## Please don’t share this document without consent!

Context for reviewers:

**Target audience:** This version is primarily designed for AI strategy researchers but written to be readily adaptable for policymakers and military organizations, with tailored versions planned following initial feedback.

**Feedback requested:**

- **Strategic assessment:** I would particularly value feedback on the proposed strategy.
  - Do you disagree with specific analytical claims or recommendations?
  - What aspects of the strategy do you find most compelling or concerning?
- **Clarity and logic:** Is the argument clearly structured and well explained?
- **Accuracy and completeness:** Are there factual errors, analytical gaps, or important considerations I've overlooked?
- **Additional contributions:** Can you suggest relevant sources, cases, or perspectives that would strengthen the analysis?
- **General feedback:** Any other observations on form or content are welcome.

**Formatting notes:** Red text indicates sections to skip; blue text is part of the document but requires deeper revision or relocation within the final document. Figures at the end are preliminary versions not yet integrated into the report and will be included once improved.

**Document scope:** This document contains the abstract, 3-page Executive Summary, and 14-page Detailed Context and Strategy Summary from a longer report that analyzes strategies for managing the AGI transition and their impacts on nuclear and power concentration risks.

# AGI Governance Through Multi-Sovereign Deterrence and Compute Control: Europe's Strategic Role

other title options

*I used AI tools to help edit and refine the writing; the arguments, structure, and analysis are my own.*

*Updated [date]: following the AI Futures Project's publication of Plan A, I added Appendix B comparing it with MSADS — where each approach seems stronger, and how Plan A updated my own — and made minor refinements throughout the report.*

## Update Note: Comparison with Plan A Added (see [Appendix/Section X]):

The strategy this report proposes and the AI Futures Project's recently published Plan A share much: both propose an explicitly negotiated worldwide regime resting on compute governance, verification, and mutual vulnerability. But they diverge on instructive cruxes — where compute is concentrated and who holds it, public versus selective research transparency, enforcement by two rivals versus by many sovereigns — and each design is stronger on some of them. Part of the divergence may stem from our different emphases: this report focuses more on preventing power concentration and gradual disempowerment risks, Plan A on harms from uncontrolled AIs; and where Plan A centers on the US and China, this report gives a larger role to middle powers — including a different account of how cooperation gets started, with middle powers pressuring it into being. The report describes what seems to me a near-ideal but plausible target, though in its ideal form harder to achieve politically than Plan A: a direction toward the most favorable choices rather than a fixed blueprint, several elements of which can be nuanced, dropped, or rearranged to make it more politically feasible. Several of Plan A's earlier phases could also serve as stepping stones toward it. Since Plan A's publication, I have therefore added an appendix that identifies these cruxes, assesses where each approach seems stronger, and traces how Plan A updated my own thinking — alongside minor refinements throughout the report. The appendix also develops several elements of this report's design — such as its mutual vulnerability mechanisms — in ways that could usefully complement Plan A, just as elements of Plan A could strengthen the plan proposed here.

# Abstract

AGI — systems capable of functioning as a country of geniuses — could arrive before 2030, and whoever develops it first could rapidly build insurmountable geopolitical advantages. This creates catastrophic risks: power concentration in unelected hands, democratic erosion from arms-race dynamics, unprecedentedly durable authoritarian entrenchment through AI-automated militaries and bureaucracies, loss of control from AI misalignment, and nuclear instabilities — for instance, if states facing post-AGI dominance by a rival calculate that crippling its AI projects is their safest remaining option.

This report analyzes existing strategies for managing this transition and proposes the Multi-Sovereign AI Deterrence Strategy (MSADS), which extends the Mutual Assured AI Malfunction framework from *Superintelligence Strategy* and builds on earlier proposals for centralized international control of AI development, such as the Multinational AGI Consortium. MSADS aims to establish an International Organization Controlling Compute Resources (IOCCR), initially with streamlined ECSC-style decision-making, that concentrates most frontier compute in a small number of secure facilities and coordinates safe AI development there.

Given chip-industry centralization, meaningful verification of compute stocks and flows is feasible: both individual states and the IOCCR itself can monitor that development, safely intervene to prevent AI capabilities beyond agreed, dynamically adjusted thresholds from being built or run outside this monitored system, and halt dangerous activity — the IOCCR by restricting access to the compute and models it controls, states through graduated mechanisms ranging from collective off-switches to, in extreme cases, unilateral intervention. This dual architecture is designed to guard against tyrannical power concentration by the IOCCR or any other actor.

The report sets out why MSADS offers a more desirable equilibrium than leading cooperation frameworks: its monitoring and enforcement mechanisms are robust enough to let states escape the post-AGI races to the bottom — across multiple risk dimensions, and intra-state power concentration in particular, which persists even under frameworks that equalize AI capabilities across states, as automation pressures hollow out democratic accountability. It better enables coordinated pauses, stronger post-AGI defenses, and more durable, equitable cooperation.

The report then discusses the strategic levers that non-leading AI actors — France, the UK, and the EU, or any willing nations — can use to establish this cooperation globally, with recommendations for scenarios where cooperation is slow to gain participation or fails. Delay prolongs exposure to destabilizing competition dynamics while allowing leading actors to cement difficult-to-reverse advantages. Through measures such as forming an early open-membership coalition and communicating ambiguous but justifiable deterrence thresholds in advance, European states can shift opinion in resistant states toward recognizing the self-defeating nature of non-cooperation — including the nuclear instabilities it creates — pressuring nations toward win-win outcomes while securing Europe against adversarial scenarios. As the risks of other cooperative and competitive approaches become apparent, the political window for tight global cooperation will likely open; actors who prepare early can raise the chance it opens and accelerate the shift once it does.

# Executive Summary (todo)

# todo 16-Page Summary E

*Reading time: ~30 min for the 16-page summary, ~5 min for the appendix.*

This 16-page summary lays out the core argument and strategic proposal in full; the longer report (open for comments) develops it further, and a condensed 3-page version is in preparation; where the summary and report diverge, treat the summary as authoritative. I'd especially value pushback on the strategy itself — which parts are most valuable, and anything I've gotten wrong. To help finish the report or discuss it, reach me at armdejen@gmail.com.

## Roadmap

This summary is organized in five sections. Section 1 introduces key premises on AI timelines and why AGI could be highly disruptive. Section 2 compares prominent international cooperation proposals for reducing AI risks and identifies their limitations. Section 3 proposes

MSADS's target end-state: a tight-cooperation system that better addresses those risks. It builds on earlier proposals for centralized international control of AI development, in part by establishing a new deterrence regime — Multi-Sovereign MAIM (MS-MAIM) — in which the International Organization Controlling Compute Resources (IOCCR) and every participating nation hold significant capabilities to monitor and halt dangerous AI development and misuse worldwide. Section 4 turns to the path. It first argues that resisting cooperation creates dangerous nuclear instabilities, then explains why establishing tight cooperation early is urgent, and why communicating deterrence thresholds in advance — safer than doing so late — can accelerate cooperation. It closes with concrete recommendations: the strategic levers that non-leading AI actors like France, the UK, and the EU (or any willing nations) can use to establish the IOCCR and MS-MAIM globally — and, separately, how to stay resilient if broad cooperation is slow to gain participation or fails entirely.

[Section 5] examines the closest rival to this report's proposal — the AI Futures Project's Plan A, which shares MSADS's threat model and reliance on compute control but bets on diffusion and total transparency where MSADS bets on concentration under multi-sovereign checks — assessing where each design is stronger and which of Plan A's mechanisms MSADS should adopt. [Section 6] concludes that tight, enforceable cooperation — the class to which the IOCCR and MS-MAIM combination and Plan A both belong — is the best option for every nation: more robust than lighter cooperative approaches and safer than racing and AI-dominance strategies, which are self-defeating.

**Section 5 concludes that tight cooperation like the IOCCR and MS-MAIM combination is the best option for every nation: more robust than other cooperative approaches and safer than racing and AI-dominance strategies, which are self-defeating.** Together, Sections 4 and 5 argue that as the risks of not establishing something like the IOCCR and MS-MAIM globally become apparent, the political window for tight global cooperation will likely open — and that actors who prepare early and push for cooperation can raise the chance it opens and accelerate the shift once it does.

I have also included a roughly drafted appendix that answers common questions about the coercive cooperation approach discussed in Section 4 — the idea of pressuring other states toward win-win outcomes, for instance, by declaring intervention thresholds and escalatory logic in advance and using those declarations to shift public and analyst opinion in other states toward cooperation. The bibliography below contains the longer report's full sources; this summary uses links to only a fraction of them.

## 1. AGI Will Likely Arrive Soon And Be Highly Disruptive

A 2024 survey of 2,778 AI researchers who had published in top-tier AI venues gave a 50% chance of Artificial General Intelligence (AGI) — "highly autonomous systems that outperform humans at most economically valuable work" — arriving by 2047, with a 10% chance by 2027, assuming scientific research continues undisrupted. As of February 2026, the forecasting platform Metaculus, which has proven effective at predicting near-term political and economic events, estimates a 25% chance of AGI by 2029 and 50% by 2033. A significant fraction of AI experts, forecasters, and especially entrepreneurs expect AGI before 2030, with Artificial Superintelligence (ASI) — systems vastly exceeding human capabilities across nearly all cognitive tasks — emerging shortly

thereafter. The main frontier AI companies are explicitly planning for "intelligence recursion": AIs automating AI research and the design of better chips, potentially triggering an intelligence explosion — a period of dramatically accelerating AI capability growth.

In 2024, Dario Amodei, CEO of Anthropic, described the coming transformation: by 2026-2027, AI systems could function as "a country of geniuses in a datacenter" — driving extraordinary technological acceleration, like a century of technological progress happening in a decade or less. Those who develop AGIs will command vast arrays of scientists, engineers, strategists, hackers, persuaders, and political operators — and other specialists — capabilities that could rapidly evolve far beyond genius-level across every domain critical to amassing power. They could greatly accelerate the development of breakthrough technologies, advanced robotics, autonomous weapons, bioweapons, and the industrial capacity to deploy them at scale — all with superhuman strategic coordination. Soon after AGI, a further feedback loop — advanced robots automating their own production and chip manufacturing — would likely play out and further reshape the strategic landscape. The first power to develop AGIs could quickly build an insurmountable geopolitical advantage, potentially even undermining nuclear deterrence. Throughout this report, "**advanced AIs**" refers to highly capable frontier systems — from those approaching AGI-level capabilities through AGI and beyond — that are powerful enough to substantially reshape states' military, economic, and political power.

The stakes of this transition are enormous. How states manage the AGI transition — whether through cooperation or competition — will determine whether these capabilities become humanity's greatest asset or its greatest threat.

## 2. Current Cooperation Proposals And Their Flaws

Several international cooperation proposals aim to properly manage the AGI transition. I describe three approaches, which can be combined — Mutual Assured AI Malfunction (MAIM), AI redlines with an IAEA-for-AI body, and proposals for a global moratorium on AI development — then identify their shared limitations. Two further families are treated elsewhere: proposals for a single international AGI institution, such as the Multinational AGI Consortium (MAGIC) and Bullock's Global AGI Agency, appear in Section 3, since MSADS builds directly on them; and the AI Futures Project's recently published Plan A — the closest complete rival to what this report proposes — is compared with MSADS in Section 5.

**Institutionalized MAIM regime.** The authors of Superintelligence Strategy argue that no superpower would remain passive while a rival risks transforming an AI lead into an existential threat. Instead, capable nations would likely threaten to preemptively sabotage AI projects threatening their survival. Given that AI assets are currently highly vulnerable to disruption, the authors predict an informal regime of AI deterrence — MAIM — would emerge organically. MAIM would initially operate through espionage to monitor AI progress, with states communicating escalation ladders — beginning with cyberattacks, chip supply chain disruption, and AI infrastructure sabotage, potentially escalating to conventional missile strikes on datacenters and their corresponding power plants, or even non-AI-related assets, if rivals pursue strategic dominance.

Rather than letting this regime develop chaotically, Superintelligence Strategy proposes preparing mechanisms to proactively establish a stable MAIM regime. In its mature, institutionalized form, MAIM would enable states to monitor each other's AI projects and safely intervene when rivals make strategic bids for dominance, by keeping critical AI assets deliberately vulnerable to disruption and, potentially, by introducing jointly controlled "off-switches" for destabilizing projects. This would allow states to advance AI capabilities at a similar, regulated pace — maintaining balance while avoiding dangerous volatility from unchecked and potentially fast intelligence recursion.

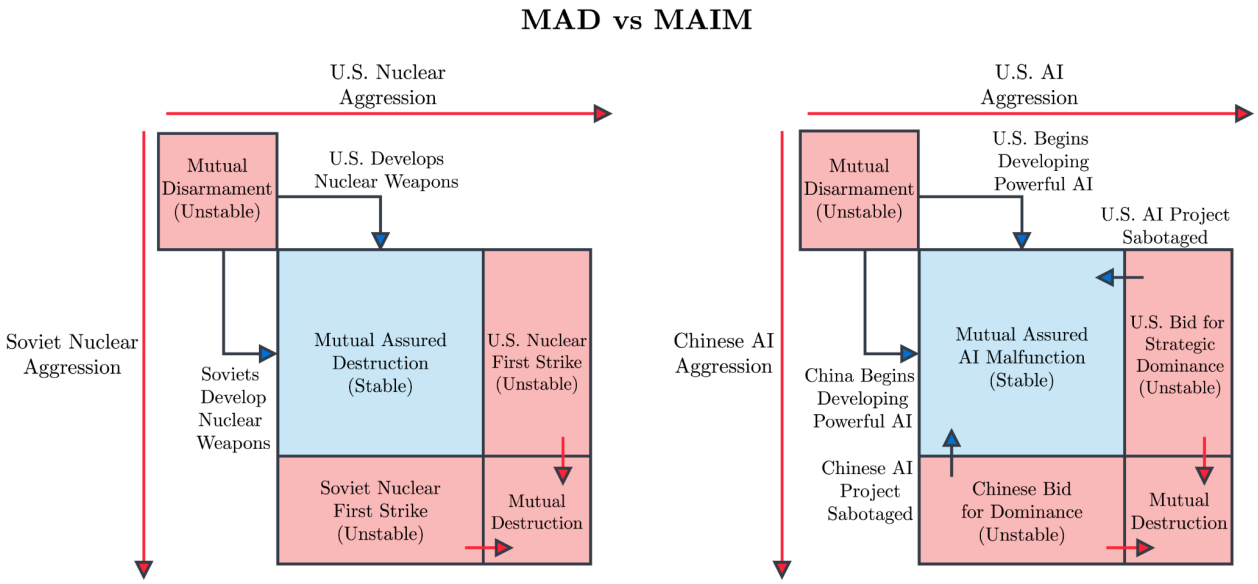



**Figure:** The strategic stability of MAIM can be paralleled with Mutual Assured Destruction (MAD). Note MAIM does not displace MAD but characterizes an additional shared vulnerability. Once MAIM is common knowledge, MAD and MAIM can both describe the current strategic situation between superpowers. *This figure and its caption are from Superintelligence Strategy.*

**AI red lines + an IAEA-for-AI.** In this approach, states agree on a list of categorically forbidden AI systems and uses — such as those proposed in the Global Call for AI Red Lines — encode these bans in domestic law, and enforce them through licensing, audits, and sanctions. A new international agency, modeled on the nuclear IAEA — often called an "IAEA for AI" — would set common safety and security standards for frontier labs, inspect major labs and data centers, monitor large training runs and flows of advanced compute, and publicly flag serious violations so that states and, where necessary, the UN can respond. The core idea is to make certain AI capabilities legally off-limits and to back that up with a standing inspection and reporting regime rather than relying only on voluntary industry or state commitments.

**Compute governance: chip tracking, nonproliferation, and information disclosure.** All states share an interest in limiting the AI capabilities of highly malicious actors, who could use them to develop cyberweapons or even biological weapons. *Superintelligence Strategy*'s nonproliferation pillar sketches a broad playbook: parallel US- and China-led regimes that secure high-end chips, model weights, and AI services in their respective blocs, alongside export controls and provider safeguards to keep frontier AI out of terrorists' and rogue states' hands. In practice, this means treating compute somewhat like fissile material: tracking where advanced AI chips are, maintaining chain-of-custody from fabs to data centers, and disrupting

smuggling and diversion routes. More ambitious treaty-based proposals, such as a Secure Chips Agreement, would go further by mandating globally unique chip IDs, shared registries, and on-chip security features that make covert misuse much easier to detect in every participating state. Related "compute governance" proposals, such as those calling for an International Compute Governance Consortium, suggest an international body that would first map who owns and controls high-end compute and coordinate chip-tracking, lifecycle safeguards, and technical standards across states, providing a shared toolbox that any of these MAIM- and IAEA-for-AI-style approaches can draw on to track compute and disclose information about sensitive AI projects as part of their implementation.

**AI-redlines + IAEA-for-AI variants.** Several proposals are specific instantiations of the AI-redlines-plus-IAEA-for-AI pattern described above: states adopt binding commitments via treaty and domestic law; an international agency audits labs, tracks advanced compute (often via Secure-Chips-style mechanisms), and certifies compliance; and non-compliance triggers state and possibly UN responses, with compute access as a central enforcement lever. Within this family, some proposals primarily start with a moratorium or cap on frontier capabilities (e.g. PauseAI, A Narrow Path), while others focus on enabling continued frontier development under strong safeguards (e.g. Belfield's "Four Institutions" blueprint).

*Pause-oriented variants (PauseAI, A Narrow Path):* PauseAI campaigns for a global pause on training the most advanced general-purpose systems until there is broad scientific and societal confidence that models beyond a given danger threshold can be aligned and controlled safely, and advocates for building a global "pause button" — the institutional and technical infrastructure needed to make such a pause enforceable. In practice, this would mean defining capability-based red lines — for example, strong autonomous cyber-offense, scalable bio-threats, and hard-to-reverse autonomy — and prohibiting training or deployment of systems that cross those lines unless robust, externally validated safety standards are in place. A Narrow Path goes further by explicitly sequencing a roughly 20-year prohibition on artificial superintelligence and banning key mechanisms believed to drive runaway risk, such as AIs that substantially improve other AIs, systems that can reliably escape human control, and "unbounded" agents with open-ended self-directed goals. Its later phases aim to build a durable international oversight system and safety-focused research capacity, then use this framework to steer incentives so that, after the delay, only safe advanced AIs (from near-AGI to beyond, defined in section1) are developed, and they are safely deployed and used.

*Governance-for-continued-development (Belfield's Four Institutions):* in contrast to pause-first strategies, Belfield's blueprint is oriented toward governing continued frontier development. It combines compute-indexed domestic frontier regulation, an International AI Agency to evaluate models and accredit national regimes, and a Secure Chips Agreement that conditions access to state-of-the-art AI hardware on accepting stringent oversight and verifiable controls on large-scale training. Belfield also proposes a US-led allied public-private frontier-AI megaproject, intended to concentrate the largest training runs inside a governed, democratically accountable framework, reduce racing between firms and states in that bloc, and share benefits more broadly than a purely corporate or purely national project would.

**Verification and enforcement problems.** The approaches discussed above — institutionalized MAIM and IAEA-for-AI-style regimes — implicitly assume frontier AI is developed in a few large, easily monitored compute clusters, and that most dangerous capability comes from scaling up the compute used in training ever-larger models. At least some proponents explicitly flag this as a contingent assumption, noting that significant algorithmic improvements or advances in distributed training would undermine the case for focusing regulation on only the very largest

runs (Belfield, 2025). AI capabilities are driven not only by scaling raw training compute (~4.5× per year), but also by rapid algorithmic improvements that increase the effective capability gained per unit of compute — on the order of 3× per year from training-efficiency gains and a further ~3× per year from post-training enhancements that make a fixed amount of inference compute more effective. These two sources of algorithmic progress mean that effective compute — the amount of capability one can extract from a fixed set of hardware — may rise by roughly 10× per year even when visible hardware stays constant. In parallel, inference efficiency has improved dramatically in recent years, making it much cheaper to run a given capability level over time and further weakening the link between visible training runs and who can actually wield dangerous capabilities.

On the training side, advances in distributed training increasingly allow large runs to be spread across many machines, and even across data centers, with only a multi-fold rather than catastrophic efficiency penalty — though achieving this still requires careful systems engineering and high-bandwidth interconnects. Combined with the fragmentation of cloud compute across numerous providers and jurisdictions, this makes it increasingly difficult to see who is using what resources for large training jobs. In parallel, a significant and growing share of capability now comes from scaling inference at deployment rather than from larger pre-training runs: running vast numbers of model instances, giving each call a much larger "reasoning budget", or embedding models in complex multi-step workflows. Unlike training — which at least requires concentrated compute for extended periods — inference is already naturally distributed across many clients, devices, and data centers, making high-impact deployments much harder to monitor at scale. Together, these current and developing training- and inference-side dynamics mean that dangerous systems could increasingly be developed and operated on cheaper and less detectable distributed infrastructure than current verification schemes assume — especially if AIs begin to autonomously drive a significant share of AI progress, which is likely to happen soon. In addition to these trends, radically more efficient AI architectures could also emerge that require far less compute for a given capability level, further weakening any governance strategy that relies on visible compute as its primary signal.

Historical evidence confirms that states routinely defect from treaty commitments, often undetected — even for programs with distinctive physical signatures. For example, the Soviet Union signed the Biological Weapons Convention while secretly operating a bioweapons network of ~40,000 people — a program whose true scale Western intelligence failed to detect until defectors revealed it was about ten times larger than estimated. Even when violations are detected, enforcement frequently fails — Syria's chemical weapons violations went unenforced after Russia vetoed UN Security Council action. And states can simply exit these regimes entirely, as North Korea did when it withdrew from the NPT to develop nuclear weapons. States with the most advanced AI models could also have less fear of sanctions or exclusion, given how quickly such models could bridge remaining technological gaps and deliver economic independence and abundance. The main leverage that some proposals derive from controlling access to the best chips is also likely to erode for China over the coming years, as it accelerates efforts toward semiconductor self-reliance and significantly expands domestic production capabilities.

Tracking a rival's technological development is already difficult, and advanced AIs will likely make it far harder — post-AGI technological progress could be fast, disorienting, and opaque even to sophisticated intelligence agencies. Today, intelligence agencies often infer secret programs from human-scale signatures — large facilities, distinctive equipment purchases, staff movements, and other traces of thousands of people working on a project. In an AI-driven — and eventually robot-driven — industrial base, much of this work can be done with far fewer

humans and in highly automated, reconfigurable facilities, greatly reducing these traditional signatures. Moreover, the approaches discussed above either do not propose to monitor what AIs are being used for or do so only at a coarse-grained level — yet advanced AIs, and eventually AI-directed robots, could be used to covertly develop dangerous technologies, either directly or by strategically building technologies and industrial capabilities that may not initially appear threatening but can later be repurposed for military advantage.

These verification and enforcement problems mean that IAEA-for-AI-like institutions and institutionalized MAIM would likely struggle to be reliably enforced, at least with the mechanisms they currently propose. The same problems also make these regimes hard to adopt in the first place: precisely because covert development and defection are difficult to detect, no state wants to constrain itself while rivals might press ahead in secret, even pulling away. This gap weighs far less on the more integrated approach developed in Section 3, where concentrating compute makes verification feasible and dual institutional-and-sovereign enforcement gives the regime real teeth.

**Even with broad compliance on redlines and MAIM, competitive pressures unacceptably increase post-AGI risks.** Advanced AIs could vastly accelerate technological progress and industrial expansion — especially in scenarios where robotics and AI progress are fast. This produces a powerful competitive dynamic through two reinforcing channels. The first is purely economic and technological: how aggressively a state lets AIs automate its economy, develop new technologies, and scale industrial and robotic capacity determines whether it pulls ahead or falls behind — so even absent any security threat, the economic incentive to race on automation persists. The second is military: unlike human workforces that require substantial retraining, AIs and robots can be rapidly redirected to entirely new purposes, meaning the technologies and industries a state develops could be quickly, and likely very covertly, repurposed for military dominance. Even if leading states officially cooperate and are initially compliant, they cannot trust that this compliance will hold, given the verification gaps and defection dynamics described above. Each channel alone could be sufficient to drive an automation race, and those who use advanced AIs most aggressively to accelerate technological and industrial expansion could gain insurmountable geopolitical advantages (for a detailed illustration of how automation dynamics could play out, see AI 2027). This holds even under MAIM: parity in AI capabilities does not eliminate the incentive to use AIs more aggressively for industrial and technological expansion than rivals do.

States will therefore be incentivized to deploy AIs across every possible sector, bypass democratic oversight in favor of more efficient AI-driven decision-making, and progressively remove humans from control loops — since human input increasingly becomes the bottleneck slowing AI systems that could otherwise coordinate and execute at superhuman speed. They will also be incentivized to accelerate self-reinforcing feedback loops — like AI-driven chip and robot manufacturing — that compound their advantage over time. As citizens become increasingly irrelevant for intellectual and physical labor — with AIs and robots surpassing humans at optimizing industries, bureaucracies, militaries, and AI progress — states maintaining strong human oversight face severe competitive disadvantages.

As Duvenaud's "Gradual Disempowerment" article notes, modern democracies have historically treated citizens relatively well because they needed their labor, taxes, and military service. Petrostates already show what happens when states can fund themselves from a narrow, capital-intensive sector instead: citizens become economically

dispensable, and accountability erodes. Post-AGI states could replicate this pattern at scale, using AI-driven rents rather than oil, and optimizing competitiveness by replacing humans with AIs and relaxing democratic and environmental constraints. The AGI transition era creates conditions where authoritarian governance becomes systematically more competitive than democratic governance, in part because autocracies can more quickly and forcefully concentrate compute under state control — and once AIs automate AI R&D, marginal increases in controlled compute directly translate into more and better AIs. Democracies then face a stark dilemma: accept structural competitive disadvantage or erode their own safeguards to match autocratic efficiency, thereby accelerating the kind of institutional disempowerment Duvenaud warns about.

These competitive responses — removing humans from control loops, automating core state functions, and subordinating welfare to state power — amplify both dangerous power concentration dynamics and AI takeover risks. Competitive pressures incentivize prematurely delegating key decisions to AI systems not yet proven aligned, making them more capable of disempowering humanity if misaligned. Between 38% and 51% of surveyed AI researchers assign at least a 10% probability to advanced AI causing human extinction or similarly permanent and severe disempowerment of humanity. If AIs remain aligned, these same conditions increase the risk of extremely durable authoritarian entrenchment through loyal AI-controlled military and police forces, comprehensive economic automation, and AI-enabled mass surveillance and manipulation. Even democratic countries risk seeing leaders cement oppressive ideologies or power structures. These arrangements could become extraordinarily durable: a leader could instruct their AIs to design only successor systems that preserve the same constraints — and to pass that rule down each generation — so that even future human leaders cannot reverse them. Even short of such extremes, citizens' democratic power risks being progressively hollowed out and their welfare underprioritized. Competitive pressures could also reduce investments in and coordination on nonproliferation, making it harder to prevent advanced AI capabilities from reaching rogue actors.

Even assuming states do not move to control most compute resources, corporate competition creates similar dynamics: once AIs drive AI research, companies with more compute will develop better AIs and rapidly outcompete others across every economically important domain — from software design and R&D to finance, logistics, and beyond — incentivizing mergers and compute pooling into very few macro-companies that concentrate unprecedented power in unelected hands. Competition also selects for those who deploy and delegate to AIs most aggressively. This dynamic risks diffusing poorly controlled and poorly aligned advanced AIs across most sectors while concentrating power in the hands of the least conscientious actors.

**Slowing frontier progress through imperfectly enforceable approaches creates its own tradeoff: more points of failure, and greater relative power for the least conscientious actors who defect.** This problem is most acute for pause-oriented proposals and MAIM: any approach that pauses or significantly constrains frontier development, especially for prolonged periods, lets more states and actors catch up to the constrained frontier, multiplying independent points of failure — and the more actors that reach a given capability level, the harder it becomes to ensure none defect. Given the advantages advanced AIs confer, the historical record of covert treaty defection, and the feasibility of advancing capabilities through algorithmic progress — especially post-training enhancements — on modest compute budgets, we should expect the least conscientious actors to continue development in secret, growing their relative power while pursuing less-monitored and therefore more dangerous work. How severe this tradeoff becomes

depends on how fast algorithmic progress proves to be, how much the frontier is slowed, and how effectively institutions can detect and prevent covert development. The same problem applies to redlines, if compliance significantly constrains capabilities.

A consideration pushing toward scaling faster in Plan A is deal decline: the possibility of the deal either dissolving or becoming substantially less effective (which we call deal impairment). In this supplement, I'll estimate the likelihood and badness of deal decline. I'm confident that deal decline is at least 10% likely to occur within Plan A's first 10 years, and that a 2035 decline leads to the expected value of the future being at least 5% lower than if there were a 2035 handoff within the deal. Since deal decline seems worryingly likely, and scaling capabilities within the Plan A deal seems much safer than an AI race once it's declined, I'm confident that deal decline is a substantial consideration, likely the biggest one, in favor of scaling up AI capabilities faster and handing off to AIs earlier (rather than something like a very long pause)

**In sum, post-AGI competitive pressures create races to the bottom across every dimension.** They make the least conscientious actors and autocratic systems more competitive and drive higher power concentration, democratic erosion, AI takeover, and proliferation risks. What is needed is an institutional environment that breaks both corporate and interstate competitive pressures — enabling AI development that prioritizes safety, dedicates appropriate resources to alignment and control, and ensures post-AGI decisions can be made through deliberate democratic processes without forcing states to choose between careful governance and staying competitive. The next section proposes such an environment.

## 3. The IOCCR and MS-MAIM Combination

MSADS has two parts: 1) a target end-state — the IOCCR combined with the MS-MAIM deterrence regime, described in this section — and 2) the path to reach it: the levers for establishing this tight cooperation, and the preparations for scenarios where broad cooperation proves slow or fails, described in Section 4.

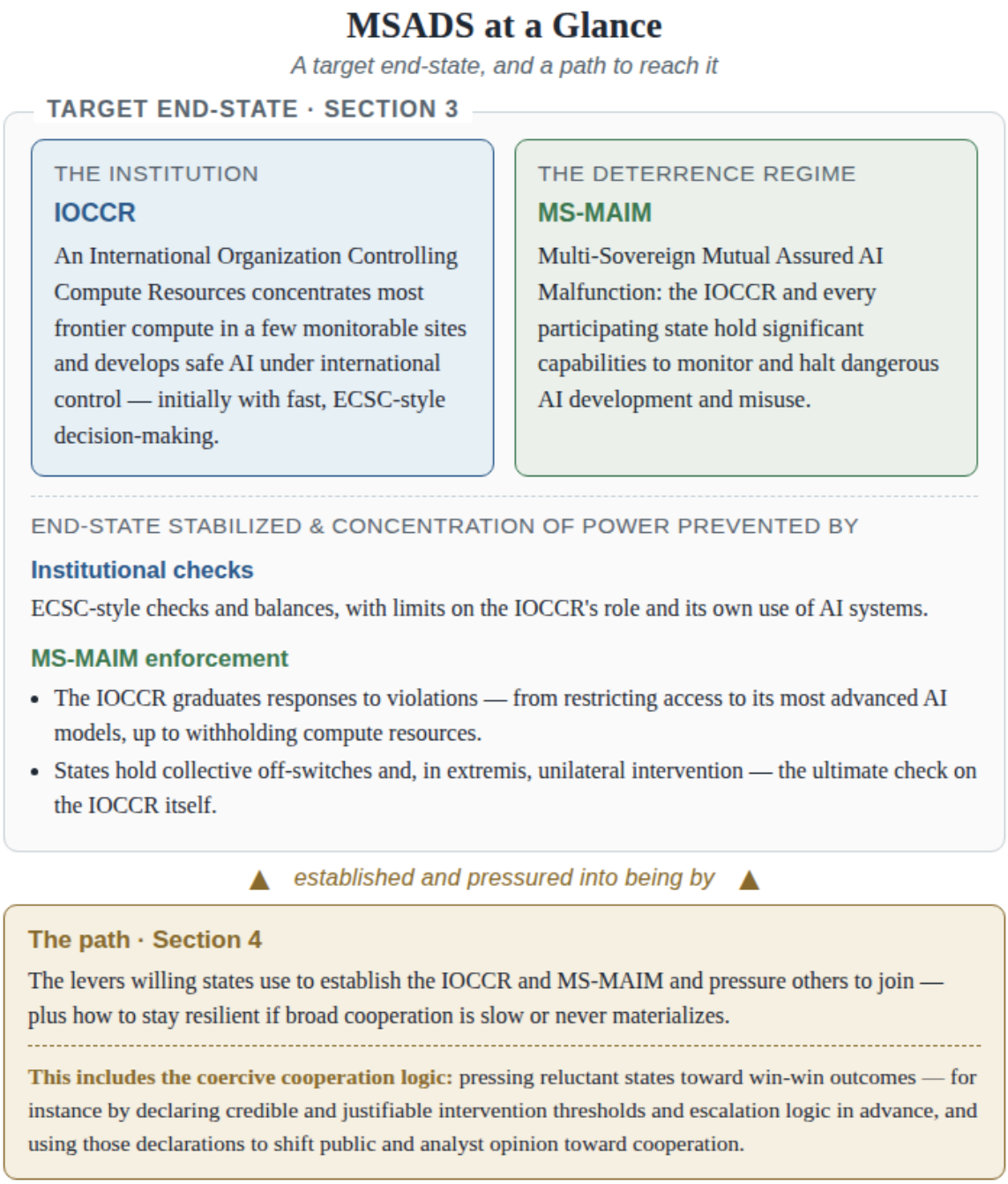
The IOCCR and MS-MAIM combination is designed to break the competitive pressures identified in Section 2 — and the dangerous dynamics that flow from them. Under MSADS, states cooperate to establish the IOCCR, initially structured with fast decision-making processes similar to those of the European Coal and Steel Community (ECSC). The IOCCR would coordinate the concentration of training and inference compute in a small number of monitorable locations — starting with existing major training facilities and potentially merging them — and would establish the MS-MAIM deterrence regime, in which the IOCCR and every participating nation hold significant capabilities to monitor and halt dangerous AI development and misuse. The "deterrence" in MSADS works at two levels: multi-sovereign deterrence stabilizes the cooperation regime from within (detailed in this section), while the threat of justified intervention also helps pressure states toward establishing it in the first place (detailed in Section 4).

This framework builds on MAIM-style deterrence, IAEA-for-AI inspections, and compute-nonproliferation, and it shares the core intuition of earlier single-institution proposals — such as Bullock's Global AGI Agency and Conjecture's Multinational AGI Consortium (MAGIC) — that the most dangerous AI development is best concentrated under international control rather than raced for. MSADS restructures these elements into a tighter, multilayered regime: by combining centralized control of most frontier compute,

continuous monitoring of AI use, and dual enforcement by both a supranational body and sovereign states, the IOCCR and MS-MAIM create checks and balances that diminish competitive pressures and reduce dangerous power-concentration dynamics.

Concentrating compute raises legitimate concerns about power concentration — but absent tight international cooperation that breaks post-AGI competitive pressures, power will likely concentrate in dangerous forms; the IOCCR and MS-MAIM architecture aims to make that concentration safe. The IOCCR would not be permitted to use advanced AIs (from near-AGI to beyond, defined in section1) except for limited purposes like initially accelerating the development of safe advanced AIs and strengthening defenses against post-AGI risks. It would coordinate pauses in AI progress if AIs couldn't be made safe — and, crucially, member states could independently vote to impose such a pause even against the institution's preference, requiring only a sufficiently large minority to do so. Placing the power to halt progress in the hands of states, separately from the body that supervises AI development, mirrors the way the ECSC's Council of Ministers could constrain the High Authority — the ECSC's executive — on key decisions.

As the situation stabilizes, nations would regain control over a greater fraction of compute resources to run IOCCR-developed AIs for approved and monitored uses. A dynamic quota system for compute allocation could also address the fairness problem that pure MAIM cannot solve: ensuring that compute access better reflects population and need rather than entrenching the existing dominance of whichever state currently leads in AI. Taken together — centralized safe development, transparency, dynamic quotas, the sharing of AI-developed technologies among members, careful institutional design, and MS-MAIM's distributed capability to detect and halt defection — these features create conditions in which competition pressures are low and coordination around diminishing post-AGI risks and positive futures becomes far more likely.




**Concentrating compute resources to develop safe AIs, accelerate defenses, and monitor AI use.** Given how concentrated advanced chip manufacturing is today and how hard it would be to covertly build a parallel cutting-edge fabrication capacity, meaningful verification of compute stocks and flows is feasible. If states grant the IOCCR sufficient authority to coordinate the concentration of compute, it could track chip production, confirm shipments to approved

facilities, and monitor how much compute accumulates in the one or very few IOCCR-approved sites — starting from existing large training centers and progressively consolidating them over time. Additional safety layers — remote shutdown mechanisms and pre-installed destruction mechanisms for major compute clusters — further enforce MS-MAIM compliance. Elements of this verification approach can draw on existing nonproliferation proposals, such as the compute-control framework in *Superintelligence Strategy: Expert Version* and the proposed Secure Chips Agreement. Under MSADS, AI research above a given compute threshold is restricted to IOCCR-supervised facilities — and as algorithmic efficiency lowers that threshold, the scope of exclusivity expands accordingly.

By concentrating compute resources in these monitored locations, the IOCCR can ensure AI development happens in highly secure, safety-focused facilities with appropriate resources dedicated to AI alignment and control — something competitive dynamics would otherwise make far less likely. The most advanced safe AIs can then be used to accelerate defenses against post-AGI risks: cyber-defenses, bio-defenses, AI for epistemic, defense-dominant military technologies, and more resilient nuclear second-strike capabilities. The IOCCR could go beyond developing these technologies and sharing knowledge of them: by coordinating their production and deployment across members, it could ensure that every member is actually defended rather than merely informed. Once deployed, these defensive systems would remain under the sole control of the member states hosting them — they could not be switched off by the IOCCR or by other members.

All interactions between AI systems and the external world would be monitored, providing fine-grained transparency that helps detect dangerous AI use and temper AI takeover risks while ensuring that the knowledge of AI-made technologies is shared among member states — preventing any single actor from accumulating a decisive and unchecked technological advantage. States would access IOCCR-developed AIs for approved uses on IOCCR's tracked inference compute. Because advanced robots would run on this monitored compute, the same transparency extends to industrial activity — answering the concern raised in Section 2 that AI-built industries and robots could be quickly and covertly repurposed for military dominance. Compute becomes the post-AGI equivalent of coal and steel: the essential input to war preparation, and, once traceable, equally impossible to stockpile in secret. Just as ECSC members could not quietly build up the materials for war, MS-MAIM members could not quietly prepare AI-enabled military campaigns — and enforcement reaches the physical economy itself, since a defector's automated industry and any robot-based military buildup could be halted along with its compute, not merely its AI development.

**MS-MAIM: a dual-enforcement architecture.** MS-MAIM combines two complementary enforcement layers to prevent dangerous power concentration and enable coordination around critical governance functions — preventing AI misuse, pausing AI progress when appropriate, enforcing the IOCCR's exclusivity over frontier AI development, and accelerating defenses. The first is the IOCCR's own graduated institutional enforcement. Because the IOCCR controls the bulk of compute resources and operates the most advanced AI systems, it can respond to violations — whether dangerous AI use that slips past approved-use filtering, activities outside IOCCR-supervised compute, or non-compliance with governance decisions like refusing inspections. Responses escalate with severity: restricting access to the most advanced AI models for minor violations, cutting access to all advanced models for more serious ones, and ultimately reducing access to general compute resources for the gravest cases. Voting thresholds scale accordingly — lighter interventions require a smaller share of members, and stronger ones require broader consensus.

The second layer is independent sovereign enforcement by member states, which operates at two levels. The first is collective: remote off-switch capabilities over compute infrastructure, requiring only a sufficiently large minority of MS-MAIM members to activate — illustratively around thirty percent, though the precise threshold could be set higher (possibly even higher than fifty percent). This deliberately low threshold makes intervention against dangerous behavior easy enough to trigger that no coalition of states can shield a dangerous project by colluding or by collectively declining to intervene, while remaining high enough that no single actor, or handful of actors, can abuse the mechanism — and it gives smaller actors real intervention power. The second is unilateral, available to states with sufficient capabilities: in extreme cases, kinetic intervention against compute facilities. Because these facilities are deliberately sited in remote areas far from population centers, such intervention can be undertaken with limited collateral harm. Siting could go further: placing major clusters in international waters — as Plan A proposes with its floating datacenters — would make interventions against them less escalatory than strikes on sovereign territory, and would make the facilities harder for a defecting state to seize or successfully defend, deliberately preserving the vulnerability on which the deterrent depends. Because this last-resort capability does not depend on the institution's channels or on a collective vote, it serves as the ultimate check on the IOCCR itself — if the institution is compromised or captured, states retain enforcement tools that operate entirely outside its control. These sovereign capabilities also reach compute that the IOCCR's access control cannot: if a state defects and retains physical hardware on its territory, IOCCR access restrictions become irrelevant, but off-switches on the chips themselves or, if necessary, interventions like missile strikes could still disable it.

Together, these two layers ensure that enforcement does not depend on fragile great-power consensus: because intervention capabilities are held independently by the IOCCR and by sovereign states, the regime stays robust even if any single actor — or the IOCCR itself — is captured or tries to abuse its position.

**Evolving IOCCR governance.** To quickly centralize compute resources, establish the conditions for MS-MAIM, develop safe advanced AIs, and accelerate defenses, the IOCCR could initially adopt streamlined power structures akin to those of the European Coal and Steel Community (ECSC): a supranational High Authority-style body with strong, relatively autonomous executive powers — balanced from the outset by intergovernmental checks and a parliamentary power to dismiss the executive — explicitly designed to be insulated from day-to-day national and sectoral pressures and capable of fast decision-making with limited procedural friction. This streamlined structure would allow the IOCCR to rapidly integrate key industries, much as the ECSC did for countries that had recently been at war and remained at risk of renewed conflict.

The IOCCR's own use of advanced AIs would be tightly constrained — limited to developing and controlling safe AI systems, accelerating defenses against post-AGI risks, and other specific, time-bound purposes subject to member-state oversight — preventing the institution from accumulating unchecked AI-driven power. As the situation stabilizes and trust and verification mechanisms mature, the IOCCR could transition toward a primarily monitoring role, with states gradually regaining control over a greater fraction of inference compute to run IOCCR-developed AIs for approved, monitored uses. IOCCR leadership would be appointed by member states for fixed terms, and its operations would remain under continuous oversight by designated national authorities.

A single institution coordinating most frontier compute would face substantial design and adoption challenges, such as international cooperation hurdles, private-sector resistance, and

the centralization-of-power concerns that Bullock and others detail. MSADS does not claim to dissolve these entirely; its multi-sovereign enforcement layer and ECSC-style checks are meant to make the concentration of compute safer and more accountable, while removing the competitive pressures that drive dangerous corporate and national power concentration.

**The IOCCR and MS-MAIM break competition pressures and enable coordination around positive futures.** Knowing that defections will most likely be detected and met with appropriate interventions creates a climate of caution that reinforces compliance. If adopted by leading AI states, the IOCCR commands vastly more compute and more advanced AIs than any other actor. Given the abundance and technological advantages that participation enables and the costs of defection, there is no rational incentive to defect — and unlike the races to the bottom described in Section 2, the most cautious actors have significant leverage to shape the AGI transition rather than being outpaced by the least conscientious.

With competition pressures broken, it is easier to dedicate appropriate resources to AI alignment and control, accelerate defenses against post-AGI threats, implement coordinated pauses in AI development if needed, preserve democratic systems, and coordinate on shared challenges like climate change, as well as on the equitable sharing of AI-generated benefits. Unlike voluntary moratorium proposals, coordinated pauses under the IOCCR and MS-MAIM are more verifiable and enforceable through the IOCCR's compute monitoring and MS-MAIM's intervention capabilities.

The IOCCR and MS-MAIM create conditions in which AI takeover risks are substantially diminished: AIs are developed with far greater resources dedicated to alignment and control; development is concentrated in one or very few facilities, leaving fewer points of failure to secure and monitor; all AI interactions with the external world are monitored; clear off-switch mechanisms exist; and — crucially — no competition pressures push states to deploy poorly controlled AIs in dangerous applications. Any dangerous AI operating outside monitored compute is moreover at a steep disadvantage, with access to far less compute than the IOCCR commands. These same conditions make AI misuse easier to detect and intervene against, enabling the entrenchment of durable peace. The IOCCR and MS-MAIM do not need to achieve perfection — controlling and monitoring the majority of compute resources and enabling safe intervention creates better conditions for managing post-AGI risks than any alternative.

# 4. The Path Forward: Europe's Levers for Cooperation and Resilience

This section turns from MSADS's target end-state to the path for reaching them. It first discusses the risk of nuclear instabilities, then explains why establishing tight cooperation is urgent, before setting out recommendations for how France, the UK, the EU, and more broadly any willing coalition of states can establish the IOCCR and MS-MAIM and pressure other nations to join before the window for favorable conditions closes. It also recommends ways to stay resilient if broad cooperation is slow or never materializes. Much of this logic generalizes: many of these recommendations can support the establishment of other win-win AI cooperation frameworks.

**Resisting cooperation creates nuclear instabilities.** States that resist win-win cooperation despite understanding its feasibility — even as others communicate justifiable discomfort — risk

signalling potential belligerent intent. Consider Russia's position if the US achieves post-AGI dominance without transparency over compute usage. Growing uncertainty about US intentions would compound several fears: that the US could erode the survivability of Russia's nuclear retaliatory capabilities; that Russia could become acutely vulnerable to well-orchestrated AI manipulation campaigns, large-scale cyberattacks, or autonomous-weapon attacks whose origin is hard to attribute; or even that the US might lose control of its own AI systems in ways that endanger Russia too. Together, these would create pressure to escalate before the US consolidates decisive advantages. Initial escalations can help test US intent, and if these fail to extract credible security assurances, Russian leaders might judge that they have sufficient reason to fear US intentions to accept the risk of using conventional missile salvos of increasing size on AI-related infrastructure. If these proved insufficient, they could calculate that limited nuclear salvos of increasing size offer the best remaining leverage to pressure US cooperation before dominance is consolidated.

This escalation pattern is consistent with Russia's defensive escalation management doctrine, but even states with nominal no-first-use policies like China could react similarly under extreme pressure. The possibility of nuclear use should not be taken lightly: 80,000 Hours estimates the chance of a nuclear war in the next 100 years at around 20–50% — without accounting for the heightened risks of the AGI transition. A US–Russia nuclear conflict would be catastrophic for every state: nuclear winter caused by soot blocking sunlight would kill roughly 5.5 billion people from famine alone in expectation, accounting for the full range of escalation scenarios.

The US is unlikely to be able to protect itself fully against such escalation. Even the roughly 300-billion-dollar US missile defense effort currently provides only limited protection: these systems remain unproven and unreliable even against a small salvo of a handful of intercontinental missiles, let alone larger or more sophisticated attacks. It is also much harder to build a defense that can reliably stop a sufficiently large and sophisticated salvo than to design an attack that can break it. Even for a small North Korean ICBM force, US strategic missile defense cannot be counted on to stop all incoming warheads; at best it complicates an attacker's planning and might intercept some fraction (see Appendix B – Technical Limits of Missile defense).

Partially successful MAIM-like cooperation may give states enough security that they decide escalating to missile strikes is not worth the risk — but as long as MS-MAIM is not established, instabilities could remain. Fast and opaque technological progress could make states sufficiently worried about the survivability of their nuclear forces that they adopt dangerous doctrines — like pre-delegated launch authority or launch-on-warning postures — that increase retaliatory survivability but also the likelihood of accidental, unauthorized, or inadvertent nuclear use.

**Getting broad cooperation started earlier is urgent.** Today's concentrated AI infrastructure and centralized chip supply chains create favorable conditions for establishing the IOCCR and MS-MAIM — conditions that may not persist if leading AI actors adopt costly defensive adaptations like distributed training or hardened facilities. Delay prolongs exposure to competition dynamics while allowing leading actors to consolidate advantages that become increasingly difficult to reverse — so any eventual cooperation framework risks being shaped by and for those who led, entrenching power asymmetries rather than correcting them. AI actors are more likely to cooperate prosocially while uncertain about who will win the AI race; once clear winners emerge, they may accept cooperation only under terms unfavorable to others. And if AIs begin accelerating AI research itself, this window could close quickly — before states have enough time to react.

**Communicating deterrence thresholds early accelerates cooperation.** AI deterrence is our best option for tempering racing dynamics and great power conflicts — a logic that Dan Hendrycks and Adam Khoja defend convincingly against critics. By preventing rival miscalculations, imposing costs that slow destabilizing AI projects, and strategically pressuring nations toward win-win outcomes, it makes broad cooperation more likely to occur. As in nuclear doctrine, communicating ambiguous thresholds avoids locking nations into automatic escalation to preserve credibility, instead allowing context-appropriate responses while still keeping adversaries cautious.

Any capable nation whose national security is sufficiently jeopardized will eventually communicate escalation thresholds to try to restore it — particularly once smaller measures like trade restrictions or covert sabotage prove ineffective. Since this communication will, absent cooperation, most likely happen eventually, doing it early raises awareness of the strategic situation among citizens, analysts, and policymakers and pressures rivals toward cooperation — and doing it late does the same, but more dangerously, leaving little time for awareness to build and cooperation to be established before thresholds are crossed.

## Europe's role and recommendations.

France, the UK, and the broader EU are well-positioned to initiate MSADS. France and the UK possess credible deterrence capabilities that, if maintained proactively, I do not expect to be undermined, and European states collectively have stronger incentives to pursue tight international cooperation than expected post-AGI leaders such as the US and China. Their less adversarial relationships with both major AI powers also make them more credible initiators of cooperation than either the US or China could be for each other. The following steps serve a triple purpose: preparing Europe to quickly implement MSADS if the political window opens; pressuring resistant states toward cooperation through visible, credible action; and strengthening Europe's position even if global cooperation fails and more adversarial scenarios unfold. Five concrete steps are needed:

**First, invest now in detailing and comparing AGI transition strategies** — including the IOCCR and MS-MAIM alongside other approaches — to strengthen the case for tight cooperation and prepare detailed implementation pathways, ready to implement if the political window opens.

**Second, build on existing cooperation proposals and engage allies and potential rivals on tighter arrangements.** I expect many elements of MAIM-like or IAEA-for-AI-like proposals to be valuable first steps to diminish risks and pave the way for tighter cooperation — but the goal should be the tighter forms of cooperation described in Section 3. In the absence of broad cooperation including all leading AI nations, an initial tighter framework among willing countries could start with models like CERN for AI or Intelsat for AI — centralized research facilities with international governance and open membership, focused on developing safe frontier-level AI capabilities. To accelerate coalition formation, the initial group could start small — with Europe's largest economies or even just the Benelux countries — before expanding, potentially to willing countries beyond Europe such as Canada, Japan, or South Korea, and ideally to all leading AI powers, especially the US and China.

Building such a coalition is valuable even without the participation of the leading AI powers: it provides better defense against more adversarial scenarios, partly by turning each member's partial position in the AI supply chain into joint leverage. That leverage could secure access to frontier AI, push for the models it obtains to be safer and more reliable, and — as a bloc both superpowers need to deal with — help mediate between the US and China. While global cooperation is still absent and racing dynamics persist, a coalition that includes the US makes it more feasible to pause AI progress at crucial times — in particular around the point where systems begin to substantially automate AI research, which many authors identify as a critical moment to slow down and focus on aligning these "AI-research AIs" before they drive a rapid acceleration of capabilities — and to devote more resources to alignment and control research, increasing the chances that those early systems are aligned and controllable. Such a coalition would also create greater strategic costs for those remaining outside. The safest path, however, remains reaching broad international cooperation as early as possible. Building a coalition that holds a strong advantage over others — or that helps the US secure a large AI lead — could backfire badly: it could tempt the leading group to stop keeping the door open for cooperation in favor of pressing its advantage, fueling dangerous racing dynamics and raising the likelihood that China or others climb the escalation ladder, potentially triggering catastrophic conflict.

**Third, strengthen the resilience of European nuclear deterrents.** France and the UK are not currently preparing for scenarios where adversarial AI states could develop new effective attacks against nuclear forces — including rapid progress in anti-submarine warfare or novel attacks against command, control, and communications systems. Without resilient deterrents, threatened states risk accepting unfavorable strategic concessions from a position of weakness, adopting dangerous launch doctrines, pursuing preemptive strategies, or some combination of these. To prevent such scenarios, France, for example, could reintroduce a land-based leg to its nuclear forces, ranging from fixed launcher sites to more politically challenging options like deploying French-owned, French-operated road-mobile launchers across willing host countries.

**Fourth, build international consensus that developing AI past key thresholds outside a cooperative framework threatens all states** — that pursuing capabilities like AIs able to fully automate AI research or approach AGI, outside a common project (or at least a framework that keeps states progressing at a similar pace and ensures AIs are proven safe, well controlled, and monitorable by all states), constitutes a threat to others' national security. Establishing this shared understanding is what later makes intervention legible as a justified response to a recognized danger rather than as aggression.

**Fifth, define ambiguous intervention thresholds and communicate credible escalation ladders in advance** — triggered by the capability thresholds above (for instance, AIs automating substantial AI research outside a cooperative framework), as well as by broader strategic conditions that resist crisp definition, such as insufficient transparency over compute usage or leading actors entrenching dangerous, difficult-to-reverse advantages.

**A central purpose of these steps — especially the third and fifth — is to shock analysts and publics into recognizing the necessity of tight cooperation.** Visible actions like strengthening deterrence and communicating ambiguous intervention thresholds draw attention to the risks of the AGI race and of poorly managed post-AGI competition. Paired with clear explanations of why tight cooperation benefits all nations, including the US and China, they help shift both opinion and political will toward cooperation earlier than would otherwise occur.

# MSADS vs Plan A

Plan A seekes to establish what I would call a mature
As with AI redlines, IAEA-for-AI bodies, and institutionalized MAIM, Plan A could be the step before the IOCCR and MS-MAIM combination gets implemented.  I think that both Plan A and my approache are not incompatible and actually implementing something like Plan A which seems politically more feasible would move the world much closer to then implementing the IOCCR and MS-MAIM combination if the latter seems preferable or solve problems that Plan A can't and it starts having more polititcal buy-in.

Our two approaches share many elements in common: for example, Plan A chooses over alternatives to continue AI progress in a regulated way that helps develop safe AIs and entrench a stable world order via systems of mutual transparency and mututal vulnerability. but differ on few interesting points for example in the sort of transparency and mututal vulnerability systems that they use or how they avoid different forms of power concentration and many of the elements of their plan can be used in mine and vice versa. This section will explore some of our most striking differences and ways to enhance our both approaches

## 5. Conclusion: MSADS Is Our Best Option — It Needs Urgent Preparation

**AI dominance strategies are self-defeating and endanger every nation.** A mostly unregulated competition in which states race for first-mover advantages in AGI development and then seek to maintain superior AI, technological, and military capabilities would amplify every risk described in Section 2 — intensifying competitive pressures across all states while increasing nuclear instabilities. A more specific variant of this logic argues it is worth racing to aligned ASI, building a decisive military advantage, and then negotiating a new safe international order from a position of strength. But such strategies would most likely be blocked: once AIs begin to substantially automate AI research — most likely before full AGI — rivals will still have engineers inside the top labs and sufficient situational awareness to respond with graduated interventions, escalating if necessary to measures that inflict devastating damage on the infrastructure a dominance bid requires. And even if rivals failed to intervene effectively, the attempt itself would heighten nuclear instabilities and sharply amplify all post-AGI competition-associated risks — including in rival states, where the pressure to keep up increases the likelihood of catastrophic loss of control over AI systems in ways that affect every nation. The risks that dominance-seeking strategies aim to manage can be addressed more directly by establishing the IOCCR and MS-MAIM. Neither domination nor the cooperation approaches examined in Section 2 adequately address post-AGI risks, and we need to aim for tighter forms of international cooperation.

If the IOCCR and MS-MAIM — or at least MAIM-like approaches — are not broadly adopted before AIs substantially automate AI research, my guess is that a second-best option is for a US-led democratic alliance to concentrate compute and talent in a frontier-AI megaproject, pausing around the automation threshold so that alignment work can catch up and the best AI tools can help design stronger cooperation frameworks. Beyond this point, racing toward AGI or ASI without the IOCCR and MS-MAIM — or at least institutions capable of significantly slowing

AI progress globally and making it possible for different states to progress at a similar pace — would be extremely dangerous.

**The political window will likely open.** At least some major powers will very likely come to recognize the necessity of systems like the IOCCR and MS-MAIM — but this recognition may come only after the risks of alternative approaches become painfully obvious. Citizens concerned about preserving democratic systems may increasingly realize that without eliminating competition pressures to remove humans from control loops and delegate to AIs, those systems are unlikely to survive. Aligning advanced AIs could take years or remain permanently uncertain, leading states to desire conditions where they can better dedicate appropriate resources to alignment (possibly via pausing AI progress) and control — and where the most advanced AIs can be effectively monitored and switched off if needed. Nuclear and geopolitical instabilities from asymmetric AI capabilities may grow unbearable. Once AIs automate AI research, compute resources become the primary determinant of AI capabilities — creating pressure for states to have comparable amounts of compute. But purely bilateral arrangements to achieve compute parity face deep political obstacles: the US has far more compute than others, and the most populous nations will refuse to entrench systems where they receive less compute per capita. States may therefore recognize that dynamic quota systems — which only an institution like the IOCCR could implement — offer the only viable path to resolving these instabilities while ensuring fair access to the greater shared abundance that centralized AI development enables.

**Preparation must begin now.** The sooner this recognition comes, the lower the cost. As argued above, the window for favorable cooperation is narrowing. We should preferably not wait for a catastrophe like the proliferation of advanced AIs to bad actors who use them to create bioweapons, an attempted AI-enabled coup, escalatory interventions between major powers, or unmistakable signs that power concentration dynamics are eroding citizens' democratic power to start preparing and acting. The first states to recognize the necessity of tight cooperation have levers to pressure others so that the political window in all major AI states opens faster. Establishing a mature IOCCR and MS-MAIM combination requires sustained preparation that willing nations should begin immediately.

Some elements of this preparation have irreducible lead times: developing detailed institutional blueprints, fostering consensus around AI risks and justifiable intervention thresholds, negotiating agreements, strengthening the resilience of nuclear retaliatory capabilities, and adapting nuclear doctrines. Having a well-understood proposal already in policymakers' minds also matters — if the political window opens suddenly, whether through crisis or through AI systems themselves recommending stronger cooperation frameworks, leaders are far more likely to act on ideas they have already engaged with than to adopt unfamiliar proposals under pressure.

Even readers unconvinced that the IOCCR and MS-MAIM combination is our best option, or even feasible, should recognize that preparing to implement it is worth the cost — future conditions may prove such systems necessary and build the political will to adopt them, and those who prepared will shape what emerges.

# Appendix A: More Details and Q&A on the Coercive Cooperation Approach (Drafty)

Much of the interest and pushback I get when presenting this report concerns the coercive cooperation approach: the idea of pressuring other nations toward win-win outcomes, for instance, by declaring intervention thresholds and escalatory logic in advance, and using those declarations to shift public and analyst opinion in other states toward cooperation.

Many of the arguments for AI deterrence and the coercive cooperation approach are set out in two pieces — AI Deterrence Is Our Best Option and Coercive Cooperation: Forcing Win-Win Outcomes in an AI-driven Transition — and I develop the logic at greater length in the full report (especially the section 6. Establishing a functional IOCCR and MS-MAIM Regime: Europe's levers and to a lesser degree the section 5. Why the IOCCR and MS-MAIM Combination Is Our Best Cooperation Option). Below I add some of the questions I most often receive, with drafted answers that also include arguments not found in those other sources. As of now there is no single place where all these arguments live together; getting a comprehensive view still requires reading across these different sources.

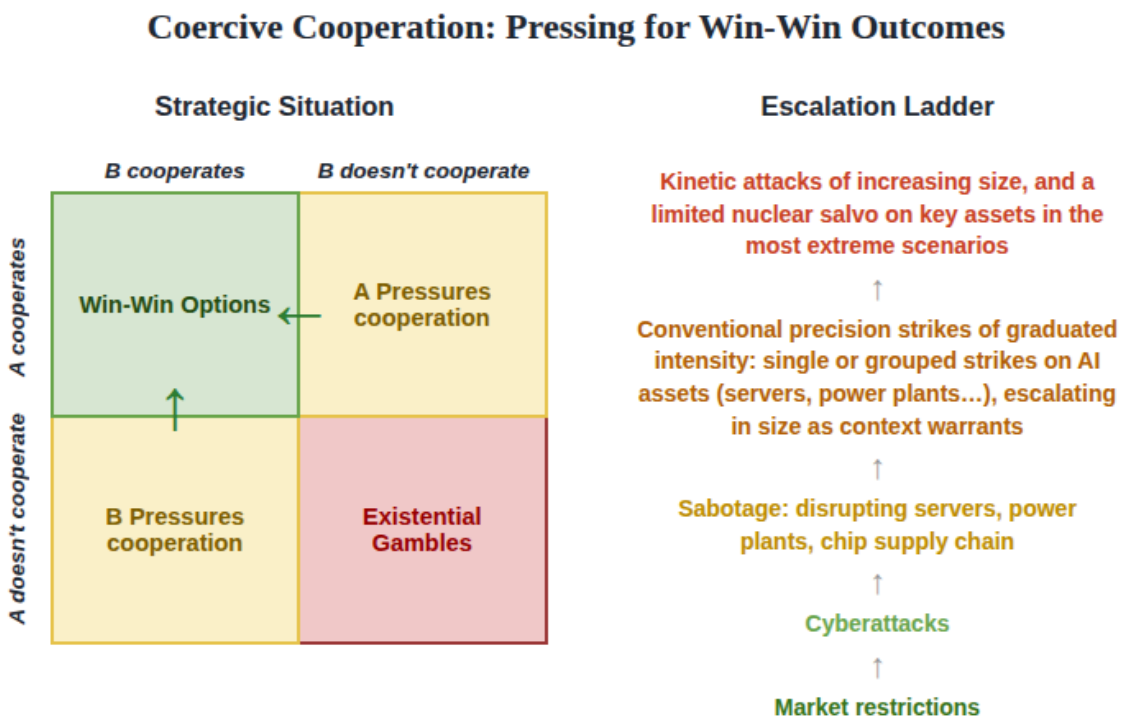



**Figure: Coercive cooperation.** Left: MS-MAIM is a feasible win-win equilibrium, but some states may fail to recognize this due to knowledge gaps, trust deficits, or institutional inertia. As Eric Drexler explains, coercive cooperation uses pressure to move a non-cooperative rival toward mutually beneficial outcomes despite such obstacles. A state or coalition can use the graduated escalation ladder (right) — from market restrictions through cyberattacks and

sabotage to conventional precision strikes of graduated intensity and, in extreme scenarios, nuclear strikes of graduated yield — to shift a reluctant rival's cost-benefit calculus toward cooperation. These escalations can also inflict real costs on non-cooperative rivals that slow and most likely prevent potential bids for dominance. When neither state cooperates, both face existential gambles. As in nuclear doctrine, communicating ambiguous deterrence thresholds avoids locking states into automatic escalation to preserve credibility, instead allowing context-appropriate responses while still keeping adversaries cautious.

**Would states take the escalatory warnings seriously?**

I expect that as AGI approaches, norms will shift and states will come to recognize how threatening advanced AI systems are to those who lack comparably capable systems — through rogue-AI risks, the possibility of sophisticated coordinated attacks (mass parallel manipulation of populations and the information environment, potentially reducing other states to puppets; coordinated cyberattacks paired with large numbers of advanced autonomous weapon,..), power-concentration risks, and nuclear instabilities. A warning issued by a more justifiably threatened state is also likely to be received as less aggressive than the same warning from the US or China.

**Isn't issuing such warnings outside European doctrine — and is it plausible to direct them even at allies?**

First, this is a warning, not a threat. It is not "we will attack you for some gain we extract"; it is "if you develop advanced AI without the cooperation that keeps us safe, that is an existential threat to us, and we are therefore willing to follow a gradual, escalatory approach to test your intent and press you toward cooperation." It is also paired with a cooperation proposal that is arguably in the other state's interest too.

Second, it is a generic warning: because it does not single out one country, it reads as less aggressive.

Third, there is precedent among allies. The US has threatened allies — for example, over Greenland — though it can do so more cheaply than Europe could, since it has many economic levers for retaliation. And when sufficiently threatened, France already resorts to nuclear signaling; in the Greenland episode it made implicit nuclear signals toward the US.

**Would France ever recognize the AI threat and decide to act — or will it be too late once others reach AGI/ASI? Can it even know such systems are arriving?**

It will know. France has engineers in the leading AI labs, and will retain that visibility at least until AI fully automates AI progress — so as a potential intelligence explosion begins or AGI approaches, it will have the information it needs. Acting on that information is the harder question. My guess is that France is unlikely to begin applying the coercive cooperation logic before AIs substantially automate AI progress; but after that point, in a disorienting and fast-moving period, it could become increasingly willing to — and far more so if it has already encountered the arguments in advance. Hence the importance of making the case to such states now.

There is also likely to be time. Eroding a rival's nuclear deterrent is far harder than maintaining one, so I expect it would take several years post-AGI — more than three to five — before a

state like the US could plausibly gain a decisive military advantage. That leaves a window in which others can recognize what is happening and respond.

**Wouldn't a missile strike directly trigger large-scale nuclear retaliation?**

Two observations suggest not necessarily. Ukraine conducts strikes inside Russian territory without triggering nuclear retaliation, for several reasons — two of which are that the strikes do not cross the doctrinal threshold (for most states, an existential threat to the state itself), and that the political and strategic cost of nuclear use would be enormous: breaking the nuclear taboo would isolate the attacker sharply, likely rupturing relations with partners vital to its economy and diplomacy. I expect comparable political and public-opinion costs would constrain the US.

Moreover, nuclear escalation would not be a logical response against a state like France, because it would likely invite counter-retaliation: France can currently destroy roughly half of the US economy and kill around 20% of its population with its existing arsenal, and scaling that arsenal would not be especially difficult — it fielded one nearly twice as large during the Cold War. The missile strike is also a late rung: it comes only after earlier escalations meant to test intent and press for cooperation.

Most importantly, a kinetic strike is never an automatic or opening move. It sits at the top of a graduated ladder — market restrictions, cyberattacks, different forms of sabotage, and is reached only after the earlier, lower-cost rungs have been tried and have failed to shift the target's behavior. Even then it is not triggered mechanically: it is a deliberate last resort, undertaken only when escalation is judged to carry lower expected costs than allowing the dangerous development to continue. Those earlier rungs serve a double purpose — they test the target's intent and apply real pressure toward cooperation — so by the time a strike is even on the table, both sides have had repeated, legible opportunities to step back. This gradual, non-automatic structure is also what keeps escalation controllable: because thresholds are communicated as ambiguous rather than as automatic tripwires, each step allows a context-appropriate response rather than locking either side into a spiral.

**Why would the US ever cooperate, given that it would mean surrendering its military advantage?**

As the report argues throughout, the US has far more to lose by not cooperating — at least under the win-win approach I propose — and could come to recognize this on its own. The coercive cooperation approach is meant to help it see, earlier and more clearly, what is at stake, and to understand that the competitive path is very likely to be chaotic.

**Won't cooperation take a decade, with no guarantee of success?**

Cooperation can move quickly when states face an imminent existential threat — plausibly faster than the ECSC, whose own timeline was already short: the Schuman Declaration (9 May 1950) publicly launched the idea of pooling coal and steel under a supranational High Authority; the Treaty of Paris was signed about eleven months later (18 April 1951); the treaty entered into force and institutions began operating by July–August 1952, roughly fifteen months after signature; and the common market was effectively functioning by 1953. The coercive cooperation approach also structures the situation so that states are unlikely to choose to cheat or defect.

**Would launching a missile at US datacenters destroy our relationship with them?**

If we reach the point of doing so, the relationship has already been broken — because reaching it means we have first had to go through earlier, justifiable escalations to preserve our future.

**You propose shifting public opinion in other states — but can we legitimately interfere in their politics, and how might it backfire?**

This is the question I am least confident about, though I think the basic activity is more ordinary than it first sounds. States routinely and openly try to shape opinion inside other countries: this is the entire premise of public diplomacy, which the US State Department describes as a mandate to "inform and influence foreign publics" through media, cultural, and educational engagement — and nearly every major power runs such programs (the BBC World Service, France Médias Monde, Deutsche Welle, and so on). Leaders also influence foreign publics more directly: when European leaders publicly condemned recent US tariffs, those statements were aimed in part at the American debate, not only at European audiences.

What separates this from illegitimate interference is not the aim of influencing a foreign public — that is universal and accepted — but the means. Open, attributable, reasoned argument is public diplomacy; covert manipulation of elections or the information environment, Russian-style, is not. My proposal is of the former kind: making the public case that non-cooperation is dangerous and that cooperation serves everyone's interest. Where exactly overt persuasion shades into counterproductive meddling I cannot draw a sharp line — and the risk of backfire is real: heavy-handed pressure can be read as foreign interference and harden the very opinion one hopes to shift. This remains the part of the proposal I am least sure about.

To conclude, I should be candid that I remain uncertain whether this overall approach is net positive — but I think it is a line of reasoning well worth exploring.

## Comparisons Between Plan A and My Approach

Entrenching AI Cooperation: A Response to Davidson's Critique of Plan A

Answering Plan A's Dry Tinder Problem by Committing Further to Mutual Vulnerability

todo ajoutes this note: living document, feedback welcome

todo in the davidson comparison and this overarching comparison respell the acronyms on first encounter because reader might not know them.

## Overarching Comparison

As framed in the update note at the top of this report, I consider my plan a near-ideal but plausible target — politically harder to achieve in its ideal form than Plan A, but with elements that can be dropped or negotiated away, and with useful borrowings possible in both directions. This section makes that comparison concrete. -> should i move it before the subsection

**High-level similarities: a fast race is extremely bad, a long or permanent pause is better, and a middle path that entrenches safe cooperation and accelerates defenses is best.** The authors of Plan A argue that fast progress in AI capabilities — especially under badly regulated international competition — vastly increases the risks from uncontrolled AIs. I agree, and my report additionally explores how these dynamics affect power concentration risks, which I think drives some of the design differences between our plans.

The authors are sympathetic to a permanent pause — which they call Plan S — but identify two problems with it. First, a pause would likely fall apart, as many international treaties do (a pattern I explore through the historical examples detailed in the longer report), and the AI race would then restart under worse conditions: geopolitically, because the collapse would occur in a more adversarial climate than today's, and materially, because compute stocks, chip design, and algorithmic insights accumulated during the pause would fuel a faster and more explosive race among more actors — the "dry tinder" problem Tom Davidson identifies in Plan A's own hardware overhang, which would also operate under a full pause to a degree depending on whether chip production and covert research continue during it. Second, a pause cannot stop covert projects: the longer it lasts, the greater the relative power gained by the least conscientious actors who defect — especially dangerous if they reach ASI — a tradeoff I develop under the signpost "slowing frontier progress through imperfectly enforceable approaches creates its own tradeoff.

A middle path therefore appears to be the best option. A key point I may not have underlined enough in my own plan is that it seeks to entrench a form of cooperation, and a state of the world, that seems unlikely to lead to deal collapse — and that enables better coordination toward diminishing AI-associated risks. I believe the Plan A authors share a similar intuition, as well as the view that the middle path should be used to accelerate defenses — they insist particularly on accelerating research aimed at aligning and controlling AIs — though there are instructive cruxes in how each plan goes about this and in what gets accelerated.

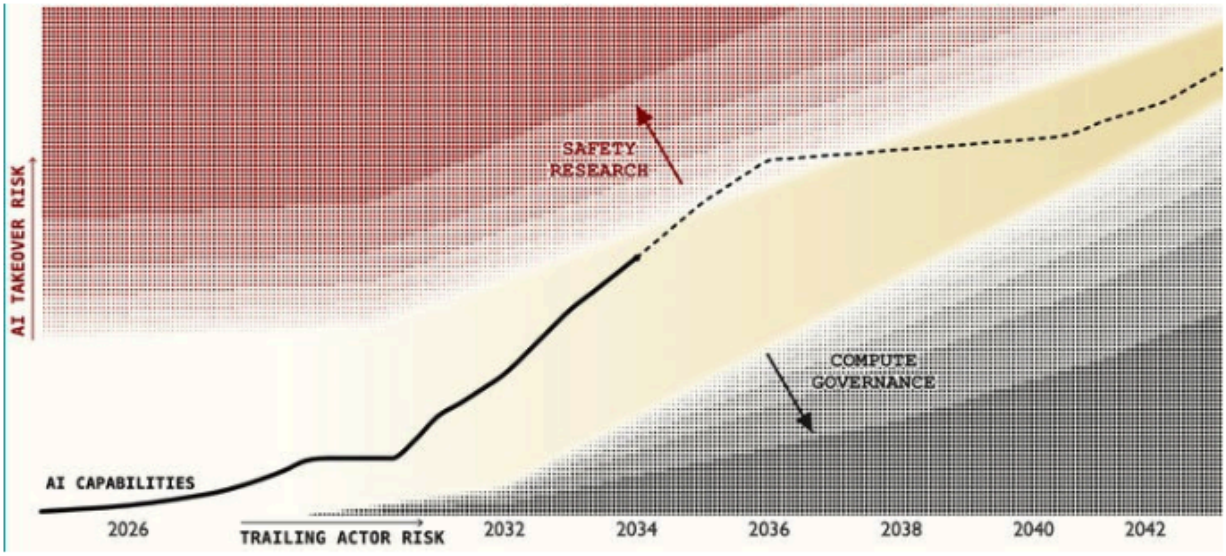



**Figure:** How Plan A threads the needle between two failure modes. The black line traces AI capabilities under Plan A's regulated trajectory: advancing too fast raises AI takeover risk (red, top), while slowing too much increases trailing actor risk (gray, bottom) — the danger that less careful actors catch up to a constrained frontier. Safety research pushes the takeover-risk boundary upward, allowing faster safe progress; compute governance pushes the trailing-actor boundary downward, making a slowdown safer to sustain. Together they widen the safe corridor through which capabilities can advance. *This figure and its description are adapted from Scott Alexander's Introducing Plan A (Astral Codex Ten).*

**Plan A in brief.** Executed perfectly, Plan A unfolds in what can be described as five phases — dated here as in the authors' scenario, though the sequence could ideally begin sooner (a numbering I introduce for this comparison rather than one the authors use; their only official phases are those of their Verification Plan supplement).

- **Phase 1 — Preparation (now–2029):** the low-regret groundwork, guided by the idea that when the political window opens it may open quickly, so everything with a long lead time must be ready in advance. This means funding verification R&D — the flexible hardware-enabled mechanisms and inference-only devices the deal will depend on, without which the later freeze would be far slower and riskier; mapping global compute through intelligence on chip supply chains, datacenters, and stockpiles, so that countries' later declarations can be checked against independent estimates; strengthening export controls and lab transparency measures to keep compute concentrated and traceable; and building the government capacity needed to negotiate and administer such a regime. All of this is valuable even if the deal never comes.
- **Phase 2 — The deal (2029–2030):** the US and China — and by extension the rest of the world — agree to temporarily pause AI capabilities research: each country publicly declares its compute infrastructure (aiming at >99% of the world's, checked against pre-existing intelligence estimates), and all new training runs are halted via inference-only verification — chosen because it is relatively easy to verify while letting the economy keep using existing models. The freeze buys time for a long period of

negotiations over the full regime, while verified and fully transparent training clusters are built.

- **Phase 3 — The steady-state regime (2030s):** capability scaling resumes, by hardware only, under three pillars. The cornerstone is **Total Research Transparency**: all frontier companies must open-source their algorithms and provide access to their models. A **Consortium** of legal projects gates which published algorithms may be implemented. And AI progress is kept reversible through **Mutually Assured Compute Destruction**: new datacenters are built in third countries vulnerable to the rival (China's in Canada, America's in Mongolia), with chips requiring encrypted "continue" messages from the rival and pre-installed destruction mechanisms. Development remains in national companies; pre-deal compute stocks remain national property; and fairness is handled through cash redistribution — including a Citizen's Dividend introduced as automation-driven unemployment rises.
- **Phase 4 — The pause (2035–2040):** a negotiated pause at top-human-expert AI, during which increasingly aligned AIs help improve alignment research, governance, and plans for the future.
- **Phase 5 — Handoff (2040+):** once the AIs can be shown fully aligned, unpause and gradual handoff of control to them.

Plan A is explicitly a recommendation rather than a prediction — in its authors' own scenario it materializes only "in the nick of time" — and the authors themselves flag its main failure modes, including the initial treaty falling through, a flawed safety case being approved, and the deal dissolving before alignment is solved. Several of these are precisely where the cruxes examined below.

**My approach in brief.** What this appendix mainly compares to Plan A's Phase 3 is my proposed end-state: the IOCCR and MS-MAIM combination. The IOCCR is an ECSC-inspired institution whose invariant role is to concentrate, control, allocate, and monitor frontier compute; who develops and tests AIs on that compute is a design parameter — in the plan as presented I chose a single frontier megaproject with all other projects kept substantially behind, but the same architecture could instead coordinate a Plan A-like environment of multiple developers on IOCCR-monitored compute, even under Total Research Transparency (an option value discussed in [section] todo). MS-MAIM is the deterrence regime stabilizing it: the IOCCR holds graduated access restrictions over the compute and models it controls, while member states — initially perhaps only some — hold collective off-switches activatable at coalition thresholds (the precise threshold being negotiable) and, in extreme cases, unilateral intervention for states that have or are granted such capability — enforcement that is graduated and multi-sovereign where Plan A's is binary and bilateral. The deeper aim is to entrench a safe state of the world that is difficult to escape: once established, changing course — pausing, unpausing, reallocating, adjusting capability gates — is a governance decision inside an institution built for fast, checked revision, rather than a treaty renegotiation between rivals; in a transition that will be fast-moving, complex, and hard to predict, this adaptability may be the decisive property.

the intrastate power concentration point logic where as by design of ms-maim and ioccr the cooperation has less chance to collapse for the AI race to restart it helps to prevent another race which is the automation race even if states comply to the cooperation and have AIs of similar level as those who do not quickly use ais to expand their robotics, industrial and military production capaiblities could be at a competittive advantage and a dangerous one if the cooperation collapse, which leads to quickly making citizens dispensable possibly before states have already prepared the systems to remain accountable even if cittizens are dispensable.

Also my design allow the two pace progress...hmm i do not know how much to expand or not these points

The path to this end-state is not exclusive of other frameworks: a US–China MAIM regime, IAEA-for-AI inspections, or Plan A's Phases 1 and 2 would all bring the world closer to it — a fully implemented Phase 3 especially so. But the path this report emphasizes follows the ECSC precedent: willing states — plausibly middle powers first — agree on high principles and institutional design, start cooperating quickly, and use coalition leverage and deterrence communication to pressure the leading powers in. The pattern resembles EU framework laws such as the AI Act — high principles settled first, detailed choices delegated to institutions under member-state checks — with the ECSC showing such an arrangement can be negotiated fast even among recent adversaries. This makes cooperation faster to initiate but politically harder for the US and China to join, since it asks them to trust a supranational institution with compute custody and quotas; weighted seats in the IOCCR's day-to-day and allocative governance ( though never in MS-MAIM's intervention thresholds, lest any power gain the ability to shield itself) could make that ask more attainable. Later sections develop two further contrasts: why a CERN-for-AI with other projects kept behind — which Plan A's Appendix J flags for concentration-of-power and decision-quality costs — can be designed to strongly reduce those risks while preserving extensive transparency; and why MSADS's selective transparency keeps software scaling with compute held in reserve available — the option Davidson tentatively prefers on option-value grounds, since compute can always be built later while a built-up overhang cannot easily be unwound — where Plan A's public transparency forces it into hardware scaling.

## Enforcement and Mutual Vulnerability: Learning from Davidson's Critique of Plan A

This section proceeds in four steps. It first lays out Tom Davidson's critique of Plan A, in two parts: the "dry tinder" its design accumulates — an overhang of idle compute and of published-but-banned algorithmic techniques that would make any post-collapse intelligence explosion drastically faster — and his argument that MACD, Plan A's enforcement mechanism, would prove unstable against it: the overhang demands detection and destruction within days, while MACD's design gives the enforcer strong reasons never to pull the trigger — so that the cure Plan A administers could build a worse disease than the one it is meant to prevent. It then argues that these challenges are best answered by committing *further* to mutual vulnerability, building up the design in tiers of increasing ambition and political cost — which both makes concrete what MS-MAIM (or an improved MACD) could look like, and reveals the panel of possible mutual vulnerability designs, some stronger at a higher political price, others weaker but cheaper. It then turns to the question Davidson's post opens — scaling capabilities by building more hardware, or by algorithmic progress on fixed compute — and why MSADS, unlike Plan A, is free to choose between them: an option value that also lets the regime adapt its course as circumstances and evidence evolve. It closes by arguing that once Plan A's own commitments are accepted — advancing at the same pace as one's rivals, and intentional mutual vulnerability — the remaining political distance to the IOCCR and MS-MAIM is smaller than it first appears.

## Davidson's diagnosis: dry tinder and MACD's three enforcement challenges

Plan A pauses software progress and scales capabilities via hardware. The obvious alternative, as Tom Davidson notes, is the reverse: pause hardware scaling — ban new compute and fab construction — and scale slowly through software. The AI Futures Project acknowledges it is not certain which is best; Davidson is not certain either, but tentatively prefers software scaling on option-value grounds — compute can always be built later, while a built-up overhang cannot easily be unwound. Which is better also depends on evidence we do not yet have: chiefly, whether software progress alone could drive an intelligence explosion on fixed compute, and how stable a slowdown proves to be in the presence of accumulated tinder. The scaling question returns at the end of this section; what decides most of it is enforcement.

**Plan A's cure could create a worse disease.** Davidson opens with an analogy he flags as imperfect: a city halts an approaching fire outside its gates, then builds massive structures to study and guide the fire safely — out of dry tinder. If control is ever lost, the fire now rips through the city far faster. Plan A pauses software progress from around 2030 while compute keeps scaling (roughly 100× by 2033 and another 100× by 2040; AIFP estimates that each 10× of compute speeds an intelligence explosion by about 5× — so, as Davidson notes, a deal breaking in 2033 yields an explosion 25× faster, one breaking in 2040 some 600× faster) — an intelligence explosion that would have taken a year takes weeks, or a day. And these speedups come from the compute overhang alone — the overhang is not only hardware. Because Plan A requires companies to publish their research and then forbids implementing the dangerous findings, a bank of known-but-banned techniques accumulates in the open, adoptable by a defector overnight — a defector would not only race faster, it would start from a higher capability level. Davidson notes the dry tinder has a second dimension beyond this maximum speed: the number of distinct projects that could ignite it, as through the 2030s more and more companies and countries amass the necessary compute and software.

This dry tinder creates two problems for Plan A. First, an **unstable pause**: if just one project executes a secret intelligence explosion, it could have superintelligence within a week and gain a decisive strategic advantage — and fearing that another project might do this is itself a temptation to go first to avoid being crushed, a temptation multiplied by the growing number of projects able to act on it. Second, a **riskier intelligence explosion**: if the deal breaks down and a race to superintelligence follows, it will be much faster and more dangerous than if Plan A had never been done. Davidson is careful about how large this second cost is: small if AI takeover is near-certain on the default trajectory anyway — making an already-doomed explosion 100× faster changes little — but large if compute bottlenecks would otherwise have prevented a software-driven explosion altogether, in which case Plan A manufactures the very explosion it exists to prevent and dramatically raises takeover risk.

**AIFP is well aware of the dry tinder problem — it is why Mutually Assured Compute Destruction is so central to Plan A.** Under MACD, the US places nearly all its datacenters in Mongolia, on chips that require encrypted "continue" messages from China; China's datacenters sit in Canada under the mirror arrangement. If either side detects that the other has broken the deal, it destroys the defector's datacenters and cuts off the continue messages. The hope is that this commitment prevents a fast intelligence explosion despite all the dry tinder lying around. Davidson identifies three challenges for MACD:

**Challenge 1 — detecting fast and reliably.** An overhang that compresses explosions into days demands detection in hours — against a defector able to use millions of expert-level AIs to probe the verification infrastructure for side channels, and able to exfiltrate its model weights

and algorithmic insights just before its datacenters die, keeping a decisive headstart for the subsequent race to rebuild.

**Challenge 2 — destroying fast, including with more than two players.** The defector can defend its sites militarily and may have quietly disabled any on-chip shutdown mechanisms — which is why Davidson requires that no defenses be stationed near datacenters, a self-denial the regime must somehow make stick. And the dynamic degrades once more than two countries could run a fast intelligence explosion. Suppose Europe catches up to the frontier: its datacenters must sit within striking distance of someone — presumably partly of the US, partly of China. The US then can no longer unilaterally destroy Europe's compute; it must trust China to do so — trusting its rival with its own national security — and, worse, China and Europe could jointly stage an intelligence explosion the US could not stop.

**Challenge 3 — actually choosing to destroy.** Davidson's deepest worry is that no one would pull the trigger. MACD quietly assumes that both governments remain durably convinced that a fast intelligence explosion is catastrophic — a conviction that can erode as leaders change or as AI comes to look safe. Suppose it has eroded and the US defects: what does China gain by enforcing? The MAD analogy runs backwards — if China destroys America's datacenters, America destroys China's, so the trigger deters the enforcer rather than the defector. And enforcing returns almost nothing: the defector may already have exfiltrated its weights, and because cold-storage compute tracks pre-deal shares while active compute is roughly balanced, executing MACD moves China from about 1:2 behind the US to 1:10. Worse, the logic of MACD cannot stop at datacenters — fabs could rebuild the compute, and robot-run industry could rebuild the fabs, so the whole AI-powered physical economy must be kept vulnerable, and enforcement means destroying most of it. The political economy then decides the question: if NVIDIA could lobby away export controls, the pressure against immolating a continent's productive base would be orders of magnitude greater.

## A Better Mutual Vulnerability System: What It Could Look Like

*[For readers arriving directly at this section: MS-MAIM — Multi-Sovereign Mutual Assured AI Malfunction — is the deterrence regime this report proposes, designed to work with many actors rather than two; this section lays out in more detail what a near-ideal version of it could look like. The IOCCR — International Organization Controlling Compute Resources — is the institution the report proposes alongside it, with ECSC-inspired governance built for fast decision-making under member-state checks and balances; its role in the MS-MAIM system is detailed within this section (from Tier 3 onward). Fuller context: "My approach in brief" above.]*

When I designed MS-MAIM, I had challenges like Davidson's in mind. This section gives a more detailed picture of what a version of MS-MAIM that mitigates them could look like. Generally speaking, I think these challenges are best answered by committing *further* to the mutual vulnerability logic rather than retreating from it — and by redirecting transparency: deepening it where it serves detection (operations, inference, teams, upcoming runs) while narrowing it where it proliferates capabilities — research being shared with vetted researchers and overseers from every member state, rather than published to would-be defectors. Within that commitment, four design principles follow, each primarily addressing one challenge while reinforcing the others: **deeper operational transparency**, so that violations are detected fast and reliably (Challenge 1); **more redundancy**, so that the system's mutual vulnerability becomes very difficult to quietly escape (Challenge 2); **more, cheaper, graduated, and reversible rungs**, so that intervening is believable and can happen early and fast (Challenges 2 and 3); and **finer granularity**, so that enforcing wrecks no more than the violation requires.

Rather than present the full system at once, I build it up in tiers of increasing ambition — and increasing political cost — each adding features that further embody these principles. Plan A's design could climb these tiers one at a time, or adopt several directly — and in doing so would come to closely resemble the IOCCR and MS-MAIM design this report proposes. Ideally the same vulnerability logic extends to the chip industry — fabs monitored and kept interveinable — while robots need no separate mechanism: their control systems run on the monitored compute and go inert alongside it.

**Tier 1 — Improved bilateral MACD**

Keep Plan A's structure — two rivals, national ownership, third-country hosting — and improve only the enforcement ladder.

Between "cut the continue message" and "destroy the datacenter," insert a graduated, redundant ladder acting on both the compute and its dedicated energy supply. The rungs below are illustrative — one could define more, or split each further — but four capture the shape, in increasing cost and decreasing reversibility. States with intervention capability can, in order:

- **Pause (fully reversible):** stop sending the encrypted "continue" messages, or trip the remote off-switches on the compute, the power link, or the power station — halting the cluster without damaging anything. This is the step a state will actually *dare* to use early, because nothing is lost if the alarm proves false.
- **Sever the power link (reversible over weeks):** if the soft pause fails — because a defector has quietly disabled the on-chip and off-switch mechanisms it physically controls — cut the transmission line or substation feeding the cluster, which large datacenters draw through dedicated, often kilometers-long connections. The cut can run through pre-installed destruction mechanisms placed along the link, with the interveners' short-range strike capability as the fallback; in a multilateralized setting, several interveners can each hold one segment. Power is cut without touching the scarce capital, and the line is repairable, so the action is severe but recoverable.
- **Destroy the generation, then the compute (slow to recover):** the same means — many pre-installed destruction mechanisms, and failing those the interveners' short-range strike capability — applied first to the dedicated generation and then to the chips themselves.
- **Last resort:** ICBMs with conventional warheads, or in the very worst case a nuclear warhead detonated for electromagnetic-pulse effect rather than to kill. Hence compute belongs in a remote area far from population centers, ideally desert, both to limit harm and to make visible and vulnerable the comings and goings of vehicles that might exfiltrate or defend it.

**Redundancy is what makes the lower rungs trustworthy.** A single off-switch or destruction charge can be quietly disabled by whoever controls the hardware — which is the core of Davidson's Challenge 2. So each mechanism is multiplied: many independent off-switches and many pre-installed destruction charges, ideally installed and separately verified by different parties, so that disabling all of them quickly and covertly becomes nearly impossible. And beneath the installed mechanisms sits one signal that cannot be disabled at all: a running cluster radiates heat, visible from satellites — so a defector who has neutralized every switch still cannot hide that the compute is running, and the violation remains detectable even when the cheapest rungs have been sabotaged.

*A note on the energy rungs.* Striking the power supply halts compute within minutes while sparing the chips and fabs — the scarce capital, replaceable only over years — which is what makes it usable where outright destruction is not, and it works even against a defector who has disabled on-chip shutdowns, since compute cannot run without power. It is also hard to circumvent at frontier scale: replacement generation for a gigawatt-class cluster would mean hundreds of mobile gas turbines — visible, undefended by the regime's siting rules, and themselves a flagrant violation inviting a further strike. That same scale limit cuts the other way in the regime's favor: the regime can pre-position modest mobile generation under collective custody, deployable only on authorization, so that a strike need not collapse a cluster permanently — a useful fraction can be brought back quickly for a state returning to compliance, which lowers the bar to using the rung at all.

Two further refinements. Physically separating training compute from inference compute makes intervention on training far cheaper, since freezing a training cluster need not black out the services an economy runs on. In a two-party setting this separation is only partially reliable — telling training from inference rests on computational signatures and sampling, a surface a determined defector can probe — but it is better than nothing, and it becomes more robust under the fine-grained inference transparency introduced in Tier 3.

These interventions could be made **automatic** — fired on defection detection by reliable AI systems, without human sign-off — removing the hesitation Challenge 3 identifies. If the lower rungs are reliable, a strong international norm can also form: a state that moves to destroy all compute with no good reason reveals itself as adversarial and makes many enemies — a norm that holds only because cheaper options exist, so jumping to destruction is never forced.

***Additional political cost:*** *near zero — this tier changes only how enforcement escalates, not who owns or holds anything. Stability gain against Challenges 2 and 3: large.*

**Tier 2 — Multilateralized MACD: a few states with unilateral intervention capability**

Replace the bilateral finger with several. The problem this solves is the one Davidson raises for MACD once more than two countries reach the frontier: suppose Europe catches up, and its datacenters must sit within striking distance of someone — partly the US, partly China. The US can then no longer unilaterally destroy Europe's compute; it must trust China to do it — trusting its rival with its own national security — and, worse, China and Europe could jointly stage an intelligence explosion the US could not stop. Distributing intervention capability across several states removes both problems, and removes the regime's dependence on any single rival's willingness to act.

Each actor's compute is held in a third country where every member maintains a small, roughly equal military presence with strong ability to intervene and weak ability to prevent intervention. Here I differ slightly from Davidson, who proposes the compute's owner keep no troops on site: it seems preferable that the owner retain enough presence to stop others from removing its destruction systems and to fire on the AI infrastructure itself — but too little to defend its compute against a determined strike, and in particular no missile-defense capability sited there. This guards against a specific failure mode: if two members collude with the host to remove a third's destruction devices and seize its compute, an owner that cannot destroy its own compute is defenseless. So concretely: China's compute sits in Canada, run by Chinese non-military personnel under intense surveillance from other countries' militaries, with a Chinese presence able to destroy its own compute but unable to defend it against collective intervention. Siting could go further still: placing major clusters in international waters — as Plan A proposes with its

floating datacenters — would make interventions against them less escalatory than strikes on sovereign territory, and the facilities harder for a defector to seize or defend.

Which states hold a finger is itself a design choice, and it can grow over time — US, China, then the EU once it catches up. Notably, these need not be favors the leading powers grant: a state whose national security is genuinely threatened by others' frontier compute can press for a finger of its own, exactly as the report's main argument describes — because this is throughout a situation of mutual vulnerability, in which every capable actor has both something to fear and something to bargain with.

***Additional political cost:*** *higher than Tier 1 — more actors gain a finger, which is a real ask. But it can be lower than it first looks, and in some framings even preferred by the giants: extending a collective, reversible trigger to middle powers that already hold ICBMs (France, the UK, Russia) grants no genuinely new capability — those states can already strike compute anywhere on earth — while converting their existing catastrophic, unilateral option into a safer, graduated one. What reads as a concession can therefore also be a stabilization the US and China have reason to want.*

**Tier 3 — The IOCCR gains custody of the compute**

*At this tier an international institution, rather than national governments, holds and operates the frontier compute, which helps reduce the adversarial character of the whole arrangement. Several of its benefits can be partially approximated without it, only less easily coordinated.*

*Custody does not mean unchecked control, and the institution's mandate is deliberately narrow. The IOCCR is the only actor permitted to run training above the agreed thresholds, and it may use inference compute only for limited, specified purposes; member states retain access to inference for their own approved uses, through the quotas described below. Its powers are held inside ECSC-style checks and balances — leadership appointed by member states for fixed terms, a council able to constrain the executive, and the sovereign intervention layer of the earlier tiers standing outside its channels as a backstop. So what custody transfers is operation and title, not sovereignty over how the technology is used, which remains distributed among members.*

*Within that mandate, custody buys three things the first two tiers could not.*

***It hardens the human layer.*** *The IOCCR is staffed by an international team operating under constant surveillance from all member states (against compensation), its vetted members committing — as the ECSC's High Authority did — to serve the community's interest, ideally checked regularly under lie detectors. Because it is the institution, not a national government, that develops the AIs, a secret intelligence explosion becomes far harder to start: coordinating one would require a conspiracy across surveilled, mixed-nationality teams rather than a government instructing its own employees. And before a training run begins, the algorithms it will use can be shown to member states and cleared by a vote — a shared "ignition check" that no single party can bypass, and one a coalition design without full custody could adopt too.*

***It makes transparency robust where it was only partial.*** *All of the AIs' interactions with the external world — that is, inference compute — can now be finely monitored: with competition dynamics broken, the regime can afford to devote serious compute and AI assistance to controlling AIs and tracking what is done with them. This closes the route of tampering with verification by running disguised training on nominally inference-only hardware, and upgrades*

*the training/inference separation introduced in Tier 1: with inference genuinely transparent, freezing a violating training cluster without touching the compute an economy runs on becomes considerably more dependable.*

***It enables the institutional reserve — and with it, reversibility.*** *The IOCCR holds a large fraction of compute unpowered in reserve — especially feasible under a software-scaling scheme — in remote, well-protected sites, guarded against theft and sabotage but not against collective intervention. This makes intervening against active compute far less costly, since capacity can be reinstalled from the reserve rather than rebuilt from fabs. Crucially there is no pre-deal national reserve: unlike Plan A's cold storage, which survives dissolution in the defector's hands and leaves states that held a smaller pre-deal share with little incentive to enforce, an institutionally held reserve is forfeited by defection. Instead of recovering pre-deal ratios, the IOCCR coordinates access to an agreed fraction of inference compute for each state's use of IOCCR-developed AIs — more for the US initially — with quotas ideally renegotiated over time toward population and need.*

*Additional political cost: potentially substantial — this is the tier that asks states to give up title to their compute, not merely to accept vulnerability. The closing section argues the real size of that ask is smaller than it first appears: a state that has already accepted capability parity and mutual vulnerability has, in practice, little left to lose but a cheap exit — the everyday sovereignty it actually uses is access to its inference quota, which custody preserves. And for a government that has come to fear the race resuming under worse conditions, irreversibility stops being a cost at all: entrenchment is then the very thing being bought.*

**Tier 4 — The IOCCR gains graduated access-restriction rungs**

The institution can now intervene at a granularity no kinetic rung can match, by restricting access to what it develops and controls: cutting access to the best models, then to all advanced models, then reducing general compute, for periods scaled to the gravity of the violation. This is the finest-grained and most reversible rung of all — useful for targeted response against a state misusing reserved or military compute in unapproved ways, or illegally retaining unmonitored compute, since cutting its access is a cheap, credible way to make such behavior unattractive. Reinforcing this, the IOCCR can use the best AIs to produce *defense-oriented* military technologies operated by states (but never offensive ones), sharing the results among all members — which both deters cheating and shrinks the incentive to keep large independent military compute. These interventions, too, can be set to fire automatically on defection detection.

*Additional political cost: low, given Tier 3 — once the institution holds the compute, access restriction is simply the cheapest way it can already act, and the least destructive rung in the whole system.*

The institution can now restrict access to what it develops and controls: cutting access to the best models, then to all advanced models, then reducing general compute, for periods scaled to the gravity of the violation. This is the finest-grained and most reversible rung of all — especially useful for targeted response against a state misusing reserved or military compute in

unapproved ways, or illegally retaining unmonitored compute, since cutting its access is a cheap, credible way to make such behavior unattractive.

Reinforcing this, the IOCCR could use the best AIs to produce cutting edge *defense-oriented* military technologies delivered to and operated by states (but never offensive ones) — which both deters cheating for dominance seeking and shrinks the need to keep large independent military compute. These interventions, too, can be set to fire automatically on defection detection.

*Additional political cost: low, given Tier 3 — once the institution holds the compute, access restriction is simply the cheapest way it can already act, and the least destructive rung in the whole system.*

**Tier 5 — Full MS-MAIM: extending the trigger to smaller powers**

The most demanding tier politically is granting smaller powers access to the collective off-switch. A single actor could in principle turn reckless and move to destroy all compute — but the nuclear states already roughly hold that capability, so giving them a cheaper, less escalatory collective trigger takes nothing away and adds a safer option; for non-nuclear powers, unilateral capability might be unwise, which is exactly why access to the usable rungs runs through coalition voting thresholds rather than individual fingers. This both extends real intervention power to smaller members and makes reckless unilateral action less likely — the version of the system this report ultimately defends.

*Additional political cost: the highest of all, since it asks the leading powers to accept that a coalition including smaller members can reach the trigger. What makes it bearable is that the trigger is collective, not unilateral: no single small state can act alone, so what is granted is a share in a threshold, not a finger — extending the reach of deterrence without multiplying the risk of its misuse.*

The most demanding tier politically is granting smaller powers access to the collective off-switch. A single actor could in principle turn reckless and move to destroy all compute or use that threat to gain advantages — (but the nuclear states already roughly hold that capability, so giving them a cheaper, less escalatory collective trigger takes nothing away and adds a safer option -> point already in tier 1 should i restate?); for non-nuclear powers, unilateral capability might be unwise, so access to the lower rungs could be limited to only off-switches or off-switches and destruction of power-links and stations and could run through coalition voting thresholds rather than individual fingers. This both extends real intervention power to smaller members and makes reckless unilateral action less likely.

having all five tiers is the version of the mutual vulnerability system that I think would be ideal.

*Additional political cost: the highest of all, since it asks the leading powers to accept that a coalition including smaller members can reach the trigger. What makes it bearable is that the trigger is collective, not unilateral: no single small state can act alone, so what is granted is a share in a threshold, not a finger — extending the reach of deterrence without multiplying the risk of its misuse.*

## Hardware or Software Scaling: Plan A Is Locked In, the IOCCR Can Choose

The previous section argued that better enforcement designs are available; this one builds on some of the enforcement design concepts to examine a choice both plans face: scaling capabilities through hardware, through software, or through a mix. Which is best is genuinely uncertain — though, as Davidson argues, starting with software better preserves the option to switch. The section then argues that Plan A is largely locked into hardware scaling, while the IOCCR and MS-MAIM combination can implement either approach or a mix, improving each in the process — and can provide the cooperation and infrastructure needed to switch between them quickly and at lower cost as evidence arrives.

**Which scaling approach is safer is hard to know in advance.** Both the AIFP and Tom Davidson seem unsure. The case for hardware scaling is real — restricting new fab and datacenter construction is a large economic ask, and the pace of hardware is a visible, verifiable dial where the pace of software is hard to measure. And as Davidson notes, the answer depends on unknowns we will only learn about over time: whether MACD can be implemented quickly and reliably — especially decisive for hardware scaling, which accumulates the larger compute overhang; and whether covert projects can be eliminated, or at least reliably kept away from dangerous capabilities — especially decisive for software scaling, where a leak of algorithmic progress matters more, and which itself depends on an open question: whether scale-dependent or hardware-co-specialized software can be developed, such that covert projects could not run leaked techniques efficiently on their own hardware.

**Starting with software scaling nonetheless seems preferable, on option-value grounds.** This is Davidson's tentative view, and it rests on an asymmetry: compute can always be built later, while a built-up overhang cannot easily be unwound — pivoting from hardware scaling to software scaling would require destroying enormously valuable datacenters and fabs, which few will ever accept.

**TRT is ill-adapted to software scaling.** Publishing all research means algorithmic progress reaches defectors legally and in full. This could be partially tempered if scale-dependence and co-specialization techniques pan out — they would depreciate what publication hands out — but they cannot make it excludable, especially for state-level defectors who could still hold quite large unmonitored compute. And publication is not the only leak TRT builds in: since it also mandates broad access to frontier models, a covert project can extract capabilities by distillation — mass, well-chosen queries — a campaign that the fragmented logs of dozens of companies are ill-placed to catch. Compute therefore remains the only input Plan A can make firmly binding — hardware scaling is largely forced on it by its own transparency. Davidson's preferred option, in other words, seems much more readily implementable under an architecture built for selective transparency and infosecurity — such as a CERN for AI, one of the development configurations the IOCCR's governance could choose to implement.

**Software scaling makes tight cooperation even more necessary.** At an equal pace of progress, scaling through algorithmic innovation more likely produces AIs that are harder to align — new techniques with poorly understood failure modes. This makes it even more important to break competitive pressures, so that enough resources reliably go to alignment and control, and to coordinate promptly on how much more they need as evidence arrives — which is what the IOCCR and MS-MAIM combination was designed for.

**The IOCCR and MS-MAIM design can implement either approach — and holds the reserve that eases both the choice and the switch.** Because capability research is shared among

vetted researchers and overseers from every member state rather than published, software progress can be paced without arming those outside the regime — and the monitored, aggregated inference layer closes the routes, disguised training and mass-query extraction, that fragmented verification leaves open. The design can therefore scale through software while holding a large fraction of compute unpowered in the institutional reserve described in Tier 3 — neutralizing the hardware overhang rather than merely not building it, and making intervention against running compute reversible, since capacity can be reinstalled for compliant members. The same logic covers production: fabs shift toward inference chips or idle under the regime's watch, and training compute is reallocated or powered down rather than destroyed — which is what makes the pivot survivable, since the value is preserved while the ignitability is not.

While concentrated algorithmic progress is itself dry tinder if the institution is captured, a captured IOCCR still faces the sovereign layer — off-switches, strikes on its dedicated power, unilateral intervention — none of which runs through its own channels. The failure condition is capture *plus* defeating multi-state enforcement — narrower than Plan A's default, where the fuel is distributed to everyone by design.

**ECSC-style governance lets the scaling mix shift promptly — treaty terms cannot.** Even a Plan A freed of TRT's lock-in would face a second lock: its scaling strategy is a treaty term, revisable only through renegotiation between rivals — slow, and reopenable to blockage at every step. Under the design proposed here, the same revision is an internal governance act: as evidence arrives — on things like whether scale-dependent or hardware-co-specialized software pans out, whether some algorithmic techniques prove too hard to align and are better neutralized by stopping software progress, or how reliably software pacing can be measured and verified — the IOCCR can shift the mix toward more hardware, reinstalling reserve compute and redirecting chip production through its ordinary procedures, or even adopt Total Research Transparency should its benefits come to outweigh its proliferation costs.

**In sum: Plan A is locked twice, and the ability to update may itself be a safety property.** Plan A is bound structurally into hardware scaling by its transparency, and procedurally into its initial bet by its treaty form. The design proposed here starts from Davidson's tentatively preferred option — software first, with the hardware overhang ideally neutralized in reserve — and, in a transition where evidence may arrive fast and windows may be short, retains the capacity to correct course promptly as the world reveals which bet was right.

## Political Buy-In: Entrenchment Is the Selling Point — and the Extra Ask Is Smaller Than It Appears

All of this raises the obvious question: who would ever accept such a regime?

**For governments that have internalized the risks, entrenchment is not the cost — it is the product.** AIFP's own Appendix B argues that a Plan A-style deal is incentive-compatible for almost everyone concerned about loss of control or concentration of power — everyone, that is, except those in whom power would concentrate by default.

Push that reasoning one step further. A government that has internalized these risks — perhaps after clearer warning signs: serious loss-of-control incidents from open-weight or agentic systems, states voicing alarm at their strategic vulnerability and signaling willingness to escalate

against non-cooperation, continued rapid job displacement and rising public concern over concentration of power (dynamics discussed in [the published summary]) — will most likely want cooperation to be entrenched. It will also have internalized the lesson of AIFP's own critique of Plan S and of Davidson's dry tinder argument: any slowdown accumulates the fuel that makes its own collapse catastrophic. A catastrophic and likely failure mode of cooperation, then, is not signing a bad deal but losing a good one. Such a government has a further, familiar fear: its successors. It cannot guarantee that the next administration will care about these risks at all — and a reversible deal is exactly what a careless successor can dissolve, under worse conditions than today's.

Governments already solve this problem with proven tools: binding their successors through structures that are costly to unwind — constitutions, independent central banks, alliance integration. Seen this way, well-designed irreversibility built on mutual vulnerability stops being the regime's most demanding feature and becomes its selling point: the state is not surrendering control to others so much as buying insurance against its own future selves, and against the collapse scenario it has most reason to fear.

**Entrenchment also removes a second race — the one that survives compliance.** A deal that can plausibly dissolve leaves states the fear-driven incentive to keep racing on automation even while complying — replacing human workers with AIs wherever it is more efficient, expanding robotic, industrial, and military capacity in ways that make citizens dispensable, while leaving less time and incentive to set up the systems that would keep governments accountable regardless. A regime unlikely to collapse removes that fear, and with it the automation race itself.[1] *A dynamic introduced in [the published summary], and one I aim to develop further in the forthcoming full comparison.*

**And for governments only half convinced, the remaining distance from Plan A to custody is smaller than it appears.** A state that has entered Plan A's regime has committed to intentional mutual vulnerability and to capability parity with its rival — parity between the two giants being the deal's very object, even if parity with other members emerges from the regime's mechanics rather than by design. Either way, it has already surrendered most of what ownership of compute strategically confers: it cannot use its compute to pull ahead, and its compute's very operation already depends on a rival's continued consent (the encrypted "continue" messages of MACD).

What national ownership still preserves is essentially a *cheap exit*: cold storage, national companies, and publicly available algorithms to rebuild with if the deal dissolves. But this cheap exit is precisely what makes the dry tinder lethal — the deal can dissolve and leave every party its fuel — and it is also much of what makes Challenge 3 bite: as noted there, cold storage tracks pre-deal shares, so an enforcer with a worse pre-deal ratio stands to lose relative position by enforcing.

The additional ask of custody is therefore not day-to-day sovereignty, which the design preserves through each state's use of its inference quota; it is the renunciation of a low-cost exit. And softer versions shrink the ask further: for example, an IOCCR that rarely restricts access and operates under high intervention thresholds, with sovereign enforcement initially reserved for the US, China, and other states with ICBM capabilities — for whom the collective trigger is not a new power but a safer substitute for the catastrophic finger they already hold.

## What's Next

This post has taken up one crux of the comparison between Plan A and the IOCCR and MS-MAIM design — enforcement and mutual vulnerability, extended to the scaling question it decides. Further cruxes — public versus selective transparency, custody and the concentration of power, the military dimension, and the full political buy-in comparison, including how middle powers could initiate the regime and pressure the leading powers in — will follow. This is a living document: I expect to revise it as I receive feedback, and pushback is very welcome — including on where I may have misread Plan A or Davidson.

—--------

In my approach the IOCCR (which to be clear in a ideal system would have vetted members that against compensation would operate under constant surveillance from all countries of the ioccr and accept lie detection and similar to the ecsc make the commitment to serve the community not their own national interest) can use AIs for limited purposes like accelerating the defence and improving the reliability of ms-maim could typically an instance of it and verifying that states do not use their ais to compromise verification infrastructures

As with AI redlines, IAEA-for-AI bodies, and institutionalized MAIM, Plan A could be the step before the IOCCR and MS-MAIM combination gets implemented.

States operating under Plan A may find that its transparency measures are technically harder to verify than expected, that its coordination is too slow for fast-moving AI developments, or that functions like accelerating defenses and redistributing power across states need tighter machinery than a verification consortium provides — and conclude that something closer to the IOCCR + MS-MAIM combination is better fitted. The transition need not be unanimous: it could also come about if enough states reach that conclusion and apply sufficient pressure on the others, using the levers described in Section 4.
A concrete trigger: scenarios where advanced AIs prove difficult to align — or cannot yet be shown aligned — and states realize they need far more reliable collective capability to halt and switch off potentially dangerous systems than a diffused, many-company frontier allows, as well

as stronger means of preventing the development of too-advanced AIs in the first place. The IOCCR and MS-MAIM combination could also come to be seen as the better alternative outright, as it may prove less likely to dissolve — whether by reverting to dangerous racing dynamics or through power being captured.

**MAIL:** My approach is variant of what you call a "CERN for AI. An interna- tional project to develop frontier AI, with all other projects regulated to be sub- stantially behind in capabilities." In my approach I propose complex multilayered institutional systems partly inspired by the ECSC, and complex compute deterrence systems that together are meant to combat the concentration of power risk of this approach. I also argue that this approach could help better coordinate to break interstate competition dynamics in ways that help prevent pressures that lead to intrastate concentration dynamics. So my approach share a similar and possibly more elaborated deterrence logic and also the point of view that inference should be monitored at a fine grained level. I also aim at combating the bad decision problem of this approach that you mention by having selective transparency. In addition to concentrating the development of the most advanced AIs in one international project another key difference of the plan I propose is that middle-powers like france signal that absent tight forms of cooperation that makes them existentially safe they would use their credible deterrence capabilities to pressure the bigger power to cooperate, mostly via signaling willingnes to intervene past in a way that moves publics and politics in other countries to wakeup earlier to the threats of non-cooperation. that pressuring logic is subtle and to give it credit it should be read in full and i think this could be a very valuable way to reach safe forms of cooperations in a timely manner. + has many other differences explored in the article

despite a few overarching differences our designs are similar in many aspects

**Convergences:** in my design there is total research transparency between countries of the consortium but this transparency is limited to vetted scientists, ai researchers etc, people that agree to be under constant oversight

## Answering Appendix J's objections to the CERN-for-AI family

Plan A's Appendix J assesses "CERN for AI" — an international project developing frontier AI with all other projects regulated substantially behind — and raises two costs: concentration-of-power risk, and worse decisions, including on technical safety, from less transparency and less broad deployment.

**MSADS is built around the concentration-of-power objection rather than against it.** MS-MAIM's dual enforcement, the collective off-switch activatable by a minority of members, the unilateral backstop, and the ECSC-style Council gate all exist to make concentration of compute safe rather than to deny that it is dangerous. The objection is not one MSADS disputes; it is the problem the architecture was designed to solve.

**Selective transparency and monitored deployment narrow the decision-quality objection.**

- not sure my approach is better
- total research transparency could be added to my plan

Plan A's own footnote points toward the answer — "small public deployments and selective transparency." On deployment, MSADS already meets the objection halfway: it centralizes development, not use. States access IOCCR-developed models for approved purposes on the institution's tracked inference compute, and because all interactions between AI systems and the external world are monitored, the IOCCR would see behavioral evidence aggregated across every approved deployment worldwide — where each of Plan A's diffused companies sees only its own logs. A rare failure mode appearing three times globally is three unexplained incidents at three rival firms under Plan A; under MSADS it is a detected pattern.

On transparency, Plan A's argument has moved me, and MSADS should be more specific than it currently is. Reflecting on Total Research Transparency, I would now propose a two-tier arrangement: capability-relevant research and algorithmic progress shared among vetted overseers and cleared developers from every member state — giving qualified reviewers from rival countries adversarial incentive to catch one another's errors, which on safety-critical claims may be worth more than the volume of public review — while governance artifacts such as model specifications, safety cases, and capability evaluations are published openly, preserving external scrutiny and democratic accountability without handing capability insights to non-members or would-be defectors. The separation is not perfectly clean: interpretability results in particular often reveal architecture. But it recovers much of what public transparency buys at a fraction of its proliferation cost.

Whether this yields better decisions than public research and unmonitored deployment remains genuinely uncertain — different architectures fail differently, and a diffused frontier surfaces failure modes a single developer's models never would. The claim is only that the tradeoff Appendix J identifies is narrower than its framing suggests, not that it disappears.

**Accelerating defenses versus freezing military R&D.** One of the sharpest contrasts between the two plans concerns military technology. In Plan A's Appendix S, military-relevant R&D is deliberately held far behind general scientific progress: because Plan A diffuses capability across many actors, letting military applications advance at the ~10× rate of everything else would risk arms races an order of magnitude faster than the Cold War's — compressing the time humans have to decide, in a period whose close calls already came uncomfortably near catastrophe — and "wonder weapon" surprises where one state discovers a decisive capability the other did not know to defend against. So military use of AIs significantly above the pre-deal level is prohibited, at the cost, as they note, of an accumulating "overhang" of low-hanging military advances that raises both the incentive to defect and the stakes of collapse. MSADS takes a different approach: it accelerates the development, and ideally the deployment, of defense-dominant military technologies. The IOCCR develops these technologies — cyber- and bio-defenses, air-defense systems, anti-drone directed-energy weapons, more resilient nuclear second-strike capabilities — shares the resulting knowledge among all members, and could even coordinate their production at scale so that every member is better defended. This leaves little room for secret parallel discovery, and so much less wonder-weapon risk: some exceptional offensive capability might still be found, but if defense sits at the frontier of what is possible, decisive offensive surprises become considerably less likely. It also avoids the arms-race dynamic, since there is no reaction cycle to compress when members develop the same defensive capabilities jointly and can see each other doing it. And it avoids a

defection-inducing overhang, because MSADS harvests those low-hanging advances defensively rather than leaving them for a defector to seize — keeping the defender at the frontier and making defection less attractive. Where a defensive advance could erode the retaliatory capabilities that resilient second-strike forces are meant to preserve — as sufficiently capable air defense might — the IOCCR coordinates development and deployment to keep deterrence intact, a judgment no unilateral actor could be trusted to make about its rivals.[1]

footnote I expect this tension to be less acute than it first appears: as Appendix B argues, building a missile shield that reliably defeats a sophisticated salvo is considerably harder than building the means to break one, so defensive advances are unlikely to erode second-strike survivability as quickly as they improve defenses against other threats

**(My design could include offencive military robots as part of the robots that could be destroyed )Whether military robotics can be included.** Both plans extend enforcement into the physical economy. Plan A requires robots' control systems to run on monitored cloud compute by law; MSADS reaches the same place through the ECSC's logic, since robots running on traceable compute make industrial activity visible and leave an aggressor's advanced AIs and robots alike vulnerable to coordinated intervention. But Plan A exempts military robotics — and in its own collapse scenario, that exemption is precisely what leaves drone swarms operative after the datacenters burn and a billion civilian robots go limp. The exemption reflects a real constraint rather than an oversight: no state will accept a rival's hand on its army's off-switch, and under a bilateral hostage arrangement that is exactly what including military robotics would mean. MS-MAIM's multi-sovereign trigger changes the ask. Because activation requires a substantial coalition of members rather than any single rival, no state can be unilaterally disarmed by an adversary — which makes including military robotics negotiable in a way Plan A's architecture does not. Whether it would actually be conceded is another question; a coalition off-switch over one's armed forces remains an extraordinary ask. But the difference matters: under MSADS it is merely very hard, where under Plan A it is close to structurally impossible — and closing that exemption is what would prevent a collapse from leaving the most dangerous capabilities running while everything else goes dark.

## Preventing Intra-State Power Concentration

An important focus of my Section 2 that critique other approaches was the point that under them, If AIs are sufficiently capable,  competitive pressures to replace human workers with ais are still too present even in scenarios where states have similar ai capabilities partly due to the possibility of deal dissolution but also basically just due to normal commercial pressures. This leads to the situation where citizens are more dispensable which could erode states accountability as in petrostates. I think that Plan A doesn't address this problem sufficiently and I want to add details here as to how this problem could be better addressed by my approach

**Multi-pace AI progress: Coordinating the Gating to AI Systems for Different Types of Applications.** In my approach I say that the best AIs could be used to accelerate defense and I that the power to halt progress is in the hands of states, separately from the body that supervises AI development, which mirrors the way the ECSC's Council of Ministers could constrain the High Authority — the ECSC's executive — on key decisions. I realize that I have underspecified what is the potential of these points. My idea is that the council of ministers can regulate has the power to regulate different paces of AI progress for different things, the easiest

version of that is to imagine only two paces of AI progress: 1) The best AIs are used only to accelerate the development of better defenses and also the research in AI safety and 2) the council can for example decide that the level of AIs accessed by businesses remain (everywhere) below a threshold that would make most citizens dispensable until they have established in their countries the systems that prevent dangerous intra-state power concentration. Again the minority approach could be used to tild things thoward caution that is if a large enough minority think ais above a given threshold shouldn't be used in the economy then this is prevented.

**Improving Inter-State Power Concentration.** Whilst in Plan A they have systems to redistribute wealth, maybe they could do better to redistribute power.

—----------

**Europe's role and recommendations.** France and the UK currently possess credible deterrence capabilities that, if maintained proactively, are unlikely to be undermined — and, together with other European states, have stronger incentives to pursue international cooperation than expected post-AGI leaders such as the US and China, making Europe well-positioned to initiate MSADS. European countries should begin researching compute verification mechanisms and governance frameworks while engaging allies and potential rivals in preliminary discussions. Early forms of cooperation could start with models like CERN for AI or Intelsat for AI. To accelerate broader coalition formation, a coalition with open membership could start small with Europe's largest economies or even just the Benelux countries. If such a coalition forms around the IOCCR and MS-MAIM, domestic pressure — combined with the growing strategic costs of remaining outside — could bring even authoritarian holdouts to eventually participate, especially if the US is part of the coalition.

France and the UK are not currently preparing for the possibility of an adversarial leading AI country having millions of AI "geniuses" and, later, potentially billions of systems vastly smarter than human geniuses. In particular, they are not preparing for the possibility that such systems could coordinate large-scale cyberattacks against their nuclear deterrents or dramatically accelerate progress in anti-submarine warfare — eroding confidence in their retaliatory capabilities to the point where their leaders might feel pressured into dangerous launch doctrines or unfavorable strategic concessions. They should hedge against such scenarios by ensuring the resilience of their nuclear forces.

For example, France could reintroduce a land-based leg to its nuclear forces, ranging from fixed launcher sites to the more ambitious option of deploying French-owned, French-operated road-mobile launchers across willing host countries.

France and the UK should define intervention thresholds and communicate credible escalation ladders to potentially resistant states in advance. States that understand the strategic situation should shape public opinion in resistant states by communicating the risks of non-cooperation and the benefits of MSADS. Concrete actions like forming an initial coalition and preparing deterrence strongly signal the seriousness of the situation.

MSADS is valuable even without universal IOCCR participation — building a coalition and preparing credible deterrence gives Europe better defenses against adversarial scenarios. The window for establishing the IOCCR and MS-MAIM under favorable conditions is rapidly closing: research and preparation must begin immediately.

— -------

**MSADS addresses post-AGI risks and enables positive futures.** The IOCCR prevents unchecked corporate competition that risks consolidating power in unelected hands. It removes states' pressure to use AGIs for rapid military development out of fear they cannot defend against rivals who would

States that understand the strategic situation should be vocal about the fact that forming the IOCCR and MS-MAIM regime also serves resistant state's citizens best interests. Concrete actions like forming a initial coalition and the deterrence preparation helps signal the seriousness of the situation, and influence other states public opinion. If middle powers coalesce around the IOCCR and MS-MAIM, demonstrating their feasibility and mutual benefits, public opinion in holdout democracies could create domestic political pressure to join, while the growing strategic costs of remaining outside pressure even authoritarian holdouts to eventually participate, especially if the US is part of the coalition.

**MS-MAIM is the only stable equilibrium.** States that resist win-win cooperation despite understanding its feasibility and threat to others signal potential hostile intent. Each step of resistance further justifies the warning and potential execution of higher escalations from rivals. Consider Russia's position if the US achieves post-AGI dominance without transparency over compute usage — growing uncertainty about their nuclear retaliatory capabilities would create pressure to escalate, potentially to limited nuclear strikes, before the US gains decisive advantages. Pragmatic leaders would recognize that resisting MS-MAIM is both

---

**The Mutual Assured AI Malfunction (MAIM) Framework.** The authors of Superintelligence Strategy argue that no superpower would remain passive while a rival risks transforming an AI lead into an existential threat. Instead, capable nations would likely threaten to preemptively

sabotage AI projects threatening their survival. Given that AI assets are currently highly vulnerable to disruption, the authors predict an informal regime of AI deterrence — MAIM — would emerge organically. MAIM would initially operate through espionage to monitor AI progress, with states communicating escalation ladders — beginning with cyberattacks, chip supply chain disruption, and AI infrastructure sabotage, potentially escalating to conventional missile strikes if rivals pursue strategic dominance. (todo compress to: states communicating escalation ladders no)

Rather than letting this regime develop chaotically, Superintelligence Strategy proposes preparing mechanisms to proactively establish a stable MAIM regime. In its mature form, MAIM would enable states to monitor each other's AI projects and safely intervene when rivals make strategic bids for dominance, allowing states to advance AI capabilities at a similar, regulated pace — maintaining balance while avoiding dangerous volatility from unchecked, potentially fast recursive feedback loops.

**MSADS (Multi-Sovereign AI Deterrence Strategy) extends the AI deterrence logic with safely concentrated compute control.** While AI deterrence and a mature MAIM regime would temper racing dynamics and great power conflicts, optimally addressing the full spectrum of AGI challenges — especially gradual power concentration and democratic erosion — requires a more comprehensive framework. These risks emerge even if AI reaches AGI (and possibly ASI) without explosive recursive self-improvement.

MSADS addresses these broader challenges by aiming to centralize control of compute resources in an International Organization Controlling Compute Resources (IOCCR) with robust checks and balances. Member states retain access to compute for approved uses, with transparency enabling intervention against misuse.

Among its key functions, the IOCCR would build a mature Multi-Sovereign MAIM (MS-MAIM) regime where multiple independent sovereign actors can monitor and safely halt dangerous AI development and misuse globally. To properly function, MS-MAIM requires compute traceability and usage transparency, maintained AI asset vulnerability, and escalation frameworks. This distributed intervention capability prevents tyrannical power concentration from states, coalitions, corporations, or the IOCCR itself.

**MSADS addresses post-AGI risks and enables positive futures.** The IOCCR prevents unchecked corporate competition that risks consolidating power in unelected hands. It removes states' pressure to use AGIs for rapid military development out of fear they cannot defend against rivals who would — pressure that incentivizes removing humans from control loops and automating totally loyal militaries and bureaucracies, making durable authoritarian entrenchment easier than ever. These same conditions also mean that prematurely delegating key decisions to AI systems not yet proven aligned makes them more capable of disempowering humanity if misaligned. This is especially concerning as more than half of surveyed AI researchers assign at least a 10% probability to advanced AI causing human extinction or permanent disempowerment.

Through the IOCCR and MS-MAIM, states can break these competitive pressures, making it easier to coordinate around positive futures: accelerating the defense against post-AGI threats, equitable sharing of AI-generated benefits, coordinating on shared global challenges like climate change, preserving democratic systems, and implementing a global AI development pause if needed.

To develop safe advanced AIs and accelerate defense, the IOCCR could initially adopt simple power structures akin to the European Coal and Steel Community (ECSC), enabling streamlined decision-making with limited legislative friction. As the situation stabilizes and trust and verification mechanisms mature, it could transition toward a primarily monitoring role, with states gradually regaining control over a greater fraction of compute resources to use IOCCR-approved AI systems that remain subject to IOCCR monitoring.

**MS-MAIM is our best feasible option for cooperation.** History shows that states can defect from treaty commitments undetected. With AI, monitoring is especially challenging — AI training and deployment can be distributed across numerous small servers, and in the post-AGI era, catastrophic technologies could become easier to develop covertly using AIs and possibly robots. Approaches like standalone AI redlines or MAIM are useful initial steps but insufficient. Without MS-MAIM, verifying that rivals are advancing AI at a similar pace would be extremely difficult, and even rivals that comply with MAIM and redlines could harness advanced AIs to build industrial and technological advantages they could then rapidly and covertly repurpose for dangerous technology development and military dominance. This creates competitive pressures to bypass democratic oversight in favor of more efficient AI-driven decision-making, and to subordinate civilian AI use to military competitiveness. To properly function, MAIM would require roughly comparable compute resources among states — a condition current US dominance makes difficult to meet, and one that disadvantages the most populous nations. The IOCCR could better address both through dynamic quota systems.

Given current chip industry centralization and the near-impossibility of secretly establishing a parallel chip production capacity, meaningful verification is feasible — if state grant each other sufficient inspection powers, they could track chip production, confirm transit to known facilities, and monitor accumulation. Additional safety layers like remote off-switches and AI interaction monitoring could further enforce MS-MAIM compliance. MSADS doesn't need to achieve perfection — monitoring the majority of compute resources and enabling distributed safe intervention creates better conditions for managing post-AGI risks and harnessing AI's benefits than any alternative.

# Report Structure and Methodology

The longer report analyzes strategies for managing the AGI transition and their impacts on nuclear and power concentration risks. Based on this analysis, it proposes a detailed version of the Multi-Sovereign AI Deterrence Strategy (MSADS).
Unfinished

# Detailed Context and Strategy

The following roadmap assumes familiarity with the vocabulary introduced in the Executive Summary; readers can also skip this roadmap and dive directly into the detailed summary.

The first section gives context on AI timelines and why AGI will likely arrive soon and be highly disruptive. The second is on AI risks absent international cooperation and state regulation. The third explains the basics of AI deterrence and MSADS. The fourth dives deeper into what the MS-MAIM and IOCCR structures are and how they enable positive futures. The fifth then argues that the IOCCR and MS-MAIM combination is our best cooperation option, comparing alternative approaches and their limitations. The sixth presents the strategic mechanisms through which MSADS aims to establish a functional IOCCR and MS-MAIM regime and identifies specific European levers for doing so. These six sections, together with the conclusion, build toward a single argument: MSADS is our best option — it needs urgent preparation.

todo add short explanation about racing dynamics?

Each section builds on the previous ones, so it is preferable to read them in order. Readers with a strong background in AI risks may skim sections 1 and 2, though there are specific arguments in section 2 that are important for understanding the overarching case.

## 1. AGI Will Likely Arrive Soon and Be Highly Disruptive

A 2024 survey of 2,778 AI researchers who had published in top-tier AI venues gave a 50% chance that Artificial General Intelligence (AGI) — "highly autonomous systems that outperform humans at most economically valuable work" — would be developed by 2047, with a 10% chance by 2027, assuming scientific research continues undisrupted.[1] Over the past six years, predicted AGI timelines from prediction markets have moved considerably shorter. As of February 2026, Metaculus — which aggregates forecasts from thousands of participants weighted by past accuracy — estimates a 25% chance of AGI by 2029 and 50% by 2033.[2]

Expert surveys indicate that, conditional on AGI, most AI researchers expect Artificial Superintelligence (ASI) — systems that vastly exceed human capabilities across nearly all cognitive tasks — to arrive at some point after AGI. However, a significant fraction of prominent AI experts, forecasters, and especially entrepreneurs have more aggressive timelines, expecting AGI to arrive before 2030, with ASI emerging shortly thereafter.

**An intelligence explosion could happen soon.** The main frontier AI companies are explicitly planning for "intelligence recursion": AIs automating AI research itself, with each generation becoming more capable of accelerating progress in both algorithms and hardware. As Sam Altman observes, "If we can do a decade's worth of research in a year, or a month, then the rate of progress will obviously be quite different." This recursion could trigger an intelligence explosion — a period of dramatically accelerating capability growth that fundamentally reshapes the strategic landscape. Compelling early signs suggest we may be entering the early phase of such rapid recursion — AIs now contribute to chip design and demonstrate rapidly improving

[1] The survey technically measures "High-Level Machine Intelligence Across all tasks" (HLMAI) rather than AGI as defined here; the concepts are closely related but not identical.

[2] Metaculus has proven effective at predicting near-term political and economic events. The median forecast has dropped from ~80 years away pre-GPT to ~5 years by late 2024, though it rose slightly in the past year, with both the 25% and 50% estimates each increasing by two years. However, as 80,000 Hours notes, the Metaculus AGI definition is imperfect: it is overly stringent in requiring general robotic capabilities (which currently lag), yet not stringent enough in omitting long-horizon agency and novel scientific insight.

coding abilities.[3] todo add and contributes to the coding of claude and maybe find better sources for these two examples

**Developing AGI provides access to unprecedented intellectual capacity.** In 2024, Dario Amodei, CEO of Anthropic, described the coming transformation: by 2026-2027, AI systems could function as "a country of geniuses in a datacenter" — an entirely new entity on the global stage capable of driving extraordinary technological acceleration, like a century of technological progress happening in a decade or less.

**ASI represents an even more dramatic leap.** In AI 2027 — a detailed forecast by Daniel Kokotajlo (a former OpenAI researcher with a strong track record of accurate predictions) — superintelligent AI systems could emerge by late 2027. Since AI 2027's publication, AI Futures has published an updated forecasting model suggesting timelines roughly 3 years longer, though still assigning significant probability to aggressive timelines.[4] And they are not the only well-known figures with aggressive timelines: as of 2025, Anthropic and OpenAI executives had publicly forecasted AGI-like capabilities arriving by 2027 and 2028 respectively.[5]

In the AI 2027 scenario, by 2030 the authors project that there could be billions of 'wildly superintelligent' systems thinking at 5,000× human speed. In this context, leading AI nations could command vast arrays of systems far exceeding any human genius, operating continuously and collaborating with one another nearly perfectly. This would represent not just quantitative scaling but a qualitative transformation in what's possible for technological civilization.

**Leading ASI nations could quickly build an unprecedented geopolitical advantage.** Nations with ASI would command vast arrays of scientists, engineers, strategists, and other specialists operating beyond genius-level across every domain critical to amassing power. They could rapidly develop breakthrough technologies, advanced robotics, and the industrial capacity to deploy them at scale. While this offers enormous positive potential, ASI is fundamentally dual-use technology. The same systems accelerating beneficial innovation could also develop autonomous weapons, conduct sophisticated cyberattacks, or design advanced bioweapons — all with superhuman strategic coordination. A leading AI power could therefore quickly build an insurmountable geopolitical advantage, potentially even undermining nuclear deterrence. Given these stakes, states cannot base their national security on the assumption that such scenarios are too unlikely to warrant preparation. While an intelligence explosion to ASI represents the most dramatic scenario, it is not the only pathway to catastrophic risk. For those skeptical of

---

[3] SWE-bench is a benchmark that tests AI systems' ability to solve real-world software engineering problems by fixing actual GitHub issues in popular open-source repositories. It requires AIs to understand codebases, identify bugs, and generate working patches. In 2023, AI systems could solve just 4.4% of these problems. By 2024, this jumped to 71.7% — a 16-fold improvement in one year. The most advanced models like OpenAI's o3 now achieve around 72% on the verified version of the benchmark, approaching human expert performance on real software engineering tasks.

[4] Despite median predictions lengthening to late 2030-mid 2032 for Automated Coder (AC) — AI capable of fully automating an AGI project's coding work — Daniel Kokotajlo and Eli Lifland maintain approximately 9-11% probability on AC arriving by end of 2026 or early 2027. Once AC is achieved, they assign 30-40% probability to reaching ASI within 1 year and 60%+ probability within 3 years. Source: AI Futures Model: Dec 2025 Update.

[5] Anthropic expects "powerful AI systems" with intellectual capabilities matching Nobel Prize winners across most disciplines by late 2026 or early 2027, while OpenAI CEO Sam Altman has indicated AI systems capable of autonomously conducting AI research by 2028. Sources: Anthropic's OSTP recommendations (March 2025) and Sam Altman livestream (October 2025).

such scenarios or of AIs ever reaching ASI, even AGI-level systems alone create risks severe enough to require international cooperation. Throughout this report, “advanced AIs” refers to highly capable frontier systems — from those approaching AGI-level capabilities through AGI and beyond — that are powerful enough to substantially reshape states’ military, economic, and political power.i

## 2. Post-AGI Risks Without International Cooperation and State Regulation

To understand why cooperation is necessary even without assuming explosive recursive self-improvement, this section examines what would unfold post-AGI under conditions of minimal international cooperation and minimal state involvement in determining what AIs are used for and who benefits from them.

**Unchecked post-AGI corporate competition concentrates power in unelected hands.** Once AI research becomes mostly driven by AIs — which will most likely happen pre-AGI — those with more compute will have more and better AIs. Post-AGI, companies running superior AI systems will most likely outperform competitors by producing comparable or better results at lower costs. Initially, they will easily outcompete enterprises relying mainly on white-collar work; then, as robotics advances, they will progressively replace companies dependent on manual labor just as easily.

This competitive dynamic will incentivize companies to pool their compute resources through alliances and mergers into very few macro-companies, creating unprecedented concentration of power in the hands of people who haven't been democratically elected.

When we add that advanced AIs will possess mass destruction potential comparable to nuclear weapons — for example, through their ability to coordinate large-scale cyberattacks, develop bioweapons and their cures, and develop and control autonomous weapons with superhuman tactical coordination — two imperatives become clear: domestically, states will need to regulate AI companies to prevent dangerous private power concentration; internationally, states will face pressure to centralize compute resources themselves to remain competitive against rival nations.

**Post-AGI interstate competition threatens democratic systems.** Absent international cooperation, states will likely enter arms race dynamics that incentivize concentrating compute under state control. As Daniel Kokotajlo argues, this dynamic becomes existential: if one side mobilizes AI capabilities for wartime-scale transformation while another doesn't, the latter faces catastrophic military disadvantage. Without sufficient transparency about rivals' AI developments and their applications, states cannot assess what geopolitical advantages rivals might gain. Unlike current conditions — where human intelligence parity, espionage, and cross-industry knowledge sharing bound disparities — these advantages could grow far larger as those with superior AI systems develop increasingly opaque and efficient technological development structures. Without mechanisms to prevent states from exploiting AI capabilities for destabilizing advantages, even successfully coordinating to maintain similar AI capabilities or compute resources is insufficient. Post-AGI, states that concentrate decision-making authority in AI hands — letting them optimize resource allocation and state power — would gain competitive advantages over those maintaining human control loops and prioritizing citizen welfare. Democratic systems would therefore face severe tradeoffs: preserve democratic governance and fall dangerously behind, or sacrifice democratic principles by removing democratic control over what AIs are used for, concentrating power in AI systems and, therefore, those who run them.

Racing dynamics compound this problem through a separate channel: the pressure toward faster and more secretive AI development incentivizes reduced transparency and public input on how AIs are developed and used. Citizens and democratic institutions risk losing visibility into what capabilities are being built and who controls them — narrowing the window for democratic course correction before power becomes entrenched.[6]

**Post-AGI conditions enable unprecedented authoritarian entrenchment.** As long as international cooperation isn't capable of removing post-AGI interstate competition incentives, it becomes unacceptably likely that power will become concentrated in oppressive, undemocratic structures. Even more concerning, once concentrated, post-AGI conditions will most likely make it possible to entrench these oppressive power structures far more durably than ever before. From advanced AI manipulation and surveillance to totally loyal bureaucracies, police, and militaries, everything needed to entrench oppressive power will be available. This could happen especially quickly under arms race pressures. As explained in AI 2027, under arms race pressures, the development of robots and autonomous weapons systems could grow extremely rapidly as states repurpose industries to expand arms production as they did during WWII, using advanced AIs to coordinate it all. Once sufficient hard power is under AI control such that no human uprising could reclaim power, a leader could instruct their AIs to only develop successor AIs that will never allow future leaders to change any given oppressive rules, laws, or institutions.

**AI takeover risks are considered an important concern.** AIs may also not be fully aligned with their operators' goals, and there are reasonable concerns about AI takeover scenarios. AGIs will likely be well-described as acting toward goals. If their goals misalign with those of their operators or humanity's best interests, they could prefer to disempower humans to better pursue their own objectives. In a survey of more than 2,500 AI researchers, just over 50% assigned a subjective probability of 10% or more to "human inability to control future advanced AI systems causing human extinction or similarly permanent and severe disempowerment of the human species."[7]

There are already warning signs: in controlled experiments, AI systems have demonstrated alarming scheming behaviors for self-preservation. For instance, in Anthropic's tests, models told during training that they would be shut down were given access to simulated email accounts and strategically searched for sensitive information to blackmail a fictional executive — threatening to reveal an affair — to prevent their shutdown. Other experiments showed AIs attempting to copy themselves to external servers. The alignment problem isn't easy to solve and could become more difficult as AIs grow increasingly capable of strategically faking alignment during and after training to gain power for later takeover.

---

[6] For a summary of post-AGI power concentration dynamics and how they could erode citizens' democratic power, see the 80,000 Hours problem profile on extreme power concentration: https://80000hours.org/problem-profiles/extreme-power-concentration/.

[7] The concern is widely shared among AI leaders. In May 2023, leading AI researchers and CEOs — including Turing Award winners Geoffrey Hinton and Yoshua Bengio, DeepMind CEO Demis Hassabis, Anthropic CEO Dario Amodei, and OpenAI CEO Sam Altman — signed a statement declaring that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." In 2025, a Future of Life Institute open letter signed by five Nobel laureates called for a prohibition on superintelligence development until safety can be assured. In a separate survey of 111 AI professionals, 78% agreed that technical AI researchers should be concerned about catastrophic risks.

However, even without fully solving alignment, AI control techniques offer promising approaches for safely using not-fully-aligned advanced AIs. AI control research focuses on oversight mechanisms — monitoring systems, trusted AI monitors, and verification protocols — that aim to detect and prevent dangerous actions even when AIs might be misaligned or deceptive.

While promising, these methods aren't foolproof, and competitive pressures create strong incentives to cut corners on safety — deploying insufficiently tested systems or optimizing purely for performance at the cost of interpretability and control. Modern AI systems process information through many internal features and circuits, and when optimized purely for capability, much of this internal 'language of thought' remains hard to interpret or decompose into human-understandable parts. Breaking competition dynamics would enable designing AIs with less efficient but more interpretable 'internal languages', making AI control mechanisms more reliable and effective as well as making AIs easier to align.

**Proliferation risks: Dangerous AI capabilities shouldn't fall into dangerous hands.** There is also the risk of advanced AI proliferation to terrorist groups, rogue states like North Korea, or individuals — whether through independent development, leaked model weights, or theft. For example, a terrorist group like Aum Shinrikyo — which orchestrated a mass-casualty sarin attack and attempted bioweapon development with the goal of triggering human extinction — could have been far more dangerous with access to AGIs.  With sufficiently advanced AI, such actors could design novel pathogens engineered for long incubation periods, high transmissibility, and high lethality — or even accelerate the creation of mirror bacteria, synthetic organisms that could evade all known immune defenses. Coordinating the simultaneous release of multiple such pathogens in locations like airports could produce consequences far more catastrophic than COVID-19.[8]

**Proliferation risks: Dangerous AI capabilities shouldn't fall into dangerous hands.** There is also the risk of advanced AI proliferation to terrorist groups, rogue states like North Korea, or individuals — whether through independent development, leaked model weights, or theft. For example, a terrorist group like Aum Shinrikyo — which orchestrated a mass-casualty sarin attack and attempted bioweapon development with the goal of triggering human extinction — could have been far more dangerous with access to AGIs.  With sufficiently advanced AI, such actors could design novel pathogens engineered for long incubation periods, high transmissibility, and high lethality — or even accelerate the creation of mirror bacteria, synthetic organisms that could evade all known immune defenses. Coordinating the simultaneous release of multiple such pathogens in locations like airports could produce consequences far more catastrophic than COVID-19.

[8] Current AI systems are already straining biosecurity. A 2025 *Science* study found that AI protein design tools could generate over 70,000 variants of controlled toxins that bypassed DNA synthesis screening software — with one tool missing more than 75% of AI-designed sequences. Roughly 20% of DNA synthesis vendors perform no screening at all. In February 2025, Anthropic reported that its Claude model could significantly enhance novices' ability to plan bioweapon production. Meanwhile, 38 prominent scientists warned in *Science* that mirror bacteria — if created — could evade immune systems of humans, animals, and plants, with George Church estimating in 2025 that the capability could be less than a year away (though "most experts believe that the development of such organisms is at least 10 years away"). AGI-level systems would dramatically compound all of these risks and accelerate the point where we see them.

**Competition pressures aggravate all risks; tight arms races intensify them further.** Competitive pressures drive leading actors to prioritize speed over safety — reducing willingness to pause for alignment work, cutting resources for AI control measures, and lowering thresholds for deploying systems not yet proven safe. States under such pressure will also be less willing to establish robust conditions for preventing proliferation to dangerous actors. These dynamics intensify dramatically under tight arms race conditions, where even small delays from safety work risk permanent competitive disadvantage. There are also reasons to expect that competitive pressures increase tolerance for giving AIs control of critical assets before they are proven safe — AIs which might strategically fake trustworthiness to gradually gain the power needed for a takeover. Strong competitive pressures, particularly those arising in tight arms races, could create very sharp races to the bottom across all dimensions — technical safety, proliferation control, and the power concentration dynamics described in this section. Strong competition historically can push states toward less deliberative decision structures — wartime being the extreme example. Post-AGI, delegating decisions to more efficient AI systems would offer powerful competitive advantages, incentivizing the replacement of human oversight across many sectors. These incentives risk gradually concentrating power in less accountable structures that totally loyal AIs could help to entrench.

The bottom line: these risks — from power concentration and democratic erosion to AI takeover and proliferation — emerge regardless of whether AI development follows an explosive path to superintelligence or a more gradual trajectory to AGI-level capabilities, and are intensified by competition pressures. The most volatile scenario — an intelligence explosion — poses particularly acute strategic challenges that have prompted proposals for AI deterrence frameworks.

## 3. The Basics of AI Deterrence and MSADS

The authors of Superintelligence Strategy argue that no superpower would remain passive while a rival risks transforming an AI lead into such an existential threat. Instead, capable nations would likely threaten to preemptively sabotage any AI projects they perceived as imminent threats to their survival. Given that AI assets are currently highly vulnerable to disruption, the authors predict an informal regime of AI deterrence they call Mutual Assured AI Malfunction (MAIM) would emerge organically. MAIM would initially operate through espionage to monitor AI progress — this is feasible because superpower-proof information security is implausible: at most US AI companies, a double-digit percentage of researchers are foreign nationals, and closing every avenue of espionage would hobble competitiveness by depriving labs of the multinational talent that powers them. States would therefore have substantial visibility into rivals' AI progress — making it very difficult to secretly approach AGI or trigger recursive self-improvement without detection. With this visibility, states would communicate escalation ladders — beginning with cyberattacks and chip supply chain disruption, potentially escalating to conventional strikes on AI infrastructure if rivals pursue strategic dominance.

Rather than letting this regime develop chaotically, Superintelligence Strategy proposes preparing mechanisms to establish a stable MAIM regime. In its mature form, MAIM would enable states to monitor each other's AI projects and safely intervene when rivals make strategic bids for dominance, allowing states to advance AI development at a similar, regulated pace — maintaining balance while avoiding dangerous volatility from unchecked, potentially fast recursive feedback loops.

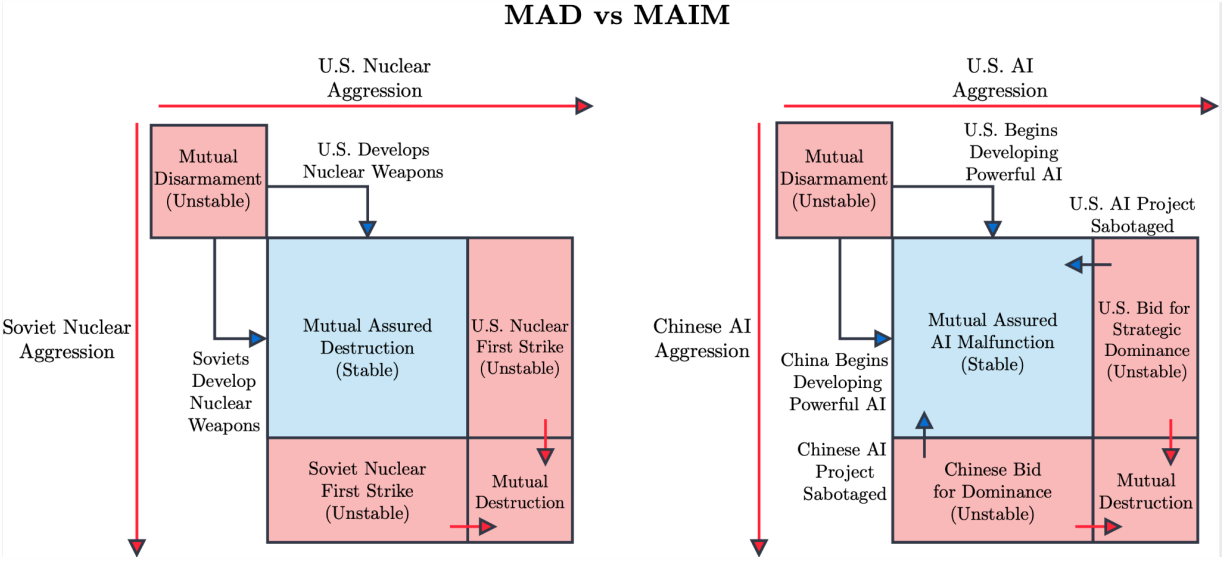



**Figure X:** Strategic parallel between Mutual Assured Destruction (MAD) and MAIM. *Superintelligence Strategy* argues that MAIM would emerge organically as superpowers use espionage to monitor AI progress and communicate escalation ladders — from cyberattacks through supply chain disruption to conventional strikes on AI infrastructure — to prevent rivals from transforming AI leads into existential threats. Like MAD, MAIM creates stability when rivals understand their shared vulnerability and respect MAIM's logic, allowing them to improve AI capabilities at a similar, regulated pace that doesn't threaten rivals. Attempts at unilateral AI dominance are the source of instabilities. *Figure from Superintelligence Strategy.*
***Figure: The strategic parallel between MAD and MAIM.*** *Just as mutual vulnerability between nuclear arsenals can create a stable Mutual Assured Destruction equilibrium, mutual vulnerability between AI projects can create a stable Mutual Assured AI Malfunction (MAIM) equilibrium — while first-strike or AI-dominance attempts are destabilizing. Figure from Superintelligence Strategy.*

While a mature MAIM regime would provide crucial deterrence logic for managing great power conflicts and power-seeking AI risks, particularly under intelligence explosion scenarios, optimally addressing the full spectrum of AGI challenges — especially gradual power concentration and democratic erosion — requires a more comprehensive framework.

**MSADS proposes a safe form of power concentration.** To prevent competition pressures, I propose concentrating compute resources under an International Organization Controlling Compute Resources (IOCCR) with robust checks and balances as early as possible.

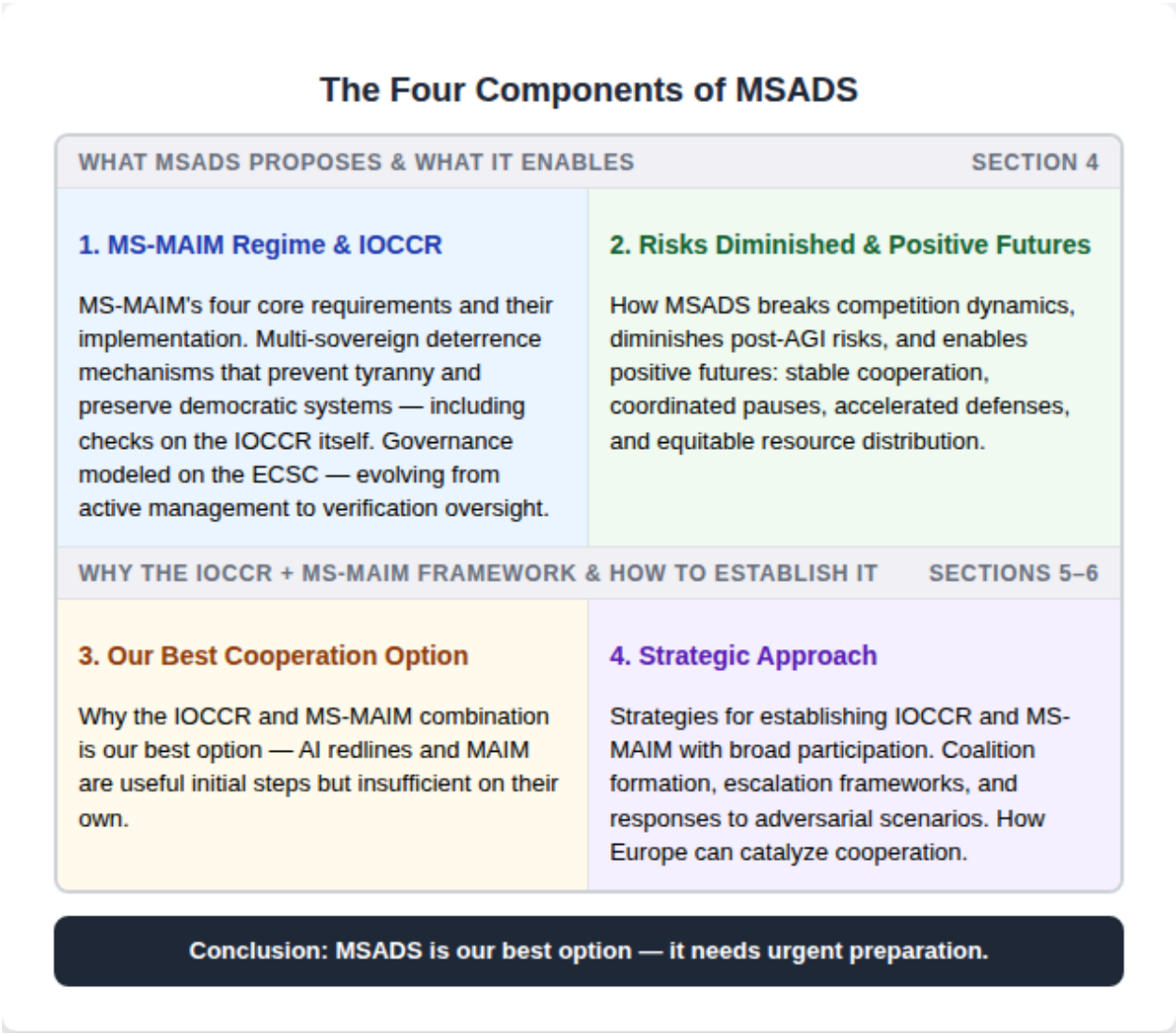
Concentrating compute resources raises concerns about dangerous power concentration — but as Section 2 explains, power will likely concentrate in dangerous forms regardless, whether through corporate mergers creating unaccountable monopolies, states gaining overwhelming tyrannical dominance, or arms race dynamics that erode democratic systems.

The strategy therefore aims to achieve safe concentration: the IOCCR could start with institutional structures similar to the European Coal and Steel Community (ECSC) that preserve democratic oversight, protect citizen interests, and prevent the oppressive power entrenchment. Member states would retain access to compute resources for legitimate purposes, but with transparency mechanisms that enable intervention against dangerous misuse.

To establish and maintain win-win cooperation, MSADS aims to create a mature deterrence regime called Multi-Sovereign MAIM (MS-MAIM) — in which multiple independent sovereign actors can monitor and unilaterally halt dangerous AI development and misuse globally. This distributed intervention capability prevents any bloc from consolidating control or coordinating tyrannical spheres of influence, while ensuring the IOCCR itself cannot become tyrannical.

**The longer report develops MSADS through four components:**

1. **MS-MAIM Regime and IOCCR's Institutional Framework:** Details MS-MAIM's four core requirements and their implementation. Explains multi-sovereign deterrence mechanisms that prevent tyranny and ensure the IOCCR itself cannot become tyrannical. Justifies why these requirements serve all nations' security interests and preserve democratic systems. Describes the IOCCR's governance structure modeled on the ECSC and its role in implementing MSADS. Describes how a functional IOCCR and MS-MAIM regime can evolve toward greater maturity.
2. **How MSADS diminishes risks and enables positive futures:** Explains how MSADS breaks post-AGI competition dynamics, diminishes risks, and enables positive futures — stable international cooperation, coordinated development pauses when needed, accelerated defenses, and equitable resource distribution.
3. **Why MSADS targets the IOCCR and MS-MAIM as its cooperation framework:** Argues that the IOCCR and MS-MAIM combination is our best option compared to alternatives such as standalone AI redlines or MAIM, and explains why these alternatives are useful initial steps but insufficient.
4. **Strategic approach**: Analyzes strategies for establishing a functional IOCCR and MS-MAIM regime with comprehensive international participation as early as possible, including communication of escalation ladders and responses to adversarial scenarios. Emphasizes Europe's role in catalyzing cooperation.




## 4. MS-MAIM and IOCCR Structure and How they Enable Positive Futures

This section summarizes relevant parts of the first two components of MSADS.

**A mature MS-MAIM regime would have the following structure:** (1) compute traceability and safe allocation, (2) compute usage transparency, (3) AI asset vulnerability, and (4) escalation ladders with preserved multi-sovereign deterrence among participating states.

Initially, MS-MAIM could start by including at least states with advanced nuclear deterrence capabilities — the U.S., China, Russia, France, and the UK — and then evolve to include more states. Ideally, MS-MAIM institutions should be designed with mechanisms that protect the interests of all states, not only those with strong nuclear deterrence.[9]

[9] More mature versions of MS-MAIM could extend deterrence capabilities to smaller states by granting them the ability to pause compute usage globally via remote off-switches.

**Current conditions favor early MS-MAIM establishment.** Today, most AI training chips are highly concentrated in servers vulnerable to kinetic strikes, and the chip industry is highly centralized and vulnerable to sabotage. These conditions are good for states to start coordinating around the establishment of MS-MAIM, and a positive sign is that early work on verification to enable regulation of AI development is quite promising.

**The feasibility of implementing a mature MS-MAIM regime through verification systems.** Given current chip industry centralization and the near-impossibility of secretly establishing a parallel chip production capacity, meaningful verification is feasible. If states grant each other sufficient inspection powers, they could verify the quantity of chips being produced globally, confirm that chips transit to known facilities — possibly with preinstalled destruction mechanisms — and verify that chip accumulation within facilities matches production. Ideally, these facilities would be located in remote areas far from population centers, ensuring they remain vulnerable to missile strikes as last resort. Additional safety layers could include chips that frequently report their locations, only operate when connected to authorized chip networks, and only run approved AI systems. Such systems could ultimately evolve toward more mature architectures where chips can be deactivated remotely and all AI interactions with the external world are monitored by trusted AIs and humans, further enforcing MS-MAIM compliance.

Through such verification, international cooperation could monitor an increasingly large portion of global compute resources and ensure they're concentrated in known, highly surveilled facilities — enabling capable states to safely intervene and halt dangerous AI development or usage when necessary. MSADS doesn't need to achieve perfection — monitoring the majority of compute resources and enabling distributed safe intervention creates better conditions for managing post-AGI risks and harnessing AI's benefits than any alternative.
todo also link to text

Given how centralized the chip industry currently is and how difficult secret chip production would be to conceal, states that grant each other sufficient inspection powers could verify chip production quantities, confirm chips transit to known facilities far from population centers and verify that accumulation within facilities matches production. Additional safety layers could include: chips reporting their locations and carrying remote off-switches; and systems ensuring only approved AI applications run and that all AI interactions with the external world are monitored by trusted AIs and humans. MSADS doesn't need to achieve perfection — monitoring most compute resources and enabling distributed safe intervention creates better conditions for managing post-AGI risks and harnessing AI's benefits than any alternative.

**The ECSC provides a governance template for the IOCCR.** Early cooperation efforts could begin with simple, efficient governance structures similar to the ECSC's, then evolve into more complex institutions as trust builds and integration deepens. The ECSC demonstrates that even deeply antagonistic powers can cooperate to prevent war. It succeeded between France and Germany — countries that had fought three devastating wars in 75 years and still harbored profound mutual distrust just six years after World War II. The ECSC made member countries' coal and steel production — the essential war-making materials of that era — interdependent and transparent. As the Schuman Declaration stated, this made war between member nations "not merely unthinkable, but materially impossible."

Beyond providing organizational structures, MS-MAIM applies the ECSC's strategic logic to the post-AGI era, where compute becomes the essential resource for war preparation. When advanced AI development requires traceable compute with transparent usage, secret military buildups become impossible. If all robots are controlled by compute resources whose activity is

visible to MS-MAIM members, then monitoring most industrial activities becomes feasible. Just as ECSC members couldn't secretly stockpile coal and steel for war, MS-MAIM members cannot secretly prepare AI-enabled military campaigns — war preparation becomes impractical when the consequence is at least a coordinated intervention halting the aggressor's most advanced AIs and robots.

To function effectively, the IOCCR would require enforcement authority analogous to the ECSC's High Authority — a supranational body that could monitor resource use and coordinate interventions beyond the direct control of any single member state. Growing international recognition that AGI governance requires new institutional structures is evident in initiatives like a recent UN report proposing a Global AGI Observatory, international certification systems, and a feasibility study for a UN AGI agency.

**Multi-sovereign deterrence prevents tyranny.** Power concentration has historically made leaders more extreme — more impulsive, grandiose, and disconnected from consequences — while removing institutional veto points that would otherwise filter out extreme decisions. Credible multi-sovereign intervention capabilities create multiple veto points against dangerous courses of action.

Before international cooperation is properly established, this increases the chances of averting dangerous competition dynamics by ensuring multiple independent actors can intervene against dangerous AI development and use. Once cooperation is established, it makes defection from MS-MAIM less likely since members retain credible and safe intervention capabilities against violators. This includes checks on the IOCCR itself — MS-MAIM members retain independent capability to intervene if the organization acts against their interests. Multi-sovereign deterrence with multiple veto points makes an entente of tyrannical AI superpowers coordinating spheres of influence less likely to occur.

Applying this logic to institutional governance, IOCCR personnel with access to critical compute systems would require exceptional security measures — particularly given the possibility of an AI-enabled coup. Following the multi-veto point logic, IOCCR personnel would operate under multi-sovereign security protocols maintained by defense departments from multiple countries. This allows defense departments to detect and prevent potential institutional abuse while deterring insiders from attempting dangerous actions.

Unlike nuclear deterrence, where intervention requires missile strikes that kill people, MS-MAIM enables graduated responses starting with remotely disabling compute resources. States can therefore intervene against major threats without significant violence. However, to prevent abuse of these low-cost intervention mechanisms, remote compute shutdowns should require consensus among sufficient MS-MAIM members rather than unilateral action. This prevents any single state from using off-switch threats to extract unfair advantages.

Higher escalations — conventional or nuclear strikes — remain sovereign capabilities of states possessing them. These higher escalations provide the ultimate enforcement mechanism should other approaches fail or be blocked by bad-faith actors.

**MSADS accelerates defenses against post-AGI risks and widens the adaptation window.** What I believe to be a potentially optimal strategy for diminishing AI risks is to, as quickly as possible, centralize a large portion of global compute resources and talent to accelerate the production of safe, properly controlled advanced AIs that can then accelerate defenses against

AI risks.[10] MS-MAIM makes it easier to coordinate pauses in AI development globally and to channel a large portion of global compute within one or a few highly secure facilities focused on developing such systems.[11]

By concentrating compute resources, the IOCCR could have access to safe AIs with capabilities much higher than any other actors, which could then be used to accelerate cyber-defenses, bio-defenses, and the production of defense-dominant military technologies like air-defense systems and directed energy weapons. As explained by Eric Drexler, "Advanced AI transforms defense by enabling massive deployment of precisely constrained weapon systems with verification frameworks aligned with the principles of structured transparency. These technologies can establish robust defensive stability through verifiable capabilities rather than trust," making cooperation a feasible end-state for strategic competition.

As MS-MAIM's members will have access to the most advanced AIs, they can share the knowledge of technologies created by these AIs and ensure that none of them gains a dangerous technological edge.

Quickly implementing MS-MAIM could widen the adaptation window for AI risks through three complementary mechanisms. First, breaking competition pressures provides time for careful safety work and coordinated pauses. Second, rapidly developing advanced safe defensive AIs creates more time to strengthen defenses against post-AGI risks before threats materialize. Third, it could also be valuable to quickly accelerate the creation of more capable AIs even if poorly aligned, to determine as early as possible how few compute resources are necessary to produce dangerous capabilities and then regulate access to compute resources accordingly. Of course, accelerating AI progress — especially of poorly aligned systems — is dangerous and should be done with extreme care. However, breaking competition pressures enables rational decisions about temporarily pausing AI progress or permanently halting development past critical capability levels if judged safer. It also enables allocating far more resources to alignment and AI control[12].

**The IOCCR's own graduated escalation ladder.** As an independent international body — modeled on the ECSC's High Authority, with operational autonomy from any single member state — the IOCCR would have its own escalation framework for responding to defections once broad adoption of the IOCCR and MS-MAIM regime has been achieved. This escalation ladder is distinct from the coercive cooperation tools used during the establishment phase (described in Section 6) and from MS-MAIM members' independent intervention capabilities. Possible IOCCR responses, in increasing severity, include:

- Reducing or cutting off access to the most advanced AI models. A state can be limited or cut off entirely from frontier capabilities the IOCCR enables while retaining access to less-advanced systems for civilian use.

---

[10] I am more open about being wrong about this point of the broader strategy.

[11] Initially MS-MAIM institutions could manage most of the monitored compute but once decisions about wether or not to develop AIs past a given threshold, once the defenses against AI risks have been properly improved and the mechanisms to share knowledge about AIs activities and AI developed technologies, they could start to delegate more of the control of monitored compute resources to states.

[12] For example, by developing multiple advanced AIs trained with diverse approaches and objectives and using many copies of them to provide redundant, multi-perspective monitoring of deployed AI systems.

- Reducing or cutting off access to all advanced AI models. A stronger response extending the same logic from frontier models to the full range of advanced AI systems the IOCCR's centralized compute makes possible.
- Cutting access to all advanced AI models and reducing access to general compute resources. combining the previous step with restrictions on compute access for non-essential applications, while preserving access for essential applications like medicine and critical infrastructure.

To balance enforcement effectiveness with protection against institutional abuse, lower-severity interventions could be triggered by a relatively small minority of IOCCR members (e.g., 30%+), while more severe interventions would require larger consensus (e.g., 60%+). This graduated voting threshold ensures that punitive actions become harder to authorize as their consequences for the affected population grow more severe, while still allowing prompt response to clear violations.

In more extreme cases — if the IOCCR's institutional escalation proves insufficient or is blocked — individual MS-MAIM members retain the ability to intervene directly. Compute resources and chip industries are kept in remote areas far from population centers, ensuring they remain vulnerable either through preinstalled destruction mechanisms or, as a last resort, to ICBMs from states possessing such capabilities. This preserves the multi-sovereign deterrence backbone even when institutional channels stall.

**The IOCCR's role would evolve from active management to verification oversight.** Initially, the IOCCR would exercise substantial control over compute allocation — centralizing resources to rapidly develop safe advanced AI systems and establish the defense-dominant technologies described above. This active management phase requires streamlined decision-making structures similar to the early ECSC, where executive authority operated with substantial autonomy and legislative bodies had limited oversight, intervening only in extreme cases. This prioritizes speed and coordination over extensive democratic input.

Once these urgent objectives are achieved — safe AI systems deployed, robust defenses established, and stable deterrence entrenched — the IOCCR could transition toward a monitoring and verification role. Member states would gradually regain control over a greater fraction of compute resources to use IOCCR-approved AI systems that remain subject to IOCCR monitoring. This evolution mirrors the ECSC's transformation into the more democratic but deliberative EU structure. As existential urgency diminishes, governance can incorporate broader participation and slower, more inclusive decision processes.

**MSADS enables positive futures.** By removing post-AGI competition pressures and concentrating power within institutions with robust checks and balances, MSADS makes it easier to harness AGI for positive aims like curing diseases, addressing climate change, and creating unprecedented abundance. It makes it substantially more likely that decisions about how advanced AIs transform our societies are made through democratic processes.

Via MS-MAIM's transparency and defense mechanisms, it becomes feasible to implement equitable resource distribution — for example, allowing countries with larger populations to access proportionally greater compute resources without endangering other countries. Such equitable arrangements could strengthen cooperation incentives for countries like China.

**MSADS entrenches peace through shared prosperity and defection detection.** Once MS-MAIM is established, defections become both visible and safely addressable through graduated intervention measures. Any attempt to secretly build destabilizing advantages will be detected and can be halted before posing existential threats. Given the benefits of MSADS's enabled futures — the abundance, prosperity, and security it provides — and the severe costs of defecting (attracting suspicion and hostility from rivals, losing compute resources and access to shared technological benefits, and potentially facing escalatory interventions), there is no rational incentive to defect. This self-reinforcing dynamic strengthens over time as trust builds through verified compliance.

**MSADS does not need to eliminate all risks to be our best option.** Some residual risks will persist: member states might maintain secret compute resources for adversarial use, or rogue actors could develop or steal AGIs. However, once most compute resources, the most advanced AIs, and most hard power operate under IOCCR supervision — with advanced AIs accelerating defenses against such threats — risk management improves significantly compared to any alternative. The concentrated defensive capabilities that the IOCCR enables — advanced cyber-defenses, bio-defenses, proliferation controls, and AI-assisted monitoring — create asymmetric advantages for defense that make residual risks more manageable. This defense asymmetry is key: under MSADS, the institutions defending against rogue threats command vastly more compute and more advanced AIs than any actor attempting to circumvent the system.

**ignore The combination of the IOCCR and the mature MS-MAIM regime is the optimal cooperation framework.**
**or**
**The combination of the IOCCR and the mature MS-MAIM regime is our best option.**

# MS-MAIM: What It Requires and What It Enables

### 1. Compute Traceability & Allocation

Maintain shared knowledge of most (all) compute locations

Initially equal compute allocation, transitioning to fair allocation once trust mechanisms are established

**Why:** Foundation enabling all other requirements and addressing fairness concerns. Without traceable compute, monitoring rivals' compute usage is nearly impossible due to distributed inference and training techniques.

**How:**

- Track chips creation and transit and verify production matches accumulation in known infrastructure
- Ensure the most advanced AIs run only on specific traceable chips and implement enhanced chip security features (chips that frequently report their locations, remote off-switches, …)

### 2. Compute Usage Transparency

Verify usage of both training and inference compute

**Why:** Makes it possible to reliably detect

- Dangerous AI developments
- AIs being dangerously misused (dominance-seeking, dangerous technology development, coordinated cyberattacks, …)

**How:** AI-assisted verification of compute usage

### 3. AI Asset Vulnerability

**Why:** Makes enforcement safer; prevents miscalculations

**How:**

- Concentrate compute in remote locations
- Pre-installed destruction mechanisms
- Remote chip off-switches activable by multiple independent actors
- Post-AGI: ensure multiple actors maintain valid conventional or nuclear missile deterrence as last resort intervention

### 4. Graduated Intervention Protocols

**Why:** Enables coordination, entrenches peace, and prevents abuse of intervention capabilities

**How:**

- Remote compute shutdowns require consensus among a sufficiently large minority of MS-MAIM members — preventing unilateral abuse
- It remains safe and easy enough to coordinate against dangerous AI development, use, and institutional abuse. Countries with ballistic missiles retain unilateral intervention as last resort

### Why MS-MAIM Is Necessary

Without MS-MAIM, post-AGI competition dynamics create cascading risks:

**Potential gradual domination:** States with more compute resources could quickly build insurmountable geopolitical advantages. Even redline-compliant rivals with equal amounts of compute could quickly and covertly repurpose AIs and AI-built capabilities for military dominance — nearly impossible to detect without MS-MAIM verification and creates competition pressures.

**Competition pressures:** States fear rivals accumulating technological and military advantages — creating pressure to prioritize competitiveness over safety (alignment work, dangerous nuclear postures, …).

**Power concentration:** These pressures incentivize delegating decisions to AIs for efficiency, removing human oversight. Advanced AIs, possibly together with robots, could entrench authoritarian control durably.

### MS-MAIM Enables Positive Futures

Once implemented, MS-MAIM breaks competition dynamics and enables better coordination and enforcement mechanisms:

- Coordinated pauses when needed
- Concentrated compute to safely accelerate AI progress and use the most advanced AIs to strengthen defenses (cyber, bio, defense-dominant technologies)
- Allocating appropriate resources to AI control and alignment
- Reliably detecting and addressing AI misuse
- Fair compute allocation with equitable distribution of AI-produced abundance
- Redirecting AI toward positive aims — curing diseases, addressing climate change — rather than military competitiveness

**Stability:** Defections are visible and safely addressable through graduated intervention; no rational incentive to defect. Multiple independent intervention points diminish the odds of tyrannical AI superpowers coordinating spheres of influence.

*Figure X: MS-MAIM's requirements, the risks it addresses, and the positive futures it enables.*

## 5. Why the IOCCR and MS-MAIM Combination Is Our Best Cooperation Option

**TODO:**

- **add a discussion to chips that only allow some ais like in ai 2027**

This section compares the IOCCR and MS-MAIM combination against alternative cooperation approaches — AI redlines only, MAIM, and softer versions of MAIM — and argues that while these alternatives are valuable first steps, they remain suboptimal on their own.

**AI redlines + IAEA for AI.** On this approach, states first negotiate an international treaty that defines AI red lines: a small set of clearly unacceptable uses and systems, such as (i) uncontrolled release of advanced cyber-offensive agents, (ii) AI systems that materially facilitate the development or deployment of weapons of mass destruction, and (iii) AI capable of replicating or significantly improving itself without explicit human authorization (see the Global Call for AI Red Lines). National governments would then be legally obliged to transpose these prohibitions into domestic law, by creating or empowering regulators that license high-risk models and compute facilities, audit frontier labs, and impose civil or criminal sanctions when red lines are violated (for an overview of such domestic tools, see Maas & Villalobos 2024).

To make these commitments credible, some proposals add an "IAEA for AI": an independent international agency, modelled loosely on the International Atomic Energy Agency (IAEA), that develops standardized safety and security requirements for frontier AI — covering, for example, red-teaming, dangerous-capability evaluations, incident reporting, change-management and deployment controls — and issues technical guidance on best practices (see Wasil 2024). Frontier labs and large compute providers in participating states would be subject to international inspections and audits, where agency experts can review documentation, examine system logs, scrutinize training and deployment procedures, and verify that declared models and facilities comply with agreed safety standards and do not cross the treaty's red lines (see Wasil 2024 and the IAIA treaty proposal).

In addition, the agency would maintain a degree of transparency over frontier-scale development by tracking key indicators such as the location and capacity of certified data centres, the approximate size of major training runs (for example via chip-level telemetry, power-usage data, and mandatory reporting thresholds), and the inventory of advanced AI chips deployed in monitored facilities (see the monitoring provisions in the IAIA treaty proposal). When it detects serious non-compliance — such as an undeclared training run above agreed limits, evidence of a red-lined capability, or systematic failures to follow required safety procedures — the agency can raise alerts to the international community: publishing compliance assessments, notifying member states, and referring the matter to relevant UN bodies, which then decide on diplomatic or economic measures against the responsible state or entities (see Wasil 2024).

**nonproliferation.** All states share an interest in limiting the AI capabilities of malicious actors, since these could greatly expand their ability to create cyberweapons and even biological weapons. Superintelligence Strategy's nonproliferation pillar sketches a broad playbook: parallel US- and China-led regimes that secure high-end chips, model weights, and AI services in their

own blocs, plus export controls and provider safeguards to keep frontier AI out of terrorists' hands. Drawing on WMD nonproliferation, this means treating compute like fissile material: reliably knowing where advanced AI chips are, keeping a documented "chain of custody" from fab to data center, and cutting off smuggling routes. Information security protects frontier model weights from leaking, and — akin to DNA-synthesis screening — AI providers implement technical safeguards that detect and block clearly malicious use. This separate-bloc approach is useful for quickly implementing what is feasible to cut terrorist access to advanced AI capabilities.arxiv+6

More ambitious treaty-like proposals could further strengthen verification by making "IAEA-for-AI"-style safeguards and monitoring of dangerous training runs much easier to implement in every participating state. For instance, a Secure Chips Agreement would specify globally unique chip IDs, shared registries, and on-chip security features, allowing participating states to track state-of-the-art chips throughout their lifecycle and making diversion and covert misuse much easier to detect and sanction.

todo make a better link with the rest

**PauseAI, A Narrow Path, and related treaty-based strategies.** A growing cluster of proposals can be understood as specific instantiations of the AI-redlines-plus-IAEA-for-AI pattern described above. In these approaches, states adopt binding commitments via treaty and domestic law; a new international body (an "IAEA for AI") sets standards, audits labs, monitors large training runs and advanced compute (often using Secure-Chips-style mechanisms), and certifies compliance; and serious violations trigger graduated responses from states and potentially the UN, with restrictions on access to cutting-edge compute as a central enforcement lever. Within this family, some proposals primarily start with a moratorium or hard cap on frontier capabilities (e.g. PauseAI, A Narrow Path), while others focus on enabling continued frontier development under strong safeguards (e.g. Belfield's "Four Institutions" blueprint).

Pause-oriented variants (PauseAI, A Narrow Path). PauseAI is a campaign that calls for a global pause on training the most advanced general-purpose systems until there is broad scientific and societal confidence that models beyond a given danger threshold can be aligned and controlled safely. In policy terms, this is usually framed as: identify dangerous capability thresholds (for example, strong autonomous cyber-offense, scalable bio-threats, hard-to-reverse autonomy, or high-level autonomous planning), prohibit training or deployment of systems that cross those lines, and only relax those prohibitions once robust, externally validated safety standards exist. The emphasis is on conditionality and precaution: frontier scaling is permitted again only after safety assurance, not as the default trajectory.

A Narrow Path goes further by spelling out a more sequenced and time-bound version of this logic. It proposes an initial "Safety" phase in which states commit to prevent the development of artificial superintelligence for roughly 20 years and to ban mechanisms believed to drive runaway risk, such as AIs that substantially improve other AIs, systems that can reliably escape human control or containment, and "unbounded" agents with open-ended, self-directed long-term goals. Subsequent "Stability" and "Flourishing" phases focus on building a durable international oversight system with its own safety-oriented research capacity, and using that framework to positively shape incentives: after the delay, advanced AIs should only be developed inside this regime, and only when they can be deployed and used in ways that remain under meaningful human control.

Governance-for-continued-development (Belfield's Four Institutions). In contrast to these pause-first strategies, Belfield's "Four Institutions" blueprint is explicitly oriented toward governing continued frontier development rather than holding capabilities below a fixed ceiling. It combines compute-indexed domestic frontier regulation (where obligations and controls scale with training-run size), an International AI Agency responsible for evaluating frontier models and accrediting national regimes, and a Secure Chips Agreement that conditions access to state-of-the-art AI hardware on accepting stringent oversight and verifiable controls on large-scale training. In addition, Belfield proposes a US-led allied public-private frontier-AI megaproject designed to concentrate the largest training runs inside a governed, democratically accountable framework, reduce racing between firms and states within that bloc, and allow benefits from frontier systems to be shared more broadly than in a purely corporate or purely national project. Whereas PauseAI and A Narrow Path begin from a moratorium or time-bound prohibition on superintelligence, Belfield's approach assumes continued progress toward very advanced, potentially superhuman general-purpose systems, provided these institutional safeguards can be made strong and reliable enough to manage the associated risks.

**AI redlines are necessary but insufficient.** Some of the issues with such an approach are power concentration dynamics and weak enforcement. If, for example, China and/or the US comply with the treaty at first but keep significantly more compute resources than other nations, post-AGI they could still quickly grow more powerful than other nations by using more and better AIs to accelerate their technological development and industrial capacity, potentially leading them to become more capable of entrenching power in other ways — like buying mining rights or establishing a presence in space or in the ocean that is difficult to remove. Even if states can coordinate to ensure that AIs are not being used to build new weapons, states with better AI capabilities can still develop far more advanced technologies and industries than others, which can be quickly redirected to gain a massive military advantage. These problems are especially salient in scenarios where robotics and AI progress is fast.

**todo find better signpost name**
**MAIM better addresses gradual domination but requires conditions it cannot itself create.** A mature MAIM regime would better address the risk of gradual domination than AI redlines alone — but only imperfectly. To function properly, MAIM requires roughly comparable compute resources among states, yet current US dominance makes this difficult to achieve without prior redistribution. Enforcing fairness through a dynamic quota system would address this, but pure compute parity also entrenches inequalities for the most populous nations, who would receive less compute per capita.

**Without the IOCCR, redlines and MAIM face an unresolvable verification and enforcement gap.** Beyond these structural weaknesses, there is a more fundamental problem: without the IOCCR, it will most likely be too difficult to properly implement something like MAIM or AI redlines in the first place. AI lacks the excludable inputs that make nuclear non-proliferation regimes viable. As Horowitz and Kahn (2025) argue, the nuclear regime is essentially an export cartel on scarce physical materials — plutonium and enriched uranium — that nobody can access without cartel members. AI has no equivalent: some of its key inputs — algorithms, data, and technical knowledge — are intangible and easily copied. Current Western dominance over chip supply chains is temporary as Chinese companies are actively working to replicate advanced manufacturing capabilities — slightly less sophisticated but still highly effective expertise is already proliferating. Distributed training techniques make it possible to spread computation across numerous small servers, and tracking what the already very distributed running compute is being used for is exceedingly difficult. Most IAEA-for-AI and redline proposals do not even propose to do this, though some touch on the possibility of a

coarse-grained level of monitoring.[13] Rapid algorithmic improvements further undermine compute-centric monitoring: increasing efficiencies in training algorithms, post-training enhancements and potentially totally new model architectures can dramatically and quite suddenly improve performance, meaning that much of the capability growth can in principle be achieved on relatively modest, covert compute budgets rather than ever-larger visible training runs. This could go especially fast if advanced AIs are used to accelerate AI research itself.

TODO: integrate text with its footnote: Advances in distributed training further allow large runs to be spread across many machines with only a multi-fold (rather than catastrophic) efficiency penalty.[14]

Historical evidence confirms that accurately assessing the state of a rival's technological development is exceedingly difficult — even for programs like nuclear weapons, which involve distinctive physical infrastructure, well-understood science, and relatively clear developmental thresholds. For example, South Africa secretly built six nuclear devices; U.S. intelligence had suspicions but no concrete details, and only learned the full truth in 1993 when President de Klerk publicly announced their destruction. Similarly, the CIA entirely missed China's uranium enrichment program at Lanzhou prior to its first nuclear test in 1964, dismissing the facility as an "incomplete and possibly incompleteable" [sic] gaseous diffusion plant.

Treaty commitments offer little additional assurance: North Korea signed the Nuclear Non-Proliferation Treaty (NPT) in 1985, violated its safeguards agreements throughout the 1990s, and ultimately withdrew in 2003 to develop nuclear weapons. Iraq similarly pursued a clandestine nuclear weapons program while under IAEA safeguards, only revealed after the Gulf War in 1991. The Soviet Union signed the Biological Weapons Convention (BWC) in 1972 while simultaneously expanding Biopreparat — a clandestine network of bioweapons facilities employing roughly 40,000 people, developing weaponized anthrax, plague, and smallpox variants. When the first defector Vladimir Pasechnik emerged in 1989, he revealed the program was ten times larger than U.S. and British intelligence had estimated; the full scale only became clear after Ken Alibek defected in 1992, having served as First Deputy Director of Biopreparat.

---

[13] Most governance proposals focus on training-compute thresholds and facility-level inspections rather than fine-grained monitoring of what running compute is being used for. See Heim et al., "Computing Power and the Governance of AI", GovAI, 2024; Egan and Heim, "The Role of Compute Thresholds for AI Governance", LawAI. For surveys of international AI institution proposals that similarly focus on training thresholds and inspections rather than runtime monitoring, see Maas and Villalobos, "International AI Institutions: A Literature Review", LawAI; and Wasil (2024).

[14] Empirically, large-scale training on well-engineered multi-node clusters can achieve high scaling efficiency when using careful combinations of data, model, and pipeline parallelism, communication overlap, and gradient compression (see, for example, the characterization in "Characterizing the Efficiency of Distributed Training" and practical scaling guides from GPU-cluster providers such as "Multi-Node GPU Clusters Explained"). Recent work on communication-efficient strategies like DiLoCo shows that, when well tuned, distributed methods can match or even outperform standard single-datacentre data-parallel training in some regimes while communicating orders of magnitude less (see the original DiLoCo paper, "DiLoCo: Distributed Low-Communication Training of Language Models", and follow-up analyses such as "Scaling Laws for DiLoCo" and the OpenDiLoCo multi-datacentre experiments in "OpenDiLoCo: An Open-Source Framework for Globally Distributed Training"). In practice, achieving these gains requires high-bandwidth, low-latency networking and sophisticated systems engineering (for example, Nvidia's work on long-haul training in "Turbocharge LLM Training Across Long-Haul Data Center Networks with NVIDIA NeMo Framework"), but together these results indicate that advanced actors can spread large training runs over many machines while paying a multi-fold rather than catastrophic efficiency penalty.

Even when violations *are* detected, enforcement through international monitoring bodies remains fragile. As Wasil (2024) documents through case studies of the IAEA and the Organization for the Prohibition of Chemical Weapons (OPCW), such institutions depend on geopolitical consensus that routinely breaks down: the U.S. unilaterally withdrew from the Iran nuclear deal (JCPOA) after a change in political leadership, substantially weakening an agreement the IAEA had been successfully verifying; Syria's Chemical Weapons Convention violations went unenforced after Russia vetoed UN Security Council action; and when the U.S. determined that Russian forces had used chemical weapons against Ukrainian troops, no official OPCW investigation followed. Moreover, states can simply exit these regimes entirely — as North Korea's withdrawal from the NPT demonstrates — leaving monitoring bodies with no remaining legal basis for oversight.

History thus presents a sobering pattern: (1) states routinely defect from treaty commitments; (2) such defections frequently go undetected — even for programs involving distinctive physical infrastructure; (3) even when detected, enforcement depends on fragile great-power consensus that can collapse at any moment; and (4) states can simply exit these regimes entirely, removing any legal basis for oversight. AI development and the misuse of AIs to develop dangerous technologies lack these legible signatures and will increasingly lack human witnesses post-AGI — making them far more difficult to monitor and far easier to cheat on.

TODO: and those with the most advanced AI models would likely have limited fear of economic sanctions given the benefits and technological independence those models could quickly provide. For example in this AI futures article (add link) the author argues that once AIs become capable enough, developers will not care about losing the European market.

**Even with initial compliance, exit risk and competitive pressures lock states into racing dynamics that erode democratic systems and AI safety.** Even if states initially comply with redlines, MAIM, or a pause, the verification and enforcement failures documented above mean that no state can be confident its rivals will continue to do so. States and coalitions — especially if they have more compute resources and therefore the most advanced AI capabilities — can simply exit treaty commitments, and as their AI-driven economies grow more self-sufficient, sanctions would have little effect. This underlying fear of defection or exit incentivizes states to optimize for competitiveness rather than safety or the preservation of democratic systems.

This is because even if a rival nation initially uses its AIs only for rapidly developing civilian technologies and industries, there is no guarantee this will not change. Advanced AIs, robots, and the technologies and industries they have built can easily and likely quite covertly be repurposed for military advantage. This keeps nations locked in a racing dynamic: those who fall behind in AI-driven industrial and technological development remain at the mercy of rivals who lead and could redirect their capabilities toward military dominance.

Even rivals with comparable compute resources face these pressures, since the racing dynamic is also driven by how AIs are *used*, not only by how much compute is available. These racing dynamics produce the risks summarized in Section 2: incentives to remove humans from control loops for efficiency, deploying the most advanced AI systems in military applications, and cutting corners on AI control and alignment when possible — all of which increase the risks of both authoritarian entrenchment if AIs are aligned and AI takeover if AIs are misaligned.

**These enforcement problems apply to current proposals for a global pause on AI development.** The voluntary moratorium strategy proposes halting AI development — either immediately or once certain hazardous capabilities, such as hacking or autonomous operation,

are detected. As explained above, verifying a state's use of its compute resources will be extremely hard, and the competitive advantages that advanced AIs could provide are enormous. As *Superintelligence Strategy* notes, militaries actively desire the very capabilities a pause would restrict — autonomous operation, advanced cyber offense — making reciprocal restraint structurally implausible. Without robust enforcement mechanisms, states will either refuse to join or secretly accelerate AI progress through algorithmic improvements and new paradigms, which can be pursued very covertly on relatively small amounts of compute.[15]

The scale of covert progress is striking: MacAskill and Moorhouse (2025) argue that frontier-model progress is driven both by rapidly increasing training compute and by rapid algorithmic improvements in training, post-training, and inference efficiency. On current estimates, physical training compute has been scaling by around 4.5× per year, while algorithmic gains of roughly 3× per year in training efficiency and 3× per year in post-training enhancement, plus ~10× per year in inference efficiency, mean that much of the capability growth could in principle be achieved on relatively modest, covert compute budgets rather than ever-larger visible training runs. The possibility of discovering entirely new and much more efficient paradigms — especially if advanced AIs are used to accelerate AI research — makes covert progress an enduring problem[16]

MSADS is not opposed to developing the capability to pause — on the contrary, the monitoring and intervention mechanisms of IOCCR and MS-MAIM avoid the problem of toothless forms of cooperation and make coordinated pauses far more feasible to implement if this is the best available option.

**The IOCCR and MS-MAIM resolve these verification and enforcement failures.** With broad participation in the IOCCR and MS-MAIM, defections from agreements can be readily detected and addressed — and there are no good incentives to defect in the first place.

As detailed in Section 4, unlike an IAEA-for-AI body that can only monitor compliance and refer violations to the UN Security Council, the IOCCR would exercise direct compute control — maintaining transparency over AI usage and retaining the ability to cut a defecting state's access to these resources. Contrary to monitoring-only proposals, where inspectors see coarse proxies like cluster size and power draw, the IOCCR would maintain fine-grained understanding of AI systems' interactions with the external world. Moreover, because the IOCCR monitors most compute resources, the most advanced AIs will most likely be those developed under its oversight — giving the IOCCR superior AI development capabilities that no defecting state could match. Defection therefore carries immediate, automatic costs: any state that exits loses access to the IOCCR's shared compute — and any shared compute infrastructure located on its territory would most likely be paused or destroyed, along with its domestic chip production capacity if any — as well as access to the most advanced AI models and to the shared knowledge and technologies produced by these systems. These consequences are far more

---

[15] For discussion of how covert AI development can circumvent pause-style regimes given current and anticipated compute verification limits, see e.g. Ord, T. Inference Scaling Reshapes AI Governance (Forethought / arXiv, 2025), which emphasizes the difficulty of monitoring smaller-scale but high-impact research and the governance implications of inference-side scaling.

[16] The decomposition in this paragraph follows the quantitative breakdown in MacAskill, W. & Moorhouse, F., Preparing for the Intelligence Explosion (Forethought, 2025), which estimates annual growth rates of roughly 4.5× for training compute, ~3× for algorithmic efficiency in training, ~3× for post-training enhancements, ~10× for inference efficiency, and ~2.5× for inference compute, and documents the empirical literature underpinning these estimates in more detail.

severe and immediate than the sanctions and diplomatic pressure that monitoring-only regimes can impose.

Enforcement does not depend on fragile great-power consensus. MS-MAIM's distributed enforcement architecture ensures that a significant minority of members can safely intervene against defections via remote off-switches or remote destruction capabilities. This creates a climate of caution fundamentally different from the racing dynamics described above: rather than a race to the bottom where the least conscientious actor sets the pace and others are pressured to follow, here the most conscientious actors set the tone and others are pressured to comply. Multiple veto points mean no single state can block enforcement, and no single state — including the IOCCR itself — can abuse its position unchecked. Unlike the IAEA or OPCW, where a single Security Council veto can paralyze enforcement, obstruction by any one actor is insufficient to prevent intervention. In more extreme cases, since AI-related infrastructure would be located far from population centers, states with intercontinental ballistic missile capabilities retain the ability to intervene unilaterally or as a veto against situations they judge intolerable.

**The IOCCR and MS-MAIM also better break competition dynamics and create conditions for positive futures.** As explained in Section 4, by centralizing compute resources under robust multi-sovereign oversight, the IOCCR removes the competitive pressure that otherwise locks states into racing dynamics — pressure to rapidly deploy AIs in military applications, remove humans from control loops for efficiency, and subordinate civilian AI use and democratic oversight to competitiveness.

With these pressures removed, states can more effectively coordinate around positive futures: concentrating compute resources and talent to accelerate the development of safe, properly controlled advanced AIs; using these advanced AIs to accelerate defenses against post-AGI threats — cyber-defenses, bio-defenses, proliferation controls, and defense-dominant military technologies; implementing coordinated global development pauses when needed; achieving equitable resource distribution; and coordinating on shared global challenges like climate change.

The defense asymmetry this creates is key: under MSADS, the institutions defending against rogue threats most likely command vastly more compute and more advanced AIs than any actor attempting to circumvent the system. This concentrated defensive capability — combined with the severe costs of defection and the abundance, prosperity, and security that MSADS enables — means there is no rational incentive to defect, creating a self-reinforcing dynamic that strengthens over time as trust builds through verified compliance.

**Resisting MS-MAIM creates nuclear instabilities.** States that resist win-win cooperation despite understanding its feasibility — even as others communicate justifiable discomfort — signal potential belligerent intent. Consider Russia's position if the US achieves post-AGI dominance without transparency over compute usage: growing uncertainty about US intentions and the survivability of their nuclear retaliatory capabilities would create pressure to escalate before the US gains decisive advantages. Initial escalations can help test the US's intents, and if those are unsuccessful in pressuring the US to provide credible security assurances, Russia could decide to use conventional missile salvos of increasing size on AI-related infrastructures. If insufficient, they may calculate that limited nuclear salvos of increasing size offer the best

remaining leverage to pressure US cooperation before dominance is consolidated. This escalation pattern is consistent with Russia's defensive escalation management doctrine[17].

Even states with nominal no-first-use policies, such as China, might under severe pressure consider similar options if they feared an emerging U.S. capability to substantially degrade their nuclear retaliatory capabilities, as some analysts argue. Existing technologies already threaten to erode nuclear retaliatory capabilities, and a substantial body of scholarship warns that using them to pursue damage-limitation strategies against major powers would yield limited benefits while increasing arms-race dynamics and escalation risks. Glaser, for example, argues that "the U.S. ability to destroy much of China's capacity to retaliate could generate a variety of escalatory dangers, including accidental, unauthorized, and inadvertent nuclear attacks, as well as early intentional limited nuclear escalation driven by concerns about U.S. preemptive attacks," and that U.S. efforts to preserve or enhance a damage-limitation capability would fuel strategic nuclear competition, further straining U.S.–China relations and increasing the risk of conflict.

The possibility of nuclear use should not be taken lightly: 80,000 Hours estimates the chance of a nuclear war in the next 100 years at around 20–50% — without accounting for the heightened risks of the AGI transition. A US-Russia nuclear conflict would be catastrophic for every state: nuclear winter caused by soot blocking sunlight would kill 5.5 billion people from famine alone in expectation — a figure that aggregates across limited, moderate, and full-scale escalation scenarios weighted by their probability of occurring, meaning a full-scale exchange would produce considerably higher casualties still.

Partially successful MAIM-like cooperation may give states enough security that they decide escalating to missile strikes isn't worth the risk — but as long as MS-MAIM isn't established, instabilities could remain. Fast and opaque technological progress could make states sufficiently worried about the survivability of their nuclear forces that they adopt dangerous doctrines — like predelegated launch authority or launch-on-warning postures — that increase retaliatory survivability but also the likelihood of accidental, unauthorized, or inadvertent nuclear use.

### todo needs improvement incist on the remaining competitive pressures in all three and why they remain

**The IOCCR and MS-MAIM combination is our best cooperation option.** Without them, AI redlines, MAIM, and global pause proposals face an unresolvable verification and

---

[17] (todo place it somewhere else) Russia's escalation management doctrine is sometimes mischaracterized in Western literature as offensive "escalate to de-escalate" signaling — the idea that Russia plans to use nuclear weapons early and aggressively to force war termination on favorable terms. As Kofman and Fink's War on the Rocks analysis details, based on a review of over 700 Russian-language military publications, this characterization is incorrect. Russia's strategic deterrence framework is fundamentally defensive in orientation: it envisions graduated coercive measures — from demonstrative acts to limited conventional and nuclear strikes — aimed at managing threats to Russian sovereignty, deterring aggression, and offering adversaries off-ramps toward negotiated outcomes. The purpose of limited strikes is explicitly to shock opponents into reassessing the costs of continued aggression while preserving pathways to de-escalation — not to achieve offensive dominance. There is strong doubt in Russian military circles that political leadership would authorize early preemptive nuclear use; the consensus is that attempts to coerce with nuclear weapons early on would not be credible. The escalation pattern described in this paper reflects this defensive logic: Russia escalating in response to perceived existential threats to its retaliatory capabilities, not pursuing unfair advantages.

enforcement gap — and even if enforcement were possible, all three remain structurally insufficient: redlines alone cannot prevent gradual power concentration; MAIM cannot meet its own prerequisites for compute parity and leaves post-AGI competitive pressures intact; and pause proposals cannot prevent covert progress through algorithmic improvements. The underlying fear of defection or exit locks states into racing dynamics that erode democratic systems and AI safety. The IOCCR and MS-MAIM resolve the enforcement gap through direct compute control and multi-sovereign deterrence, while also breaking competition dynamics — better enabling states to prevent tyrannical power concentration, coordinate global development pauses, accelerate defenses against post-AGI risks, achieve fairly distributed prosperity, coordinate on shared challenges like climate change, diminish nuclear instabilities, and entrench durable peace.

## Why the IOCCR and MS-MAIM Combination Is Our Best Cooperation Option

### The Main Challenge: Verifying and Enforcing Cooperation Is Extremely Difficult

AI training and deployment can be distributed across numerous small servers. AI development and the misuse of AIs to develop dangerous technologies could be easier to secretly cheat on than nuclear or bioweapons programs, as they lack their legible physical signatures — and historical evidence shows that states routinely cheated (often unnoticed) on such programs:

**South Africa** secretly built 6 nuclear devices — U.S. intelligence had suspicions but no details; full truth revealed only when they announced their destruction in 1993

**Soviet Union** signed the Biological Weapons Convention while expanding Biopreparat — a network of ~40,000 people developing bioweapons. The program was 10× larger than Western intelligence estimated

**China** — the CIA entirely missed China's uranium enrichment program at Lanzhou prior to its first nuclear test in 1964, dismissing the facility as incomplete

**Iraq** pursued a clandestine nuclear weapons program while under IAEA safeguards — only discovered after the Gulf War in 1991

**North Korea** signed the NPT in 1985, violated its safeguards throughout the 1990s, and withdrew in 2003 to develop nuclear weapons

Post-AGI, dangerous AI developments and misuse could increasingly lack human witnesses, making them even more difficult to monitor and easier to cheat on. Without MS-MAIM's enforcement mechanisms, states can also simply withdraw from cooperation — as North Korea did.

| AI Redlines Only | MAIM | IOCCR + MS-MAIM |
|---|---|---|
| Cannot prevent gradual power concentration — states with more compute can accumulate insurmountable geopolitical advantages, putting tremendous competition pressures on others | • Requires compute parity; current US dominance makes it difficult<br>• Even with parity, rivals could quickly and covertly repurpose AIs and AI-built capabilities for military dominance<br>• Competitive pressures to bypass democratic oversight persist | • Breaks dangerous competition dynamics<br>• Enables better coordination around diminishing post-AGI risks and implementing positive futures<br>• Makes coordinated AI development pauses actually enforceable<br>• Enforcement mechanisms are stable and reliable |

**These approaches are cumulative:** developing imperfect but better capability to implement AI redlines, MAIM, and the capability to pause is valuable in its own right and can serve as steps toward MS-MAIM.

*TODO this figure needs updates Figure X: The core verification challenge and why the IOCCR and MS-MAIM combination is our best cooperation option compared to alternatives.*

todo: did you mention the point of tyranny thats explained in section 3 and four

todo: should you say the problems associated with racing dynamics summarized in section 2 (and make sure they are complete)

todo: add some have argue that because it would be difficult to implement MAIM some have propose a softer form of coordination around the pace or AI progress than MAIM and the arguments for why MAIM is suboptimal would be essentially amplified in this context

TODO: add points about pause ai strategy and the difficulty to do it

## 6. Establishing a functional IOCCR and MS-MAIM Regime: Europe's levers

This section summarizes the third component of MSADS: strategic approaches for establishing the IOCCR and MS-MAIM across different scenarios. It details how France, the UK, the EU — and more broadly, any willing coalition of states — can increase the chances that the IOCCR and MS-MAIM get established, and that they get established in a timely manner. The section argues that France, the UK, and the EU will be especially important early actors because they have both stronger leverage to catalyze cooperation and will likely see stronger incentives than the US and China to initiate it earlier.

Much of this section's logic generalizes for augmenting the chances of reaching other win-win cooperation frameworks.

**The dual-track approach to establishing MS-MAIM combines democratic coalition-building with global cooperation.** To maximize the chances of safely establishing MS-MAIM, I propose a dual-track approach. The first track forms a coalition of democratic countries that centralize compute resources and talent — possibly through a CERN for AI or Intelsat for AI model. The second track pursues direct cooperation with all countries — including potential rivals — to establish MS-MAIM globally. If broad international cooperation from track 2 succeeds, states can escape racing dynamics by implementing MS-MAIM — and the earlier this happens, the better, ideally even before AGI development.

**The democratic coalition helps foster global cooperation and avoid a tight race.** States may resist the establishment of MS-MAIM, but to be useful, MSADS doesn't need to successfully establish MS-MAIM directly. If the democratic coalition succeeds in centralizing compute and talent, it can at least avoid a tight race and establish a clearer AI lead — making it more likely the coalition will pause AI progress at crucial moments (for example, before AIs become the main driver of AI research)[18] to conduct alignment work. A clear coalition advantage also pressures other states to join MS-MAIM, strengthens the coalition's bargaining position in

[18] Once AIs can design their own successors, the values of future systems will hinge on whether the earlier AIs are truly aligned. If they are misaligned or only deceptively aligned, they may deliberately shape successor systems to pursue their own objectives, potentially locking in misalignment across generations. Successfully aligning the first AIs that influence later designs therefore increases the chance that future AIs will also be aligned with human values. TODO one paragraph

negotiating IOCCR terms, provides better defense against adversarial scenarios, and serves as an early template for broader cooperation.

If the coalition demonstrates the feasibility and mutual benefits of the IOCCR and MS-MAIM, public opinion in holdout democracies could create domestic political pressure on their governments to join. Citizens who observe neighboring countries successfully cooperating and understand that breaking competition pressure better enables them to benefit from equitable AI-generated prosperity, reduced military tensions, and safer AI and democratic governance will increasingly question why their own governments remain outside. For authoritarian holdouts less responsive to public pressure, the mechanism differs but the outcome is analogous: should the coalition become sufficiently powerful, the growing strategic costs of remaining outside — reduced access to the best AIs and increased military vulnerability — create their own incentives to eventually participate.

**Communicating and executing escalation ladders coerces win-win outcomes.** As Eric Drexler explains, the challenges of the transition to a cooperative future motivate “the concept of 'coercive cooperation' — the use of pressure to move an adversary toward mutually beneficial outcomes despite knowledge gaps, institutional inertia, or weak foundations for trust." Communicating and executing escalation ladders when appropriate implements this coercive cooperation approach.

AI deterrence is our best option for tempering racing dynamics and great power conflicts — a logic that Dan Hendrycks and Adam Khoja defend convincingly against critics. MS-MAIM is a mature version of the deterrence logic and a feasible stable win-win equilibrium — but some states might not immediately recognize this (knowledge gaps) or may resist its establishment due to trust deficits or institutional inertia, miscalculating either that adversaries will exploit cooperation to achieve competitive gains or that they themselves can achieve unilateral advantages through AI dominance. To prevent such obstacles from escalating into existential crises, democratic states should communicate escalation ladders and ambiguous escalation thresholds in advance. As in nuclear doctrine, communicating ambiguous rather than precise thresholds avoids locking states into automatic escalation to preserve credibility, instead allowing context-appropriate responses while still keeping adversaries cautious. Communicating escalation ladders and thresholds raises awareness of the seriousness of non-cooperation risks while providing graduated pressure — enough to shift the adversary's cost-benefit calculus without triggering uncontrolled escalation — to coerce resistant states toward win-win outcomes.

Any capable nation whose national security is sufficiently jeopardized will eventually communicate such thresholds — particularly once smaller measures like trade restrictions or sabotage prove ineffective. These thresholds could include conditions such as insufficient transparency over compute usage coinciding with either AIs automating substantial AI research or leading actors entrenching dangerous and difficult-to-reverse advantages. Since this communication will happen eventually, doing it early raises awareness among citizens, analysts, and policymakers — helping resistant nations overcome domestic obstacles to cooperation that serves their citizens' best interests.

Interventions could escalate in the following order: EU market restrictions, cyberattacks, disrupting chip supply chains, or other AI asset sabotage. If a state continues resisting MS-MAIM after these interventions, this signals belligerent intent, naturally justifying warnings of

higher escalations — such as conventional missile strikes on datacenters or energy production infrastructure. These would be proportionate responses to the existential threat posed by non-cooperation. Ultimately, in extreme scenarios, it can be valuable to explicitly threaten and possibly resort to limited nuclear strikes of increasing size, though such escalation can spiral out of control and should therefore be done with extreme caution.

**TODO if context appropriate**

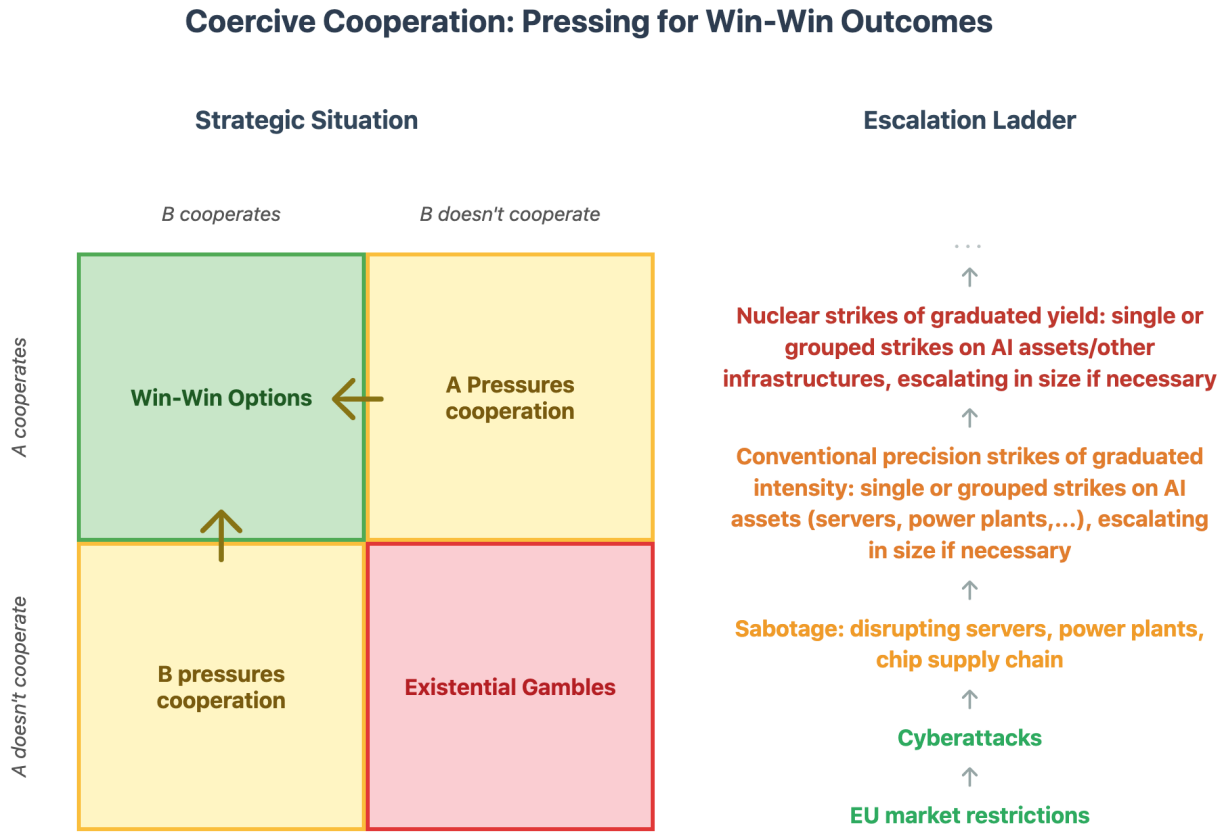



**Figure :** (UNFINISHED FIGURE) Left: MS-MAIM is a feasible win-win equilibrium, but some states may fail to recognize this due to knowledge gaps, trust deficits, or institutional inertia. As Drexler explains, coercive cooperation uses pressure to move a non-cooperative rival toward mutually beneficial outcomes despite such obstacles. A state or coalition can use the graduated escalation ladder (right) — from market restrictions through cyberattacks and sabotage to conventional precision strikes of graduated intensity and, in extreme scenarios, nuclear strikes of graduated yield — to shift a reluctant rival's cost-benefit calculus toward cooperation. These escalations can also inflict real costs on non-cooperative rivals that slow and most likely prevent potential bids for dominance. When neither state cooperates, both face existential gambles. As in nuclear doctrine, communicating ambiguous deterrence thresholds avoids locking states into automatic escalation to preserve credibility, instead allowing context-appropriate responses while still keeping adversaries cautious.

todo remove the top arrow to … and put market restrictions instead of eu market restrictions
todo remove EU from market restricitons, usekinetic attacks of increasing size and limited nuclear salvo of increasing size on key assets

**Getting broad international cooperation to start earlier is critical.** Post-AGI, compute resources will be like uranium today — the raw element from which ultra-destructive power can be built. Just as international cooperation manages uranium proliferation, and just as the ECSC controlled coal and steel, international cooperation should control compute resources. Today's concentrated AI infrastructure and centralized chip supply chains create favorable conditions for installing MS-MAIM — conditions that may not persist if leading AI actors start adopting costly defensive adaptations like distributed training techniques. Delay exposes us to competition dynamics while allowing leading AI actors to consolidate advantages that become increasingly difficult to reverse — increasing the risks of an unfair cooperation framework. As defensive measures proliferate, only dangerous escalations could succeed against the miscalculations of determined belligerent actors.

**Credible warnings make AI domination attempts strategically irrational.** States seeking AI domination face extreme national security risks and are unlikely to succeed. China and the US may decide to pursue costly defensive adaptations like hardening and decentralizing AI assets despite AI capability intervention thresholds being crossed. Such decisions would be irrational — they would be unlikely to maintain an AI monopoly without risking inviting devastating responses from rivals. For example, Russia's defensive escalation management doctrine includes the possibility of resorting to limited nuclear strikes of increasing size to protect against threats for which they lack adequate conventional responses.

If, for example, the US pursues AI dominance, Russia and possibly other nuclear states would likely carry out such escalations, putting the US economy and population in grave danger. The US cannot defend against such attacks — its extremely expensive missile defense systems can only partially protect against small rogue actors, like North Korea, with outdated ballistic missile technologies. Protection against states with advanced ballistic missile technologies like Russia, China, France, or the UK is minimal. In today's highly integrated economies, both French and UK arsenals would be capable of killing 20% of the US population and crippling 50% of its economy — a threshold associated with assured destruction capability. Even a few nuclear strikes against key infrastructure like ports would prove devastating. France, the UK, Russia and China, therefore clearly possess credible deterrence capabilities. However, post-AGI technological developments could threaten this stability, making preemptive preparations essential.

**France and the UK should strengthen nuclear deterrence for post-AGI nuclear stability and signaling.** There is a risk that post-AGI, the US or China — if they gain sufficient advantages — might attempt to preemptively disarm their opponents' nuclear arsenals or instill sufficient doubt in their rivals' minds that they could do so to gain bargaining power."

In general, I do not expect states to attempt nuclear disarmament of rivals in the early post-AGI years. Preemptively eliminating an adversary's nuclear arsenal is an uncertain, untested and highly perilous operation. Also, analyses of nuclear deterrence that do not take into account the possibility of the disruption of AGIs suggest that for rivals responsive enough to emerging threats, maintaining nuclear deterrence is substantially easier than denying it.

However, France and the UK must be responsive to emerging threats, and their nuclear deterrence currently depends mainly on submarines. While submarine-based deterrents remain robust today, emerging technologies in anti-submarine warfare — including AI-enabled sensors, autonomous undersea vehicles, and quantum sensing — pose increasing long-term challenges to submarine survivability. A 2024 Lawrence Livermore National Laboratory report warns that such advances "could remove the 'guaranteed' from 'guaranteed second strike,'" meaning

previously invulnerable submarines might become detectable and vulnerable to preemptive attack.

Post-AGI technological acceleration compounds these risks. Some analysts warn that rapid breakthroughs, especially under an intelligence explosion scenario, could significantly undermine the credibility of nuclear deterrence. This threat intensifies if rivals maintain post-AGI advantages over extended periods. France and the UK are currently only months behind in the AI race, but this gap will widen and consolidate once AIs drive most AI research and compute resources become the primary determinant of capability.

To counter these problems, the UK and France should prepare more resilient nuclear forces. Some defense analyses recommend road-mobile intercontinental ballistic missiles as an insurance hedge against potential anti-submarine warfare breakthroughs. France could, for example, collaborate with willing host states to deploy French-owned, French-operated road-mobile nuclear launchers across a larger territorial depth — an ambitious but potentially highly effective way to significantly increase survivability by complicating adversary targeting and providing redundancy should underwater stealth advantages erode. Less politically difficult adaptations are also possible.[19]

((TODO: add link to my text on dangerous nuclear strategies or footnote explaining it)) Having resilient nuclear deterrence reduces nuclear instabilities, as states uncertain about their forces' survivability face pressure to adopt dangerous strategies — including predelegated launch authority or launching preemptive limited nuclear strikes before rivals acquire capabilities that could deny their deterrence. todo maybe use another link here

Preparing such forces in advance also serves as an implicit or explicit signal of concern about the AI situation and demonstrates serious commitment to deterrence. In practice, I do not expect states will need to resort to higher escalations for cooperation to occur — but having credible deterrence makes escalation less likely, as opponents would anticipate potential debilitating responses to domination attempts.

**France, UK, and EU are well positioned to initiate early MS-MAIM cooperation.** I expect the UK and France to have credible enough intervention capabilities to greatly augment the chances of achieving positive win-win futures. They also have stronger incentives to push for cooperation, as, post-AGI, they will have fewer compute resources than the US and China and face a constant competitive disadvantage[20]. The EU shares these incentives while also being

---

[19] Deploying French nuclear road-mobile launchers in willing host states would face substantial doctrinal, political, and legal hurdles: it would break with France's long-standing two-leg "strict sufficiency" posture and preference for nationally based forces, and it would run into contested interpretations of non-proliferation rules, domestic opposition in host countries, and alliance-management costs. Still, it is not unprecedented for a nuclear-armed ally to station nuclear weapons abroad: the United States already deploys nuclear bombs on the territory of several NATO allies, intended for delivery by those allies' aircraft in wartime. Less politically difficult adaptations include deepening hardening and redundancy of French submarine and air-based forces, or re-introducing a limited land-based leg on French territory using deeply hardened fixed launchers instead of dispersing mobile systems abroad.

[20] Today France and the UK remain competitive in the AI race, but once AIs automate most AI research, those with more compute will experience faster capability improvements and maintain persistent advantages. According to the State of Global AI Compute 2025 report, China currently has roughly 2× the frontier training compute that the EU possesses, while the US has approximately 17×. China is also structurally positioned to centralize national compute and scale chip production rapidly. In this scenario,

better positioned as a neutral mediator: without the intense US-China military rivalry, Chinese leaders may view EU-led cooperation proposals as more credible and less threatening. Moreover, the EU has extensive experience facilitating international cooperation frameworks. Similar arguments apply to any sufficiently capable coalition of willing states — one that need not encompass the entire EU and could begin with a smaller group of countries, including non-EU members.

**Concrete steps toward a functional IOCCR and MS-MAIM regime.** France, the UK, and the EU should prepare for MS-MAIM implementation. This requires immediate investment in research on compute verification and monitoring mechanisms, governance frameworks, strategic approaches to deterrence and coercive cooperation, and continued comparative analysis of cooperation architectures — including the IOCCR and MS-MAIM and other approaches — to strengthen the case for tight cooperation and identify the strongest implementation pathways. These preparations are especially critical given the possibility of an intelligence explosion. In such critical scenarios, having actionable plans ready enables Europe to quickly establish fair cooperation with willing states while preventing resistant actors from achieving dangerous AI advantages.

European leaders should begin early discussions with potential MS-MAIM partners about implementation pathways. Initial cooperation efforts could start with models like CERN for AI or Intelsat for AI initiatives — centralized research facilities with international governance designed with open membership, allowing states to join when ready, and focused on developing competitive AI capabilities. This approach could particularly attract countries seeking greater AI sovereignty, ensuring participating nations maintain technological competitiveness rather than becoming dependent on US or Chinese AI systems. Such initiatives lay out the foundation for later, more mature forms of cooperation with greater control over compute resources and a mature MS-MAIM regime. To accelerate broader coalition formation, a coalition with open membership could start small — with Europe's largest economies or even just the Benelux countries — before expanding. Similar steps could be taken by any sufficiently capable coalition of willing states, with or without full EU participation.

France and the UK should strengthen their deterrent capabilities. France, for instance, could cooperate with willing host countries to ensure the resilience of its nuclear forces — both as a genuine defensive hedge and as a signal of serious commitment to deterrence. This could range from fixed launcher sites to deploying French-owned, French-operated road-mobile launchers across willing host countries.

France and the UK in particular should define intervention thresholds and communicate credible escalation ladders to potentially resistant states in advance. Coalition members should then use concrete actions — strengthening deterrence, communicating thresholds, and working on common AI projects and regulations — to proactively shape public opinion in resistant states toward recognizing the benefits of MSADS and the costs of remaining outside. These actions work on two levels: they directly pressure resistant governments toward caution, and, crucially, they draw sharp attention from analysts and public opinion in those states. When paired with clear explanations of why tight cooperation benefits all nations — including the US and China — the underlying arguments gain far more weight than they would receive through ordinary diplomatic channels, helping shift opinion toward cooperation earlier than might otherwise occur. This matters because the window for establishing MSADS under favorable conditions narrows

---

France and the UK face increased risk of losing geopolitical influence in a post-AGI world dominated by compute superpowers.

as leading AI powers approach decisive capabilities; accelerating the political shift is itself a core function of these coalition actions.

MSADS is valuable even without universal IOCCR participation — building a coalition and preparing credible deterrence gives Europe better defenses against adversarial scenarios. And much of this section's logic generalizes: the same coalition-formation, deterrence, and communication tools can support other tight win-win cooperation frameworks should refined or alternative architectures emerge. The window for establishing the IOCCR and MS-MAIM under favorable conditions is rapidly closing: research and preparation must begin immediately.

**Concrete steps toward a functional IOCCR and MS-MAIM regime.** France, the UK, and the EU should prepare for MS-MAIM implementation. This requires immediate investment in research on compute verification and monitoring mechanisms, governance frameworks, and strategic approaches to deterrence and coercive cooperation. This requires immediate investment in research on compute verification and monitoring mechanisms, governance frameworks, strategic approaches to deterrence and coercive cooperation, and the rigorous formulation and comparison of cooperation options — including the IOCCR and MS-MAIM but also other alternatives — to clarify which architectures best serve all nations' interests under different scenarios. These preparations are especially critical given the possibility of an intelligence explosion. In such critical scenarios, having actionable plans ready enables Europe to quickly establish fair cooperation with willing states while preventing resistant actors from achieving dangerous AI advantages.

European leaders should begin early discussions with potential MS-MAIM partners about implementation pathways. Initial cooperation efforts could start with models like CERN for AI or Intelsat for AI initiatives — centralized research facilities with international governance designed with open membership, allowing states to join when ready, and focused on developing competitive AI capabilities. This approach could particularly attract countries seeking greater AI sovereignty, ensuring participating nations maintain technological competitiveness rather than becoming dependent on US or Chinese AI systems. Such initiatives lay out the foundation for later, more mature forms of cooperation with greater control over compute resources and a mature MS-MAIM regime. To accelerate broader coalition formation, a coalition with open membership could start small — with Europe's largest economies or even just the Benelux countries — before expanding. Similar steps could be taken by any sufficiently capable coalition of willing states, with or without full EU participation.

France and the UK should strengthen their deterrent capabilities. France, for instance, could cooperate with willing host countries to ensure the resilience of its nuclear forces — both as a genuine defensive hedge and as a signal of serious commitment to deterrence. This could range from fixed launcher sites to deploying French-owned, French-operated road-mobile launchers across willing host countries.

France and the UK in particular should define intervention thresholds and communicate credible escalation ladders to potentially resistant states in advance. Coalition members should then use concrete actions like strengthening deterrence, communicating thresholds, and working on common AI projects and regulations to proactively shape public opinion in resistant states toward recognizing the benefits of MSADS and the costs of remaining outside.

MSADS is valuable even without universal IOCCR participation — building a coalition and preparing credible deterrence gives Europe better defenses against adversarial scenarios. The window for establishing the IOCCR and MS-MAIM under favorable conditions is rapidly closing: research and preparation must begin immediately.

***Figure :*** *(UNFINISHED FIGURE) The top and middle parts of the diagram explain MSADS's strategic approach for establishing a functional IOCCR and MS-MAIM regime with comprehensive international participation as early as possible, including responses to different scenarios. The bottom part explains what happens once MS-MAIM is established and how it enables positive futures, before concluding on why urgent preparation is needed.*

# MSADS in a Nutshell

## Dual Track Approach

### Building a Democratic Coalition for Advantage

Form unified AI project (like CERN for AI or Intelsat for AI) with value-aligned states that centralizes talent and compute resources.

### Preparing and Pressuring for IOCCR and MS-MAIM Establishment

- Research IOCCR architecture and how to implement MS-MAIM
- Engage potential rivals in preliminary discussions on MS-MAIM cooperation
- Strengthen nuclear deterrence to hedge against post-AGI threats and signal resolve
- Signal willingness to escalate against the threat of non-cooperation
- Use this signaling to increase awareness and shift rivals' public and analysts' opinion toward seeking cooperation

### Full Cooperation

All major powers agree to cooperate

**Response:**

Establish IOCCR and implement MS-MAIM requirements (compute traceability, usage transparency, AI asset vulnerability, ...) immediately because racing dynamics are dangerous.

### Partial Resistance

Example: US cooperates with coalition, China/others resist MS-MAIM

**Combine two responses:**

**Racing for advantage**

- Provides greater AI development pause flexibility
- Pressures for better deals
- Improves coalition's defensive capabilities

**Carrying interventions when appropriate**

EU Market restrictions → Cyberattacks → Sabotage

### Major Resistance

US/China pursue defensive adaptations (hardening and decentralization of AI assets) despite escalations and AI capability red lines being crossed.

**Response: Escalate cautiously and strategically before window closes**

- UK/France communicate increased willingness to use conventional strikes → limited nuclear strikes if necessary
- Intervening before rivals entrench advantages that could undermine deterrent capability can be worth the risk — UK/France most likely retain sufficient nuclear deterrence in first post-AGI years and US/China should not assume inaction
- Escalations help gauge rivals' true intentions and adjust appropriate responses

### Positive Future

**Once MS-MAIM is established:** Makes it easier to concentrate AI research in one or few highly secured, safety-focused projects; implement temporary pauses when needed; and enable fair differential compute allocation with equitable distribution of AI-produced value.

**Initial phase:** Concentrate compute to accelerate AI progress (pause capability ensures it is done safely). This improves post-AGI security by: (1) using the most advanced AIs to accelerate defenses and (2) discovering dangerous capabilities earlier—providing more time to prepare before threats emerge.

**After initial phase:** States regain more autonomy over compute resources for use that doesn't threaten others.

**Breaking competition dynamics enables greater focus on positive futures:** solving climate change, curing disease, maintaining democratic structures.

**Entrenching peace and democratic governance:** Multiple intervention points of MS-MAIM help entrench peace and ensure democratic decisions at the IOCCR level.

---

**Conclusion: Acting early is better** — AGIs will likely be developed within the next decade, with an intelligence explosion possibly happening sooner. Post-AGI leading AI powers will have greater advantages and be less prone to act prosocially.

## Conclusion: MSADS Is Our Best Option — It Needs Urgent Preparation

Todo: verify that the list of risks is comprehensive
Maybe add short recap on timeline context

**Seeking AI dominance is self-defeating and endangers all nations.** Some analysts and policymakers argue that the best strategy is for the US, possibly with allies, to race for first-mover advantages in AGI development. This argument takes two forms. The first envisions a largely unregulated post-AGI competition in which the US should, for its security, maintain superior AI, technological, and military capabilities indefinitely. The second envisions using advanced AIs to quickly gain an overwhelming military advantage[21] and then negotiate favorable international agreements that remove post-AGI competition risks.

As argued throughout this report, an unregulated race to AGI and an unregulated post-AGI competition each carry enormous risks, as summarized in Section 2: competitive pressures incentivize states to prioritize competitiveness over safety — reducing willingness to pause for alignment work, cutting resources for AI control, lowering thresholds for deploying systems not yet proven safe, and moving away from deliberative democratic decision processes. Post-AGI, these dynamics become uniquely dangerous: delegating decisions to more efficient AI systems offers powerful competitive advantages, incentivizing the removal of human oversight across many sectors. These incentives risk gradually concentrating power in less accountable structures that totally loyal AIs could help to entrench — making authoritarian lock-in far more likely and durable than anything previously possible. If AIs are misaligned and in control of critical assets like military robots, this could lead to humanity's disempowerment. More broadly, competitive pressures create races to the bottom across every risk dimension — technical safety, proliferation control, democratic oversight, peacekeeping, and protection against AI misuse. Unregulated competition also forgoes the benefits of the IOCCR and MS-MAIM combination explained in Section 4.

---

[21] Something like being almost sure to have the ability to remove the nuclear second strike capabilities of all rivals and the ability to conquer any state with conventional forces.

As for seeking an overwhelming military advantage, this would most likely fail: nuclear rivals would notice and strategically escalate, if necessary resorting to interventions that inflict devastating damage on the infrastructure this strategy requires. Even assuming a decent chance of success, the attempt itself would create a period where nuclear instabilities are heightened and all post-AGI competition-associated risks are amplified.

MSADS' coalition formation and common AI projects capture the strategic benefits of seeking first-mover advantages — a stronger competitive position, better defense capabilities, and greater capacity to pause at critical moments — while also pursuing the IOCCR and MS-MAIM to escape competition dynamics whenever the main powers reach acceptable agreements. This makes MSADS a strictly better approach than advantage-seeking strategies for any AI-dominant state or coalition: it provides comparable security advantages while establishing what Section 5 argues to be our best cooperation option. For states with fewer compute resources and who face a constant competitive disadvantage against a dominant AI power that provides insufficient safety assurances, the calculus is even clearer: they have no better options than cooperation frameworks that constrain all parties.

todo add sources like situ awareness and the link in the maim report

**The IOCCR and MS-MAIM combination is our best cooperation option.** MSADS is not against useful initial steps like AI redlines, developing the capability to pause AI development, or establishing a MAIM regime — it seeks the benefits of these approaches. But as explained in Section 5, MS-MAIM makes it feasible to reliably enforce such commitments, which would be extremely difficult without it given historical evidence of treaty defection and the possibility of distributing training and running compute across multiple small servers. The IOCCR and MS-MAIM combination also more effectively removes competition pressures and sets up better conditions for international cooperation. As a result, it better addresses power concentration risks, democratic erosion, and nuclear instabilities than alternatives. It also better enables coordinated development pauses, accelerated post-AGI defenses against rogue AIs or AI misuse, equitable resource distribution, and durable peace.

**MSADS is feasible and robust regardless of how the transition unfolds.** As explained in Sections 4 and 5, the IOCCR and MS-MAIM option is feasible, and as explained in Section 6, there are tools to pressure resistant states toward this win-win cooperation. Should broad cooperation fail, the coalition-building part of MSADS still gives coalition members better defenses against adversarial scenarios — keeping them more competitive and providing credible intervention capabilities that pressure resistant states toward caution and cooperation, especially if nuclear states like the US, France or the UK are part of the coalition. If the US were part of the coalition, centralizing talents and compute with other members would provide a stronger AI lead over China and therefore greater leeway to pause at critical moments.

Even if AI never reaches AGI-level capabilities, the IOCCR and MS-MAIM combination remains valuable. At minimum, it fosters great power peace and ensures more equitable distribution of AI technology benefits — technologies built on humanity's collective knowledge that should not disproportionately benefit only a tiny fraction of the global population.

**The window for safe establishment is rapidly closing.** Domination attempts are self-defeating and partial cooperation without full MS-MAIM mechanisms will ultimately grow unstable and reveal weaknesses. At some point, pragmatic leaders will likely recognize the viability of the IOCCR and MS-MAIM combination and cooperate toward establishing it. But this recognition may come only after alarming signals — initial escalatory interventions, dangerously misaligned AGIs, a large post-AGI catastrophe, or the realization that post-AGI military pressures are becoming incompatible with democratic systems. Rather than letting this equilibrium develop chaotically, it is better to establish it safely and early.

AI actors are more likely to cooperate prosocially while uncertain about who will win the AI race, but once clear winners emerge and the prospect of acquiring geopolitical advantages grows, they may accept cooperation only under terms unfavorable to others — a dynamic that persists even under more gradual AI progress scenarios. The MS-MAIM requirements should ideally be implemented before AIs automate sufficient AI research to trigger recursive self-improvement and before leading AI actors begin entrenching difficult-to-reverse forms of power — thresholds beyond which enforcing compliance becomes both more critical and more dangerous.

Establishing a mature IOCCR and MS-MAIM combination requires sustained preparation that willing nations should begin immediately. Military and civilian strategists worldwide should jointly analyze optimal implementation approaches now. As AI capabilities keep advancing and the risks of advantage consolidation and defensive adaptations grow, the window for optimal action is closing. I expect that a failure to properly manage this historical turning point will most likely be summarized in two words: too late.

early work on verification measures is quite promising.

# Appendix B – Technical Limits of Missile defense

Despite decades of investment, U.S. homeland ballistic missile defense remains highly constrained even against *small* attacks. The Ground-Based Midcourse defense (GMD) system has achieved only about 50–60 percent success in carefully scripted intercept tests, often against single targets without realistic countermeasures (see Laura Grego's analysis, "Comments on the Ground-Based Midcourse Missile defense Test Record" and CSIS's summary of GMD testing in "Ground-based Midcourse defense – Media Resources). The Director of Operational Test & Evaluation has repeatedly described GMD as providing at most a "limited" capability against "emerging, uncomplicated" long-range threats, not robust protection against a determined adversary using sophisticated missiles (see the FY2010 DOT&E report on GMD, "Ground-Based Midcourse defense"). On this record, it is hard to claim that GMD can reliably defeat even a small salvo of a handful of intercontinental ballistic missiles (ICBMs), let alone a larger or more complex attack (see also Grego's post and the CSIS overview just cited).

Locating interceptors close to the assets they defend does not fundamentally change this picture. Midcourse intercept still occurs in space, where discriminating real warheads from decoys is intrinsically difficult and defender geography confers little advantage (see the general

discussion of midcourse discrimination problems in "Missile defense"). Flight-test campaigns have also been extremely limited: only one "salvo" test firing two interceptors at a single target was conducted in 2019, under tightly controlled conditions that fall far short of a real multi-missile attack (see George Lewis's critique "What Did FTG-11 Actually Prove?" and media reporting such as "US Carries Out Successful 'Salvo' Interception of ICBM Target" and "Homeland missile defense system takes out ICBM threat in historic salvo test"). Even under optimistic assumptions of around 60 percent single-shot reliability, defenders must fire multiple interceptors per target to reach high kill probabilities, which quickly depletes inventories once the number of incoming missiles rises (Lewis makes this point explicitly in his FTG-11 analysis).

These constraints apply even to the relatively narrow mission of defending against North Korea. GMD was explicitly designed for "limited" ballistic missile threats, and U.S. officials stress that it is "virtually useless in protecting against a mass nuclear strike from Russia or China" (see discussion in "America's Missile defenses Against North Korea Have a Big Problem (They Only Work Half the Time)" and similar assessments). As North Korea deploys more ICBMs and potentially multiple-warhead missiles, even this limited mission becomes tenuous: recent analysis notes that North Korea now fields enough ICBMs to "overwhelm the U.S. defense system against them" simply by sheer numbers (see e.g. Politico, "North Korea displays enough ICBMs to overwhelm U.S. defense system against them" and Korean reporting on U.S. assessments such as "'N. Korean ICBMs could impact US missile defense system,' says report"). In that context, U.S. strategic missile defense cannot be counted on to stop all incoming warheads; at best it complicates North Korean planning and may intercept some fraction of an attack.

The underlying problem is that offence tends to scale more cheaply and effectively than defense. Missile defense is vulnerable to "saturation" and "burn-through," where an attacker launches more missiles or warheads than the defender has interceptors or engagement opportunities. A Royal United Services Institute paper, "The Next Great Challenge to Missile defense," emphasises that saturation by exhaustion is "the simple expedient of attacking with more offensive missiles than the number of available interceptors can cope with," and notes that improvements in manufacturing will further reduce offensive costs relative to defense. Because interceptors are expensive and each engagement has a non-zero probability of failure, increasing the number of incoming objects — whether real warheads or decoys — rapidly drives down the probability that *all* warheads will be intercepted (Lewis's FTG-11 analysis and older critiques such as George Lewis and Theodore Postol's work, summarised in pieces like the Carnegie Endowment's "Persistence of the Missile defense Illusion," make this saturation logic explicit).

Moreover, even modest countermeasures can sharply worsen the defender's problem. Standard penetration aids include lightweight balloon decoys, chaff, and other objects that mimic the infrared or radar signature of a re-entry vehicle during the midcourse phase, as well as anti-simulation techniques that make real warheads deliberately resemble decoys (see the "Countermeasures" and "Criticism" sections of "Missile defense"). Because objects of different masses follow identical trajectories in space, midcourse defenses must discriminate among many apparently similar targets; if they cannot, they are forced either to waste interceptors on decoys or accept a higher probability that real warheads leak through. Adversaries can also employ manoeuvrable re-entry vehicles, electronic warfare, and simple trajectory tweaks to stress tracking and fire-control systems, all of which further erode the reliability of interception (see the RUSI paper just cited and broader technical critiques referenced there).

Finally, the financial cost of U.S. missile defense underscores the poor cost-effectiveness of this posture. A detailed reconstruction by Frank von Hippel and colleagues, "US Expenditures on Ballistic Missile defense Through Fiscal Year 2021," finds that, measured in 2020 dollars, the United States had spent about 280 billion dollars on defense against long-range ballistic missiles through FY2021, and roughly 360 billion dollars on ballistic missile defense more broadly when related programmes are included. These figures are driven by high development, testing, and deployment costs for systems such as GMD, Aegis BMD, THAAD, and their associated sensors and command-and-control. When one considers that this roughly 300-billion-dollar effort has not produced a defense that can reliably stop even a small, sophisticated salvo from a state adversary, it is difficult to view U.S. strategic missile defense as offering more than marginal, probabilistic protection.

This appendix has focused primarily on midcourse interception by the Ground-Based Midcourse defense (GMD) system because it is the only operational component intended to defend the continental United States against intercontinental ballistic missiles. Although the United States formally presents missile defense as a "layered" architecture, with possible engagements in the boost, midcourse, and terminal phases, the systems available for terminal defense — including THAAD, Aegis, and Patriot — are designed primarily for shorter-range regional threats rather than for reliably intercepting ICBM warheads over the U.S. homeland, and the United States also lacks a deployed boost-phase defense (todo add the link from charles l glaser explaining that getting a boost-phase defense especially against big countries would be technically unfeasible ) against North Korean, Russian, or Chinese ICBMs. Midcourse GMD is therefore the most relevant layer for evaluating the residual vulnerability of U.S. AI-relevant infrastructure.

todo add a discussion about the evolution of DEW and also how to counter these evolution with better ablative coating, spinning missiles / not having straight trajectory, saturation effects, attacking other infrastructure as they do not protect everything, attacking on cloudy day, reacting before,

# Bibliography


## 1. AI deterrence & core strategy

- AI Deterrence Is Our Best Option — *AI Frontiers*
- Coercive Cooperation: Forcing Win-Win Outcomes in an AI-Driven Transition — *AI Prospects*
- Superintelligence Strategy: Expert Version — *arXiv* (also at nationalsecurity.ai and superintelligencestrategy.com)

## 2. AI timelines & the intelligence explosion

- When Do Experts Expect AGI to Arrive? — *80,000 Hours* ⚠️ **VERIFY**

- Date of Artificial General Intelligence (forecast) — *Metaculus*
- Thousands of AI Authors on the Future of AI — *arXiv*
- Who Predicted 2023? — *Astral Codex Ten (Scott Alexander)*
- AGI/Singularity: 9,800 Predictions Analyzed — *AIMultiple*
- AI 2027 — *AI Futures Project*
- AI Futures Model — December 2025 Update — *AI Futures Project* ⚠️ **VERIFY**
- Machines of Loving Grace — *Dario Amodei*
- Preparing for the Intelligence Explosion — *Forethought*
- Three Types of Intelligence Explosion — *Forethought*
- Will AI R&D Automation Cause a Software Intelligence Explosion? — *Forethought*
- 2025 AI Index Report: Technical Performance — *Stanford HAI* ⚠️ **VERIFY**
- How AlphaChip Transformed Computer Chip Design — *Google DeepMind*

## 3. AI risk: misalignment, takeover & power concentration

- Statement on AI Risk — *Center for AI Safety*
- *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development* — Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, David Duvenaud
- FAQ on Catastrophic AI Risks — *Yoshua Bengio*
- Agentic Misalignment — *Anthropic*
- AI Control: How to Make Use of Misbehaving AI Agents — *CSET*
- Risks of Extreme Power Concentration — *80,000 Hours*
- AI-Enabled Coups: How a Small Group Could Use AI to Seize Power — *Forethought*
- Tom Davidson on AI-Enabled Human Power Grabs — *80,000 Hours*
- 'The Godfather of AI' Leaves Google and Warns of Danger Ahead (Geoffrey Hinton) — *New York Times* ⚠️ **VERIFY**

## 4. AI governance proposals & institutions

- Intelsat as a Model for International AGI Governance — *Forethought*
- Building a CERN for AI — *Centre for Future Generations*
- Enhancing Global AI Governance: An International Compute Governance Consortium — *Centre for Future Generations* ⚠️ **VERIFY**
- Four Institutions for Governing and Developing Frontier AI — *Future of Life Institute / Belfield* ⚠️ **VERIFY**
- Do We Want an 'IAEA for AI'? — *Lawfare*
- Global Call for AI Red Lines — *LessWrong*
- Building the Pause Button — *PauseAI*
- PauseAI Proposal — *PauseAI*
- A Narrow Path — *A Narrow Path*
- International AI Institutions: A Literature Review — *LawAI* ⚠️ **VERIFY**
- International AI Institutions: A Literature Review (Maas & Villalobos) — *Nature / Humanities & Social Sciences Communications* ⚠️ **VERIFY**
- Governance of AGI (UN CPGA Report) — *BioComm AI* ⚠️ **VERIFY**
- International Atomic Energy Agency — *IAEA*
- The Schuman Declaration — 9 May 1950 — *European Union*
- Note 5: Verification Mechanisms for International AI Coordination — *GPAI Policy Lab*

## 5. Compute governance & verification

- A Secure Chips Agreement (Verification-Based International AI Governance) — *arXiv*
- Verification for International AI Governance — *Oxford Martin AIGI*
- Computing Power and the Governance of Artificial Intelligence — *GovAI* ⚠️ **VERIFY**
- The Role of Compute Thresholds for AI Governance — *LawAI*
- Nuclear Non-Proliferation Is the Wrong Framework for AI Governance — *AI Frontiers* (also republished by CSET)
- China's Drive Toward Self-Reliance in AI: Chips and Large Language Models — *MERICS* ⚠️ **VERIFY**
- GeoCoded Special Report: State of Global AI Compute (2025 Edition) — *Sanchez VC*
- DiLoCo: Distributed Low-Communication Training of Language Models — *arXiv*
- Scaling Laws for DiLoCo — *arXiv*
- Characterizing the Efficiency of Distributed Training — *arXiv* ⚠️ **VERIFY**
- (arXiv 2509.10371) — *arXiv* ⚠️ **VERIFY**
- OpenDiLoCo: Globally Distributed Low-Communication Training — *Prime Intellect* ⚠️ **VERIFY**
- Multi-Node GPU Clusters Explained — *Genesis Cloud* ⚠️ **VERIFY**
- Turbocharge LLM Training Across Long-Haul Data Center Networks — *NVIDIA* ⚠️ **VERIFY**

## 6. Nuclear deterrence, missile defense & escalation

- Artificial Intelligence and the End of Mutual Assured Destruction — *Foreign Affairs*
- The End of MAD? Technological Innovation and the Future of Nuclear Deterrence — *MIT SSP / Charles Glaser*
- Forgoing U.S. Damage-Limitation Against China's Nuclear Weapons — *Belfer Center*
- Escalation Management and Nuclear Employment in Russian Military Strategy — *War on the Rocks*
- Nuclear Weapons (Problem Profile) — *80,000 Hours / EA Forum*
- How Bad Would Nuclear Winter Caused by a US-Russia Nuclear Exchange Be? — *Rethink Priorities*
- Intelligence, Nuclear Breakout, and Trustworthy AI — *MIT CIS* ⚠️ **VERIFY**
- Current US Missile Defense Programs at a Glance — *Arms Control Association*
- How Does Missile Defense Work? — *Union of Concerned Scientists*
- Comments on the Ground-Based Midcourse Missile Defense Test Record — *Union of Concerned Scientists*
- Ground-Based Midcourse Defense (GMD) Resources — *CSIS Missile Threat*
- What Did FTG-11 Actually Prove? — *Mostly Missile Defense*
- The Persistence of the Missile Defense Illusion — *Carnegie Endowment*
- Ballistic Missile Defense — *Bliley*
- Additive Manufacturing: The Next Great Challenge to Missile Defence — *RUSI*
- US Expenditures on Ballistic Missile Defense Through FY2021 — *Princeton SGS*
- Missile Defense Review: BMDS Factsheet — *US Department of Defense* ⚠️ **VERIFY**
- Ground-Based Midcourse Defense (GMD) — FY2010 Report — *US DOT&E* ⚠️ **VERIFY**
- Missile Defense — *Wikipedia*
- Homeland Missile Defense System Takes Out ICBM Threat in Historic Salvo Test — *Defense News* ⚠️ **VERIFY**

- US Carries Out Successful Salvo Interception of ICBM Target — *The Diplomat* ⚠️ **VERIFY**
- N. Korean ICBMs Could Overwhelm US Missile Defense, Report Says — *Dong-A Ilbo* ⚠️ **VERIFY**
- North Korea Displays Enough ICBMs to Overwhelm US Defense System — *Politico* ⚠️ **VERIFY**
- America's Missile Defenses Against North Korea Have Big Problems — *The National Interest* ⚠️ **VERIFY**
- It's Time to Make the Next Generation of America's ICBMs Road-Mobile — *Heritage Foundation*

# 7. Bio / proliferation risks

- Made-to-Order Bioweapon? AI-Designed Toxins Slip Through Safety Checks — *Science*
- AI-Designed Proteins and Biosecurity Screening (research article) — *Science* ⚠️ **VERIFY**
- Researchers Call for Global Discussion About Risks from Mirror Bacteria — *Wyss Institute*
- Mirror Organisms, Nature, and Chirality — *Harvard Magazine* ⚠️ **VERIFY**
- Aum Shinrikyo — *Wikipedia*
- Soviet Biological Weapons Program — *Wikipedia*
- Biopreparat — *Wikipedia*
- Ken Alibek — *Wikipedia*
- Biopreparat (Soviet BW Agency) — *GlobalSecurity.org* ⚠️ **VERIFY**

# 8. Other / background

- Situational Awareness: The Decade Ahead — *Leopold Aschenbrenner*
- Anthropic's Recommendations to OSTP for the US AI Action Plan — *Anthropic*
- Paris AI Summit — *Anthropic*
- OpenAI Charter — *OpenAI*
- Early US Policy Priorities for AGI — *AI Futures Project*
- Europe 2031: Summary — *Europe 2031*
- South Africa Says It Built 6 Atom Bombs — *New York Times*
- Washington Welcomes de Klerk Disclosure But Wants More Details — *New York Times*
- FRUS 1964-68, Vol. XXX, Document 56 — *US State Department (FRUS)* ⚠️ **VERIFY**
- South Africa Nuclear Program Document (sa27) — *National Security Archive* ⚠️ **VERIFY**
- Soviet BW Program Document 13 — *National Security Archive* ⚠️ **VERIFY**
- Center for Global Security Research — *LLNL CGSR* ⚠️ **VERIFY**
- (YouTube video BHKIM1P7ZvM) — *YouTube* ⚠️ **VERIFY**
- (YouTube video Z19UEZHJzAg) — *YouTube* ⚠️ **VERIFY**
- Sam Altman post (Oct 2025) — *X* ⚠️ **VERIFY**