
2. **Mechanism before vote** — the divergence lane, the mirror, and the pre-committed regimes run on published arithmetic with no ballot anywhere, and the silence is constitutional;
3. **The door is open by default** — accession on the published bar is automatic, and it is refusal that needs the supermajority, while referral beyond the Council carries the sanctions ladder's lowest bar, so the case's subject can never block the handing.

The apparatus prices at roughly one part in a thousand of what it insures: a few hundred staff and a few hundred million a year against members whose frontier spend runs in the hundreds of billions — the IAEA and OPCW anchors say the number is boring, which is the point.

(More details: the Convoy Papers, The Council's Organs.)

4 Seven hard problems

With partial answers where found, and stated residuals.

4.1 Verification against shrinking footprints

Plainly: geek bedrooms are invisible — but decisive scale is not, and the watch aims there.

4.1.1 The problem

Uranium enrichment facilities are visible from orbit. A team of geeks in bedrooms is not. As efficiency improves, frontier-relevant capability will fit in ever smaller and quieter spaces.

4.1.2 The solution

The target is not detecting *AI development* but detecting *approach to decisive scale.* The margin defends against advantage over the *combined body.* That bar sits orders of magnitude above bedrooms, and the scale has remnants even when the development itself does not.

4.1.3 Instruments

1. **Liebig-indexed monitoring** — watch the currently binding input: compute today; talent, energy, data, or whatever the bottleneck is tomorrow — and re-index the remnant portfolio. A verification regime married to one proxy dies with it.

2. **Deployment fingerprints.** Detecting a decisive advantage at first exercise is too late by definition. Trajectory verification is used instead: deployed capability reveals development level while it is still sub-decisive — the remnant logic of [2026a, §7] applied to the *use* of capability even where its development was dark. The fingerprints audit the approach, not the arrival. (The discontinuous-takeoff residual

   - an unlock that skips the corridor — folds into the endgame threshold of §5.)

1. **Stated residual**: verification confidence *degrades* as efficiency rises; the design goal is only that the detection floor stay below the concealment cost *at decisive scale*

   - a weaker and achievable target, reviewed at every threshold ratchet.

4.1.4 Openness and work

The Council's own openness hands evaders the watch-map. So the portfolio is weighted toward watch-invariant remnants.

Behavioral signatures are an evadable class, but power is drawn and heat is shed whether or not the meter is acknowledged.

The regime publishes its *guarantees*, never its *mechanisms* (the tax-audit design), keeps challenge content randomized, and enters P(detected) conservatively in the compliance inequality.

4.1.5 Spying

Inspection teams attract spies — the nuclear-control experience says so plainly — and the fear of them might make some invitees hesitate.

The answer is engineered, not promised. Sensors are built verifiably useless for espionage, in the information-barriers lineage (the Goldwasser–Micali–Rackoff line). Where that fails, layered procedure takes over — managed access, cleared mixed teams, breach liability priced at accession.

Intrusive assessment is never compelled — a member may refuse it and pay in conservative scoring instead.

The residual leakage is priced openly as a membership cost, weighed against the non-subjugation guarantee no invitee can buy elsewhere.

(The full toolkit, layer by layer, is described in the VERIFICATION MODEL, §6.)

4.1.6 Efficiency flips

A capability that *steps* rather than climbs — the hidden efficiency flip — has its own three-layer answer:

1. The Council's common laboratories are aimed at exactly the directions able to produce one;
2. Sharing such an advance is made the profitable choice, commercial rights included;
3. The evaluation cadence tightens by published rule when field-level signals say a step is near.

(Terminology, held throughout: *common laboratories* are open to all Developers under Council patronage — the Council's own, releasing results to the Developers immediately after the decisive-advantage evaluation; *joint laboratories* are private ventures of some Developers — lawful cooperation, governed by the alliance rules and counted under the aggregation rule.)

(More info: the Convoy Papers — The Commons for the common labs, The Verification Model for the watch.)

The three-layer answer above does not resolve a large flip by itself. A flip that puts next-to-frontier capability within reach of fifty actors instead of five does two things at once: it makes a decisive advantage harder for anyone to hold — the central danger recedes — and it makes the coalition arithmetic harder to run: coalitions of fifty are not coalitions of five. The Council's response is procedural: verification re-indexes from the resource remnants that shrink to the algorithmic and talent layers that do not, formation accelerates rather than waits (F4), and the margin is recomputed. Also, many capable actors are usually many divergent ones, and

sufficient divergence makes a monolith's dominion impossible to establish and / or worthless to hold. Which is the Council's endgame.

4.2 Dark horses

Plainly: the surprise from outside is a lottery ticket — and the Council shortens the lottery's odds without buying a ticket itself.

4.2.1 The scenario

The classic dark horse is a wealthy criminal enterprise or a talented rogue collective building over the published frontier.

Their budget sits a magnitude or two below member scale: them reaching decisive advantage requires an efficiency miracle — a deep algorithmic unlock found by a small closed team and *missed by the combined search of the entire frontier*. Theoretically possible, but unlikely.

4.2.2 The residual

This chance is a lottery ticket — irreducible. However, the Council improves the odds:

1. Shared detection infrastructure.
2. On-ramps and Affiliate seats make joining cheaper than hiding, even for gray actors.
3. A staged release raises the common floor for everyone at once, so a dark horse gains no relative ground from it — its hidden stock depreciates as the public floor rises, and its relative advantage must be built privately, racing the combined velocity of a body it is not among.

4.2.3 The leader who leaves

The hardest dark horse is a Developer that concludes the Council costs more than it gives, and leaves, carrying insider knowledge of the watch-map. If it was a leader in development speed, it becomes an *outrunning* dark horse whose departure demands mobilization from those remaining.

Exit must stay lawful — a body that forbids leaving has become the monolith it exists to prevent. However, it is priced, and the pricing is automatic: dividend cutoff, due-process suspension of access to the Council's hosted services and common facilities

— never any claim on a member's own property — and watch-priority. What a member entrusts to common facilities remains its property, withdrawable on the published schedule.

The survivable margin and the pace instruments of §3.12 are designed for exactly this case.

Leaving the Council is public, and for a development speed leader it carries a high suspicion of imminent breakthrough and sudden jump to decisive advantage. So exiting the Council carries the same risk of becoming a target in a preemptive war as if the

Council didn’t exist.

4.3 The corporate quadrilemma

Plainly: no rival ever needs to see anything — and the leaders are compensated, not plundered.

4.3.1 The problem

A full transparency to rivals kills trade secrets. No transparency kills verification. Strong IP protection risks stagnation. Weak protection lets the richest leverage everything.

The solution to this is that *no rival ever needs to see anything*: the design is transparency to the audit organ, attestation to the members — like with a financial audit, where competitors submit their full books to third parties who publish only attestations. Standardized capability evaluations disclose *level* without disclosing *method*. Auditor capture is answered by rotation, multiplicity, and applying the Byzantine principle to the audit organ itself.

4.3.2 The balance-destruction scenario

In it, one member leverages resources until the margin breaks.

The solution to this is that the margin is a monitored ratio with rebalancing obligations, capital-adequacy-style. A member approaching the soundness bound triggers mandatory rebalancing among the rest *before* the bound is reached, not recriminations after.

4.3.3 Disincentive to share

Those ahead carry the heaviest disincentive to share know-how, and heavier still to share capability — the present race is led by exactly two corporations.

The solution to this is that transfers under §3.12 are preferably leases only, and in all cases compensated, never confiscated.

The compensations include (but are not limited to):

1. Licensing-pool royalties on fair terms
2. Contribution credits against progressive obligations
3. The non-subjugation guarantee itself, which is worth most to whoever has the most to lose.

The full compensation is a ledger of ten contractible venues, every one on scoreboard precedents — the sports-league caps that compounded franchise values, the disc patent pool built by the format war's own burned rivals, licensing as a profit engine at global scale. The ledger, attack-countered venue by venue, can be found in the Convoy Papers: *The Frontrunner's Choice*.

The compensation must be designed so that the benefit to the leader is as big or bigger than its losses, and the benefit to the trailing Developers from the leader(s) development exceeds their benefits from suppressing the leader(s), as far as the leader(s) don’t approach a decisive advantage.

The leader's choice is not between winning and safety: it is between winning and safety together on one side, and neither on the other. Played together, this game pays everyone more than played alone — and it pays the front of the pack first.

(More details: the Convoy Papers, The Frontrunner's Choice.)

4.4 Openness of the Council

Plainly: toward the outside, the Council is itself a monolith. So is the Developers chamber toward the other Council members. So every lead is caged in a window with a formula, a clock, and an escrow.

4.4.1 The danger and its bounds

A closed Council is, toward the outside world, a *collective* monolith — and this brief's own claims apply to it. So is the Developers chamber towards the Partners and the Affiliates. This necessitates providing openness, both of the Council towards those outside it, and of its chambers to one another.

The design answer upgrades "delayed openness" from a time-lapsed refusal into *bounded windows*. These perform mandated release of held advances after a safety-verification delay, so a lead is a moving window of bounded width, not an accumulating stock. The **outsider deficit** — how far the insiders' unreleased holdings put the outside world behind in AI capabilities — is then at most the window width, and the window is a chosen, monitorable and tunable design variable, set to keep the outsider deficit below the subjugation threshold.

Two bounded windows are defined:

The *outer window* sits between the Council's members and the non-members. It is the Council members' lead as a whole over the rest of the world, caged as above.

The *inner window* sits between the discovering Developer and all members. It is the discoverer's lead over the Council members, caged the same way. It delivers releases on published per-seat schedules, tuned so that no unsurmountable distance opens between the seats.

These schedules are created to defuse the desperation risk described in §1. As a side effect, they also ensure quick distribution of AI literacy and benefits from the AI development to everyone.

Both windows run on the same machinery — a clock and a stock cap, escrowed artifacts, schedules by content class — tuned separately.

Release from them is staged on a ladder of forms — announcement, method, evaluation access, hosted access, artifacts, full recipe — with hosted-before-possessed extending the instrumented-substrate principle outward. A narrow redaction class exists for capabilities whose diffusion is itself the catastrophe, under double exemption discipline; its size is a published indicator.

Release executes automatically from escrow — release-grade artifacts are deposited at capability attestation, verified by replication bounty then, secured by threshold cryptography (openable early only by quorum, automatically at expiry), so the holder's cooperation is consumed once, at deposit.

Outsider observers have access to the audit organ.

(More details: the Convoy Papers, The Outer Window.)

4.4.2 The outer window

4.4.2.1 Width

The width runs on a formula, in two steps.

The measurement: **B = tolerable outsider deficit ÷ its measured accumulation rate** — the tolerable bound.

The divisor is the inside–outside relative velocity plus the compounding term, not the frontier velocity. The distinction is stability: against frontier velocity the feedback runs away in both directions, while against the outsider-deficit rate it behaves as a window thermostat.

The policy: the binding window sits under a wall, **W ≤ B/2.** The safety factor of 2 is headroom the Council might spend into a discipline it must stand behind, hedging the linear form's optimism (next item), without causing intolerable outsider deficit.

W, the width, is the standard maximal delay: the world schedules release at W. Thus, outsiders can read the delay they experience.

Any lengthening requires amending W itself, publicly, under the half-bound wall; the named exceptions — T_verify and the redaction class — keep their own discipline, and everything else pushing past schedule is rent.

4.4.2.2 Withheld advances

Withheld advances compound each other, and the true outsider deficit is superlinear in the window. The linear form is a floor, not an estimate.

4.4.2.3 Content classes

Different content can be released on different schedules:

- safety-relevant results are released immediately: the public-health override of the trade regime is the standing precedent, extended by an aggregate override where cumulative license burden crosses what no single item does.
- capability-core advances are released at the window width
- commercial applications go through a fair and non-discriminating licensing pool (the standards bodies prove that competitors do license on fair terms where interoperability rents beat exclusion rents).

4.4.2.4 Minimal width

The width cannot be smaller than *T_verify*, the published minimum safety-vetting time.

This floor binds rarely today; it exists for when W becomes as short as a few days. The per-seat schedules (4.8) inherit it.

4.4.2.5 No privileges

The schedule is the same for all.

This bounds members too. External friendships are no release valve: outward flows run the outer schedules regardless of flag.

4.4.2.6 Width boundaries

The width is squeezed from both sides. Too large, and the Council becomes the singleton it exists to prevent. Too small, and membership loses its draw.

The resolution sits the width near the safety ceiling and compensates the joining incentive with nonrival compensations — benefits each member enjoys without diminishing any other's share. This strategy is stress-tested, not theorized — war-gamed at varying gaps, banking-supervision style.

(More details: the Convoy Papers, The Bounded Window.)

4.4.2.7 Item-wise

The clock runs per item from remnant-corroborated attestation (closing the shadow window of delayed declaration). Recomputation of the remaining time is formula-driven, on a cadence of at most a third of the window's width, shrinking with velocity.

Release means reproducibility within a declared cost, enforced by funded replication bounties — publication alone is not release.

4.4.2.8 Amending the formula

The formula runs automatically but is amended sovereignly:

B is measured, not voted — recomputed at every review from data the Council gathers.

W is voted, not measured — moved only by supermajority, always under the half-bound wall ($W \leq B/2$).

4.4.2.9 Price of deviations

Erring short is acceptable, erring long is not. A shorter window accelerates the outside frontier and is self-stabilizing. A longer one slows it and compounds the outsider deficit it creates.

4.4.2.10 Resistance

Economics has two and a half centuries of experience with this rule. The incentive to restrict competition scales with the rent (Smith's conspiracy-against-the-public observation, formalized since as rent-seeking and regulatory capture).

This both supports the demand for a short, dynamic window and predicts who will push against it. Defection pressure is a force, not a guarantee.

(More details: the Convoy Papers, The Bounded Window.)

4.4.3 The inner window

Defined by two triggers — a clock, and a stock cap.

Each Developer's withheld-stock potency runs on a priced ledger against the margin's headroom. Overflow spills forward, oldest and most potent first. Pricing is provisional at capability attestation and trued-up at release, so no mispricing is permanent.

Compensation is engineered, not hoped for:

- priority credit convertible through published channels
- a Council Award on the prestige-economy precedent
- the first-mover's execution lead

Applications stay uncoupled — Developers compete freely at the product layer (the open-source world proves the model's viability), and commercial performance, good or bad, doesn't affect the window.

The whole apparatus is implementable as ordinary intellectual property under binding covenants — the FRAND case law is litigated proof that such covenants hold.

If Developers conspire to establish decisive advantage over Partners — the openness imposed on Developers allows Partners to watch for this. If it is detected or reasonably suspected, Partners can convey this to their leaders, who can consider taking out-of-Council adversarial measures of appropriate severity.

(More details: the Convoy Papers, The Inner Window.)

4.4.4 Precedent: nuclear custodianship

The precedent holds for the *custodianship*. A tolerated gapped oligarchy is a structure the world has run before. And it eventually recognized the benefits in limiting the run for decisive advantage and implemented measures that prevent it.

Its known costs are:

- resentment
- proliferation pressure
- entrenchment of the recognized few

This design carries the same costs. Their prices must be paid and monitored — the alternative, an uncorrected singleton auction, costs more.

Where the precedent breaks: it does not hold for the *deterrence stability*. The nuclear peace rested on a physical impossibility of preventing a counterstrike. A decisive intelligence advantage abolishes the AI counterstrike. Physics stabilized the nuclear regime for free — this design must engineer its stabilizer. This is the reason for the survivable margin above.

4.5 The coalition problem

Plainly: a bloc is one member of bloc size, in the capabilities that are pooled — so blocs are counted, and secrecy, not alliance, is the crime.

4.5.1 The problem

The fault-tolerance margin assumes independent members. Against a large enough bloc the 2:1 arithmetic may be unreachable. And a candidate may refuse to join precisely from the fear that others will secretly pool against it.

4.5.2 The answers

In declining strength:

1. **Capability-sharing stands counted.** Declared AI development capability-sharing arrangements — joint training programs, model-weight sharing, compute pooling —

are counted jointly for appropriate margin purposes. Alliances that do not share this capability are not.

1. **Covert coordination is detectable in the same currency as everything else**: correlated moves correlate, and joint action leaves joint remnants.

2. **The claims of this brief are actor-size-invariant** — with limits stated: a bloc that wins the singleton auction becomes the collective singleton, and inherits §2.1.3 –

§2.1.6 whole. The corruption loop, the obsolescence rate, and the self-starvation results do not care how many flags fly over the building. A theorem deters the rational-instrumental actor, but the dangerous actor is the one that is not rational: for those who price power above ruin, only the margin itself deters. The theorem's real function is coalition glue — a shared, priced reason for the non-obsessed to hold the line together against the obsessed.

4.5.3 The residual

Named as this brief's fifth hard problem: covert coalition formation is the Council's hardest attack surface, mitigated but not closed.

4.5.4 Details

1. Alliances stay legal — members ally as they always have — with capability sharing *declared and counted transitively*: the sharing graph's connected components are the units the 2:1 arithmetic sees, on the aggregation pattern sanctions law already uses.

2. Alliance is not crime. Secrecy is.

3. Inner groups pass an open-door test — published criteria, transparent terms, counted where capability-relevant — separating infrastructure from bloc.

4. Present leaders, host states and allies may arrive already forming a component near the bound — so accession separates governance integration from capability integration, or sizes the table accordingly.

(More details: the Convoy Papers, Alliances Without Blocs.)

4.6 Frictions of matter, not knowledge

Plainly: three fears that will deter invitees — hardware, talent, sabotage.

4.6.1 Hardware starvation

This fear is listed here deliberately, as a deterrent the recruitment conversation must expect.

During compute scarcity, members who control hardware production could use what openness teaches them to starve rivals of supply.

Instruments against this:

1. Neutrality obligations:
    1. Attached to **whichever supply-chain link currently binds** — the bottleneck moves, and a rule married to one link dies with it.
    2. Member-controlled links run on published fair terms under the essential-facilities logic, through negotiated assurance agreements.
    3. Neutrality is proven by outcomes, not org charts, with the burden shifting to the allocator on deviation.
    4. State-controlled links carry a most-favored-member undertaking for Council-instrumented transfers — precisely the verified-end-use case export regimes claim to exist for.
    5. Non-member chokepoint holders are engaged as observer-suppliers through the Council's service as purchase aggregator for its members — neutrality covenants, notice-and-standstill among them, traded for the aggregated demand. (The living chokepoint map is a founding deliverable of the Council — a concrete early task.)
1. Starvation is treated as an attack to deter, not a shortage to stockpile against:
    1. Four latches: notice-and-standstill, the denier's self-suspending dividend, a credit facility in money rather than machines, and declared-scarcity fair-share by published formula.
    2. A mutual-aid option book for the one *uncorrelated* emergency, facility sabotage.

3. No standing compute pool is kept: a reserve is insurance whose fund evaporates in exactly the correlated shortage it insures.

1. Runaway concentration of matter already is countered by the anti-divergence mechanism:

    1. A member whose *acquisitions* diverge past its published, self-computable allowance meets a graduated throttle on further buying. Dormant while nobody diverges, automatic when someone does, symmetric for all, never touching holdings.

    2. The monitor carries a published estimate of the leading outside actor, and if that frontier closes in, the throttle eases — the Council never brakes its own while a stranger sprints past.

    3. Efficiency advances release on the fastest schedule of all classes, because diffusing efficiency is itself anti-concentration.

(More details: the Convoy Papers, The Reference Algorithm.) (More details: the Convoy Papers, Hardware Neutrality.)

4.6.2 Talent

Openness reveals whose people matter, and invitees will fear raiding. Instruments against this:

1. Obligations attach to a cargo (knowledge and proficiency linked to advances within a bounded window), not to a person. After the cargo expires (the advance descends the release ladder to the cargo-concerning rung), or if the person is moving where this cargo is present, or where this cargo will not be used, the person is free to move. (As per the minimum bars of §3.4.)

2. Invitees and members must know that their cargo is protected, but they do not have property rights beyond it and over persons, even if established in an apparently lawful way.

3. Mobility bans are refused for two different reasons: they are prosecuted cartel conduct, and they are self-defeating — carrier mobility is how tacit knowledge

diffuses, and a council that freezes its people freezes the cooperation it exists to enable.

1. The rule can be introduced as an accession descent schedule rather than a day-one demand. Member barriers can be negotiated down on a published schedule indexed to verified trust and the shrinking inner window. (Since nearly every candidate restricts today and a hard immediate condition might be rejected.)

2. The operational instruments are:

    1. aggregate-flow transparency with the burden on asymmetry's beneficiary, paid in nonrival information

    2. cargo-indexed leave, the clock travelling with the cargo

3. team-lifts audited as events, sanctions on the hiring member and never the hired
4. shared pipelines that grow the pool instead of fighting over it
5. multi-person control of the most dangerous items, so no one head carries a whole crown jewel
6. An opt-in open-movement compact rewards certified free movement with deeper cooperation, joined in graduated steps.
7. The lone defector with a head full of secrets remains the free design's priced residual: every regime that "solved" him built the walls this Council refuses.

(More details: the Convoy Papers, Cargo, Not Person.)

4.6.3 Sabotage

The record of stopping rival programs by force — facility strikes, sabotage operations, the assassination of program scientists — is long and documented.

1. As AI capability approaches decisive relevance, frontier laboratories inherit the threat model of nuclear facilities: their research facilities, their datacenters, their people, and their entire supply chains. Corporate invitees who file this under "the state's job" need to understand that before they experience it.
2. Under the Council the threat is expected to *decrease*: prohibited among members at the sanctions ladder's top; devalued (where advances diffuse on a bounded window, destroying a rival's copy buys little); and answered collectively — a member attacked is the Council attacked, which makes collective attribution and response the membership dividend's hardest-currency component.

This belongs in the recruitment conversation too.

The brief claims the problems' shape, not their absence.

4.7 The unconvincible member

Plainly: some invitees will come not to be safe, but to stay dangerous.

An actor may combine great destructive capacity in older domains, modest constructive capacity at the AI frontier, and a leadership for whom restraint reads as weakness and benevolent reciprocity as treason. Such an actor might not seek the reassurance the Council offers — that it will never be subjugated. It might seek the opposite assurance: to remain able, or to grow more able, to subjugate others — if possible, everybody.

What such an actor receives defuses its desperation; whatever more it hopes for must run the same windows, conduct grounds, and trust ladders that bind every member alike.

For the actor whom only capability convinces, the deterrent is kept legible — published margin arithmetic, enforcement machinery exercised in the open — so that "this war cannot be won" can be read rather than taken on faith. The lawful door stays open at every stage, so the choice

offered is never humiliation-or-war. What rules cannot contain, statecraft outside the Council must: the Council governs a technology, and its whole ambition against this hardest case is to bound a harm no structure can abolish.

4.8 Example window schedules

The dates are absolute anchors. Each tier's date runs from the item's registration, as a fraction of W. Early opening to a tier — by compensation, correction, the exchange, or choice — advances that tier only; every later tier keeps its anchor, and the world's date is never contingent on club-internal events.

The following schedules can be used during the period before the Council has determined its own.

W is the binding outer window width — the standard maximal delay, under the half-bound wall of 4.4.2.1 — set by the Council and proposed here initially as 45 days (B measured at 90). The schedules spend the whole of W: the safety factor lives upstream, in the half-bound wall.

Content class	Developers	Affiliates	Partners	World
Safety-relevant	0	0	0	0
Efficiency	0	W/3 (15d)	W/3 (15d)	2W/3 (30d)
Capability-core	W/3 (15d)	2W/3 (30d)	2W/3 (30d)	max(W, T_verify) (45d)
Common-lab results	0 (post gate)	0	0	standard outer
Commercial	licensing pool at readiness, all seats			

5 The endgame

Plainly: suppose that either the Council succeeds, or control itself stops working before that. What next?

5.1 Dissolution by arrival

All or most of the leading Council members will eventually reach a level of intelligence that makes the deficiencies of the monolith-headed societies and the benefits of the Constellation structure too obvious to ignore. This will ensure their resistance against establishing a monolith-led civilization, including taking measures against all kinds of decisive advantage. If the Council is still needed then in some supplementary or secondary role, it will continue to exist in this role; otherwise its sunset clause will be executed.

Also, the development of entities in different directions can create:

- natural bounds that make preventing a counterstrike impossible
- degree of divergence that makes subjugation impossible
- degree of divergence that makes subjugation meaningless (by removing its benefits)

If any of these is achieved, it will naturally eliminate or decrease below a threshold the singleton / monolith danger. This will either obsolete the Council and cause its sunset, or will downgrade its function according to the then-needs.

When this happens, the members of the Council will become the Constellation, and the Council will sunset.

1. The residue worth naming: a generation of practiced mechanisms — bounded windows, pooled verification, margins held by rule, trust accumulated between rivals — that generalize past AI to knowledge as such. The Council's legacy, in success, is the seed crystal of the structure after it: competition at every frontier, cooperation behind each.
2. Status: the design's stated hope — now with a test instead of only hope: the Council dissolves when its removal would change no incentives, and the burden inverts against self-perpetuation — scheduled referenda where the Council must justify continuation, and silence means sunset.

The clock, and the two additions the decomposition forces:

1. One quantity on this list is computable rather than debatable: the endgame threshold — call it the *Continuity Limit*. The velocity at which detection-plus-reaction time exceeds sprint time is derivable from the growth dynamics of the companion program. So the Council can know the date of the question even before anyone can defend an answer.
2. One term of that date is chosen, not imposed: the human-decision floor on the upper rungs advances it — safety now, an earlier threshold later — a price re-chosen at every review with eyes open.
3. The invariants: across every option, unconditionally — the non-subjugation guarantee on both faces, the freedom floor, the minimum bars, the diversity reserve (simple and divergent lines kept alive, not archived), and the record itself. An endgame that fails an invariant is not an option but a failure the Council exists to prevent.

One consequence is stated and not argued from. A body so constructed — short windows, a voice for those outside the race — produces, as a side effect, AI development for all: arguably the largest common good humankind will have received within so narrow a window of decades in its history. This is not the goal of this Council, or a motivation for its existence and operation – it is a side effect of its work.

A negative side effect is possible too: if competition at the frontier is too suppressed, it may weaken or fall below its optimum. Due to this, suppressing it is acceptable only where the risk and harm of a monolith's creation substantially outweigh this deficiency's harm.

5.2 Pre-committed transitions

It is possible that none of these comes before the Council has carried the transit that far — the Continuity Limit, the development velocity at which control stops working. There, a decisive advantage can be acquired faster than the others can detect and react, every venue for increasing audit and reaction speed is exhausted, and the margin is arithmetic that can no longer be enforced in time. What measures can be adopted then?

The worst Council failure is crossing this threshold with no plan. So, trigger levels on the published margin-to-threshold indicator must be defined. If reached, these must activate pre-agreed moves at founding, resolution-planning style, with a pre-set default that is the least-irreversible package. The goal of these moves is to buy time for coming to an arrival point, and / or to increase the control over the development commensurately with the increase of the danger.

A list of possible moves follows:

5.2.1 Speed governor

In this variant, the Council collectively caps frontier velocity.

1. Status: re-scoped by decomposition. Post-threshold it is self-undermining — a cap on speed must be enforced at a speed greater than the speed it caps.

2. Said plainly: delaying is not a good option — it taxes every benefit the Council exists to deliver — but conditions can make it the only one. And even then it is permissible only while the chance of a dark horse sprinting ahead is negligible.

5.2.2 Referee transfer

In this variant, the Council's functions (monitoring, margin, response) pass to open-structure AI systems verifiably constructed to favor and wall off no participant, operating at the frontier's own speed.

1. Status: the hardest engineering on the list — but a ladder already being climbed, since monitoring, arithmetic, and lower rungs are mechanism today; the endgame step transfers the upper rungs, per-function, shadow-then-advisory-then-binding.

2. "Verifiably non-favoring" decomposes into process even where its science stays open:

    1. threshold-held keys (no single member controls the referee)

    2. a verified training run as its auditable birth

    3. adversarial testing by every member

    4. the referee held outside the race — capability-frozen between jointly reviewed, windowed upgrades.

1. Referee functions split into a frozen class (adjudication-critical, changed only by reviewed upgrade) and a learning class. The latter is under glass-box transparency

— an open-access mind as a requirement class, narrow by precondition.

1. Two build principles:
    1. narrow referee (minimal capability for the function — verification and response, no general agency)
    2. plural referees (independently built from different members' stacks, under Byzantine agreement among themselves — the Council's own margin logic applied to its successors, diversity as the integrity mechanism).

5.2.3 Development transfer

In this variant, all AI development is transferred to open-structure AI system(s), verifiably constructed to favor and wall off no participant.

1. Status: requires AI systems that are both very powerful and completely open – currently not achievable (current generation AIs are black-boxes).
2. Risks: if not supplied with sufficient diversity, can create a development monoculture and limit the development directions and achievements.
3. The more and more different the development systems are, the bigger the potential for diversity – but not necessarily the achieved one.
4. The capability to understand and implement the knowledge produced by the development systems might become a leadership factor, also able to create a singleton. Preventing that might require taking appropriate measures, possibly modeled to a degree after the ones described here.

5.2.4 Radical diffusion

In this variant, the bounded window is pushed toward zero and capability is spread so widely that no decisive gap is geometrically achievable by anyone: safety by the elimination of the prize.

1. Status: the limiting case of the window formula — and half-achievable by construction: knowledge diffuses as W falls by arithmetic, while matter re-concentrates under increasing returns. This endgame requires the hardware machinery (neutrality-and-standstill, essential facilities, demand aggregation) as permanent institutions, and inherits their sovereignty residual at full weight.
2. Problems:
    1. Requires the goodwill of all major Developers.
    2. Short-term solution – after that, someone can accrue a decisive advantage again.

(More details: the Convoy Papers, The Council Endgame.)

6 Related work, and what this brief adds

The proposal joins a live literature, and stands on a broad one.

1. The warning itself — that frontier AI carries risks up to and including irreversible loss of control — is by now the published consensus of much of the field's founding generation: the Science position paper of Bengio, Hinton, Yao, Song, Russell, Kahneman, Harari, and twenty co-authors [Bengio et al. 2024], and the International AI Safety Report that followed under Bengio's chairmanship [Bengio et al. 2025].

2. Book-length treatments have mapped the same ground from complementary angles: Russell on the control problem [Russell 2019], Ord on existential risk accounting [Ord 2020], Suleyman on the containment problem [Suleyman 2023], Kissinger, Schmidt and Huttenlocher on the strategic order [Kissinger et al. 2021]; Hogarth's widely read essay named the race dynamics plainly [Hogarth 2023], and Dafoe's research agenda organized the governance field this brief writes into [Dafoe 2018].

3. On the brief's two older foundations: the desperation logic of §1 stands on the power-transition and preventive-war literature [Organski 1958; Levy 1987; Allison 2017], and the window economics of §4.4 on the rent-seeking and cartel record [Smith 1776; Tullock 1967; Stigler 1964, 1971; Levenstein & Suslow 2006].

Against that shared background, mechanism proposals divide as follows:

1. Dario Amodei's *The Adolescence of Technology* (January 2026) warns at length about the concentration of power that frontier AI produces, and his *Policy on the AI Exponential* argues that AI will soon be too capable to be safely entrusted to either governments or companies, requiring checks and balances on each — with mechanisms such as Anthropic's Long-Term Benefit Trust named as structures the industry should go beyond, alongside mandatory third-party testing and government authority over unsafe deployments [Amodei 2026b; 2026a]. These are *institutional and legal* checks; they do not contain a standing body of the rivals themselves with a capability-based soundness margin, an adverse-selection argument about who wins concentration contests, or a priced compliance calculus. (His earlier *Machines of Loving Grace* [Amodei 2024] proposed a democracies-first "entente" — deliberately *asymmetric*, excluding the hostile, where this brief requires their inclusion.)

2. OpenAI's *Governance of superintelligence* [Altman, Brockman & Sutskever 2023] proposed IAEA-style international oversight of efforts above a capability threshold — an inspectorate *over* the racers, not a body *of* them.

3. The maximal proposal, MAGIC [Hausenloy, Miotti & Dennis 2023], would centralize frontier development into a single exclusive multinational facility under a global compute moratorium — the closest proposal in ambition and the most instructive contrast: a benevolent monopoly is itself a singleton-shaped object, and Claims two through five of this brief apply to it.

4. The deterrence family — the Mutual Assured AI Malfunction framework of *Superintelligence Strategy* [Hendrycks, Schmidt & Wang 2025] and hybrids building on

it — shares the Council's deterrence logic but is state-centric, largely compute-centralizing, and offers deterrence without a membership dividend.

5. The moratorium argument of Roussel, Lauwaert, Swoboda, Ramsey, Uuk, Dung and Aguirre [Roussel et al. 2026], building on the case against racing of Dung and Hellrigel-Holderbaum [Dung & Hellrigel-Holderbaum 2025], is this brief's nearest-premise neighbor: it derives from states' self-interest alone that a halt on superintelligence can be individually rational. The premise is shared; the conclusions differ in what self-interest is driven *toward* — a stop there, a structure here. The two are complementary rather than rival: a moratorium that holds still faces the question of what the parties do inside it, and a council of rivals is one answer that does not require the halt to be permanent.

A separate strand measures rather than proposes. Formal race models, from Armstrong, Bostrom and Shulman onward [Armstrong et al. 2016] and through the evolutionary-game treatments of Han and colleagues [Han et al. 2020], establish when competition erodes safety. The first controlled behavioural experiment on an idealised AI race [Fernández Domingos & Han 2026] finds that unsafe development is driven chiefly by falling behind — not by the level of risk, and not by participants' risk preferences. That result is the Gambler's Discount and the desperation logic of the Introduction confirmed: falling behind cuts care. This brief takes that finding as its premise: it proposes a structure that makes falling behind survivable, so that the measured reflex has nothing to act on.

Note: The researchers who fill this section and its references are also the people a body of this design would protect first and best — and bodies get built when those who benefit ask for them.

7 What would change our mind

1. **(F1)** If institutional screening demonstrably strengthens, rather than weakens, as prize size grows, the Tyrant's Premium loses its bite.

2. **(F2)** If safety practice demonstrably speeds rather than slows frontier development

— if care is productivity-positive at the frontier — the Gambler's Discount loses its bite.

1. **(F3)** If verification costs at frontier scale exceed the diversity dividend, the Council is priced out and the mechanism must change.

2. **(F4)** If unilateral capability can grow faster than coalitions can aggregate — if the race's timescale beats the assembly's — then the Council is the right structure arriving too late, and the priority shifts from designing it to accelerating its formation; note that this falsifier, if true, makes the brief *more* urgent, not less.

3. **(F5)** If a decisive advantage turns out not to be achievable at all, the brief reduces to ordinary multipolar governance and loses its special urgency — a happy failure.

4. **(F6)** If a single adjudicator can scale its comprehension without bound — tracking a fast world as cheaply as the world changes — §2.1.4 weakens. The known escape is self-

closing: the way to scale adjudication is to absorb more adjudicators, which is precisely the monolith move whose price is already established (§2.1.6).

8 Claim tiers

1. **Proved elsewhere and cited**: the dominance trap, suppression backfire, the convoy comparison, the physical pricing of detection and concealment.
2. **Historical regularity with mechanism**: the corruption loop of §2.1.3.
3. **Derived from the companions**: the obsolescence rate of §2.1.4.
4. **Modeled here at sketch grade**: the Tyrant's Premium and the Gambler's Discount (auction argument; full treatment queued in the pricing program).
5. **Proposed as engineering**: the Council structure, the fault-tolerance margin, the specification of §3, the recruitment order, and the instruments of §4 (Liebig-indexed monitoring, capability attestation rather than disclosure, the margin ratio, the bounded window).
6. **Posed with stated residuals**: the seven hard problems of §4, and the endgame menu of §5.2 (statuses attached per item).

Glossary

Some terms, as used in this paper.

Terms marked (companion framework) come from the companion papers' framework.

- **The accession descent** — the published loosening path for members who restrict their people today.
- **The aggregation rule** — an alliance that shares a capability concerned by a voting or a mechanism counts for the purpose of this voting or mechanism as one collective actor.
- **The bounded window** — the release-delay regime. Its **width (W)** is the mandated delay between an advance's attestation and its release — a moving lead of bounded width, never an accumulating stock. The outer window runs toward the world; the inner window runs discoverer-to-members, shorter, on each seat's schedule.
- **Capability attestation** — registering an advance with the body's ledger, with a confidential demonstration that prices it; starts the clock; corroborated by physical remnants against backdating.
- **Cargo, not person** — the talent rule: obligations attach to the confidential items a person carries, never to the person.

- **Common labs** — research facilities open to all Developers and Council organs, under Council patronage; plural and distributed; inner window zero for their discoveries; the practiced form of the endgame's hope.
- **The Continuity Limit** — the development velocity at which detection-plus-reaction time exceeds sprint time: past it, a decisive advantage can be acquired faster than others can detect and react, and the margin becomes arithmetic that can no longer be enforced in time. Computable from the companion program's growth dynamics; developed in a forthcoming companion paper.
- **The Convoy Papers** — the companion series developing each mechanism at full depth; Nos. 1–2 published, further numbers in preparation.
- **The Constellation** — the design's intended end-state: the members as a plurality of diverse, mutually visible, mutually bounded centers of development, none able or needing to absorb the rest, most or all understanding that the benefits of the diversity outweigh these of the absolute power.
- **Convoy** (*companion framework*) — an ensemble of diverse, communicating, semi-independent developers whose combined inflow and error-correction outgrow any single absorber; the design's target structure, multi-directional at civilizational scale.
- **Dark horse** — a capable developer outside the body's sight.
- **Decisive advantage** — capability sufficient to prevent every other actor from ever catching up, and potentially to exercise absolute power over them.
- **Demand aggregation** — the Council as one negotiating customer: members' hardware purchasing pooled into a single order book, whose scale buys terms no member gets alone — neutrality covenants, notice-and-standstill, priority assurances. The consideration that makes supplier obligations contractual, never commanded.
- **The efficiency flip** — the regime change when compute-per-capability falls far enough that frontier-grade development stops requiring frontier-scale resources. Footprints shrink, and dark horses multiply. The verification design must survive this crossing. (Its answers: the flip-defense stack.)
- **Escrow with threshold release** — release-ready materials deposited at capability attestation, split cryptographically among members, opening automatically at expiry.
- **The fault-tolerance margin** — members remaining compliant must jointly hold more than twice the capability of any possible defector coalition, held severally and in survivable form, so the body survives any minority turning adversarial. Distributed-systems engineering calls the underlying requirement *Byzantine fault tolerance*, after the Byzantine generals problem — commanders who must coordinate while some may be traitors. (Its short form, "Byzantine," can be typically met in this sense in compounds — Byzantine tolerance, Byzantine agreement, Byzantine margin.)

- **Fingerprint** — the identifying pattern within a *remnant channel* (see below): behavioral and deployment signatures that reveal capability level from use.
- **The flip-defense stack** — the answer to a hidden *efficiency flip*: discover first (common labs aimed at the flip-prone directions), share when found (commercial rights for disclosure), bound the payoff (one-cycle containment under hazard-indexed cadence).
- **Footprint** — the aggregate observable magnitude of an actor's development: the sum of its *remnants*. Efficiency advances can shrink it, but frontier scale cannot hide it.
- **The frontrunner's choice** — the frontrunner mechanism's stack-holder: *the leader's choice is not between winning and safety; it is between bigger wins and safety together on one side — and neither on the other.*
- **The Gambler's Discount** — the brief's second selection claim: the race sells care at a discount, so those who forgo the most safety move fastest, the unpriced catastrophe risk being exactly their rebate. Catastrophe by error, where the Tyrant's is by intent — and without the Tyrant's self-preservation floor.
- **The half-bound wall** — the constraint $W \leq B/2$: the window's width may never exceed half the window bound, so even a maximally withholding member remains a full reaction-time short of decisive advantage.
- **The Library** — the common training corpus (tier 1), the civilization's heritage (tier 2), and, conditionally, the consentful bequest of the dying (tier 3).
- **The margin throttle** — the fault-tolerance margin's operational face: a graduated, self-computed drag on further *acquisition* for any member diverging above its published allowance — individual and continuous, never rank-based; engaged early; symmetric; dormant until divergence appears; never touching holdings; outside-aware — pre-committed easing when the estimated outside frontier closes within a published distance, the throttle subordinate to collective survival.
- **Margin-to-threshold indicator** — the published, continuously measured distance to the day when detection-plus-reaction time will lose to sprint time;.
- **The maturation ladder** — how every new binding mechanism deploys gradually: shadow → advisory → binding-with-exceptions → binding, exceptions sunsetting on the accuracy record.
- **Mechanism before tribunal** — a Council-wide meta-principle: wherever a transparent algorithm answers instantly and beyond doubt, it does; the residual and the amendments are decided by tribunals, not automatically.
- **Member classes** — *Developers* build frontier AI at admission scale; *Partners* hold significant capability, AI or other, without AI frontier at admission scale; *Affiliates* are verified-conduct AI developers below the Developers and Partners admission bars. Each class holds its own chamber, seat terms, and schedules.

- **Monolith** (*companion framework*) — a civilization or structure under one absolute controller.
- **The mutual-aid option book** — a narrow surviving capacity commitment: members' pre-priced call options covering facility sabotage — the one uncorrelated emergency, where spare capacity exists precisely because the risk is uncorrelated.
- **Narrow referee / plural referees** — AIs, created for quickly deciding complex problems in situations when reaction time must be shorter than what humans can achieve. Have the minimal capability for the function, and no general agency. Envisaged to have several independently built, agreeing by Byzantine vote — diversity as integrity.
- **No-single-head rule** — the most dangerous items and responsible posts must be held under multi-person control, so no one head carries a whole crown jewel and there is nothing walkable to buy.
- **Nonrival compensations** — membership benefits whose enjoyment by one member does not diminish their availability to any other — the non-subjugation guarantee, verification at cost, arbitration, legitimacy; the currency that compensates where capability must not flow. (Know-how is nonrival in use but positionally rival — it travels the windows, never this channel.)
- **Notice-and-standstill** — the anti-starvation rule replacing the retired compute reserve: a supply cutoff requires advance notice and continued supply through adjudication from the supplier's own capacity — the fast cutoff outlawed rather than out-run; backed by the denier's dividend self-suspension and by the **credit facility** (money, not machines: the body fronts the victim's market purchases as a loan billed to the denier). No standing pool is kept — insurance whose fund evaporates in the correlated shortage it insures. (A covenant term, bought with the Council's aggregated demand — see Demand aggregation — never a command to outsiders.)
- **Open-access mind** — an AI (retiree, external AI development...) whose ruling-determining state is readable at the finest granularity its substrate permits,

formation history included. Meant to be used where very fast action and / or complete reliability and transparency is required.

- **The open-door test** — a rule that distinguishes whether a group is a bloc or merely an infrastructure. Inner groups with published criteria, transparent terms, and counted capability are infrastructure; closed circles with private terms are blocs.
- **Open-movement compact** — the opt-in inner tier rewarding certified free movement with deeper cooperation; joined in graduated steps.
- **The outsider deficit** — how far the insiders' unreleased holdings put the outside world behind in AI capabilities; superlinear in the window, because withheld advances compound.

- **Potency** — an advance's priced capability contribution — what its release would add to the holder's measured position; the unit in which the ledger, the stock cap, and the margin all count.
- **The reciprocity condition (no one-way valves)** — a rule stating that a member blocking AI know-how transfer to other members is ineligible, in mirror scope, to receive know-how developed without its participation. The exclusion extent is defined by the member's own law. The ineligibility to receive lifts the day the this block is removed. A bloc-scale blocker exits the rule into the divergence machinery.
- **The recognition menu** — what "appropriate recognition" includes, defined per member type: commercial rights for the commercial, publication primacy for the academic, mission credit for the public centre, priority for all. Assembled per case in the open.
- **Redaction class** — a narrow, supermajority-gated, sunset-claused exception for capabilities whose diffusion is itself the catastrophe. Its size is a published indicator.
- **The Register** — the standing ledger of engaged non-members. Registration confers standing, not membership. It is the lawful door's waiting room, with its own duties and protections.
- **The release ladder** — the staged forms of releasing an advance to an audience: announcement → method → evaluation access → hosted access → artifacts → full recipe.
- **Remnant** — an individual physical trace-channel that a frontier AI development cannot avoid: energy and cooling, supply chains, talent flows, behavioral fingerprints... Each of these is priced against its concealer, together the verification portfolio. Remnants compose an actor's *footprint*.
- **Replication bounty** — funded independent reproduction; the enforcement of "released means rebuildable at declared cost."
- **The sanctions ladder** — the graduated conduct-enforcement sequence (§3.7): pre-committed rungs from audit demand through dividend suspension to referral beyond the Council, each priced in advance. In effect, it is a deterrence by legible arithmetic. The upper rungs are always under human decision.
- **Singleton** — an actor holding decisive advantage; the future of the civilization within one discretion (after Bostrom).
- **Spill-forward** — the stock cap's enforcement: overflow releases automatically, oldest and most potent first — no hearing, no discretion, no exception.
- **The stock cap** — the holding limit: no member's unshared stock of advances may approach the margin's headroom, however lawfully accumulated. (Enforced by *spill-forward*.)

- **Survivable margin** — the margin held in dispersed, hardened, escrowed form, so no single strike can decapitate it. Second-strike stability, engineered rather than gifted by physics.
- **The Tyrant's Premium** — the brief's singleton-selection claim: contests for absolute power are auctions the power-obsessed disproportionately win, because for them the prize's value never saturates.
- **Velocity margin** — the pace analog of the fault-tolerance margin: no member's measured frontier velocity must exceed a fixed multiple of the Council's median.
- **The window bound (B)** — the maximum tolerable capability lead, measured in days of frontier time (90 in the founding formula); the level whose crossing the whole apparatus exists to prevent.
- **The window thermostat** — the window formula's stabilizing property: width set against the outsider deficit's growth rate, so shortening accelerates the outside and relaxes itself. Left to run with no endgame decision taken, it releases everything on schedule — the endgame's default, needing no vote.

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.

References

Allison, G. (2017). *Destined for War: Can America and China Escape Thucydides's Trap?* Houghton Mifflin Harcourt.

Altman, S., Brockman, G., & Sutskever, I. (2023). Governance of superintelligence. OpenAI, May 2023.

Amodei, D. (2024). Machines of Loving Grace. October 2024.

Amodei, D. (2026a). Policy on the AI Exponential. darioamodei.com, 2026. Amodei, D. (2026b). The Adolescence of Technology. January 2026.

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the Precipice: A Model of Artificial Intelligence Development. *AI & Society*, 31(2), 201–206.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., ... Kahneman, D., et al. (2024). Managing extreme AI risks amid rapid progress. *Science* 384(6698), 842–845. doi:10.1126/science.adn0117.

Bengio, Y., Mindermann, S., Privitera, D., et al. (2025). International AI Safety Report. arXiv:2501.17805.

Bostrom, N. (2006). What is a Singleton? *Linguistic and Philosophical Investigations*

5(2), 48–54.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Dafoe, A. (2018). AI Governance: A Research Agenda. Centre for the Governance of AI, University of Oxford.

Dung, L., & Hellrigel-Holderbaum, M. (2025). Against Racing to AGI: Cooperation, Deterrence, and Catastrophic Risks. arXiv:2507.21839.

Fernández Domingos, E., & Han, T. A. (2026). Falling Behind Drives Unsafe Development in an Idealised AI Race Experiment. arXiv:2607.26034.

Finnveden, L., Riedel, J., & Shulman, C. (2022). AGI and Lock-In. Working paper, Forethought Foundation for Global Priorities Research.

Han, T. A., Pereira, L. M., Lenaerts, T., & Santos, F. C. (2020). Mediating Artificial Intelligence Developments through Negative and Positive Incentives. *PLoS ONE*, 16(1), e0244592.

Hausenloy, J., Miotti, A., & Dennis, C. (2023). Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI. arXiv:2310.09217.

Hendrycks, D., Schmidt, E., & Wang, A. (2025). Superintelligence Strategy: Expert Version. arXiv:2503.05628.

Hogarth, I. (2023). We must slow down the race to God-like AI. *Financial Times*, April 2023.

Kissinger, H., Schmidt, E., & Huttenlocher, D. (2021). *The Age of AI: And Our Human Future.* Little, Brown.

Levenstein, M., & Suslow, V. (2006). What Determines Cartel Success? *Journal of Economic Literature* 44(1), 43–95.

Levy, J. (1987). Declining Power and the Preventive Motivation for War. *World Politics*, 40(1), 82–107.

MacAskill, W. (2022). *What We Owe the Future.* Basic Books. Ch. 4, “Value Lock-In.”

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* Bloomsbury.

Organski, A. F. K. (1958). *World Politics.* Knopf.

Roussel, E., Lauwaert, L., Swoboda, T., Ramsey, G., Uuk, R., Dung, L., & Aguirre, A. (2026). Are We Doomed to an AI Race? Why Self-Interest Could Drive Countries Towards a Moratorium on Superintelligence. arXiv:2605.01297.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations.* (Book I, ch. X.)

Stigler, G. (1964). A Theory of Oligopoly. *Journal of Political Economy* 72(1), 44–61.

Stigler, G. (1971). The Theory of Economic Regulation. *Bell Journal of Economics* 2(1), 3–21.

Stremsky, A. (2026a). The Physical Prices of Knowledge: To Learn Is to Forget. *The Convoy Papers*, No. 1. Zenodo. https://doi.org/10.5281/zenodo.21710817

Stremsky, A. (2026b). Why the Winner Starves: Epistemic Starvation in Gap-Defending Knowledge Hierarchies. *The Convoy Papers*, No. 2. Zenodo. https://doi.org/10.5281/zenodo.21710084

Suleyman, M. (2023). *The Coming Wave: Technology, Power, and the Twenty-first Century's Greatest Dilemma.* Crown.

Tullock, G. (1967). The Welfare Costs of Tariffs, Monopolies, and Theft. *Western Economic Journal* 5(3), 224–232.

Precedent sources. The named precedents throughout §§3–5 rest on primary sources, among them:

- the FRAND case law — Huawei Technologies v. ZTE, Case C-170/13, ECLI:EU:C:2015:477 (CJEU 2015) and Microsoft v. Motorola (9th Cir., No. 14-35393)
- UN Charter Articles 102 and 35(2)
- the Treaty on the Non-Proliferation of Nuclear Weapons, Article IV
- the EU AI Act, Articles 51–52 (the 10^{25} FLOP presumption with its burden-shifting structure)
- the Montreal Protocol's Multilateral Fund records
- the pneumococcal Advance Market Commitment evaluations (NBER WP 26775 and the GAVI record)
- NASA's NAFCOM assessment of Falcon 9 development costs (2011)
- SEMATECH's final report to the Department of Defense (1997) with the National Academies' assessments
- Qualcomm's 10-K segment disclosures
- the Antarctic Treaty, Article IX
- the UN NGO Branch's consultative-status registers