WHITE PAPER

The Threshold Compact

Governing Superintelligence the Way the World Governed the Bomb

A proposal for a private-sector-led, verification-based international authority for the safe and collaborative development of advanced artificial intelligence, co-led by the United States and China.

Logan Reed

Former U.S. Army Captain | Former Intelligence Contractor Supporting U.S. and Partner Special Operations Forces | Technology Modernization and Responsible AI

Boiling Springs, North Carolina | September 2026

The views expressed are the author's alone and do not represent any employer, agency, or client, past or present.

Contents

A Note From the Author

I have spent my career inside institutions that exist because someone, at some point, decided a danger was too large to leave to good intentions. I am a former Army Captain with eleven years of military experience, most of it in operations, including coalition and NATO assignments and deployments in support of the campaign against ISIS and militia groups in Iraq and Syria, where I also took on intelligence work alongside my operational duties. I advised the commander of a combat aviation brigade of twenty-five hundred soldiers. I advised the commander of a Special Purpose Marine Air-Ground Task Force. I planned multinational exercises in Europe. I helped build information-sharing arrangements among allied and partner forces that did not fully trust one another, and I learned firsthand what it takes to get rivals to share what matters: not goodwill, but rules, verification, and a shared understanding of the cost of being wrong. I later returned to Iraq as an intelligence contractor, supporting United States and partner special operations forces with the collection and analysis of the intelligence that drove their operations. That work taught me, more than anything else in my career, the difference between what a source says and what can be corroborated, and how much it depends on getting that difference right. I have held a top secret clearance with access to sensitive compartmented information.

That experience taught me two things that shape every page of this paper. The first is that no serious government trusts another government's word about a weapon. It trusts what it can verify. The second is that when a capability is powerful enough, the people closest to it are usually the first to understand that it needs to be contained and the last to be given the authority to contain it.

I have also worked in the private sector at scale, supporting operations for one of the world's largest technology companies and consulting to federal health and defense organizations on the modernization of enterprise systems. Today I work on technology modernization and the responsible adoption of AI at a professional services company that serves the federal government. That work has included leading a Responsible and Ethical AI subgroup and facilitating an AI strategy workshop with senior executives. I also run a small company of my own. I studied philosophy before I studied anything else. I am not an AI researcher. I am a practitioner who has watched how classified programs, coalition politics, arms control agreements, and large technical enterprises actually function and who has concluded that the way we are approaching the race to superintelligence is going to fail on its own terms.

This paper is my attempt to say plainly what I believe the problem is and to propose something that could actually be built. The views here are mine alone and do not represent any employer, agency, or client, past or present.

Logan Reed
Boiling Springs, North Carolina
September 2026

Executive Summary

The most common argument against slowing the development of artificial superintelligence is also the most honest one: if the United States slows down, its adversaries will not. I accept that argument. It is true. It is also the reason the current path cannot hold, because it applies with equal force to every other nation in the race. Every capable state believes it cannot afford to be second, and so every capable state is sprinting toward a technology that its own leading laboratories describe as potentially the most dangerous humanity has ever built.

This is not a new situation. We have been here before, with a different technology, and we know what the answer looked like. When nuclear weapons arrived, the world did not rely on promises. It built institutions, inspection systems, material accounting, and treaties with consequences so that commitments could be checked even where trust was absent. The result has been imperfect and has suffered serious setbacks. But more than eighty years after the first use of nuclear weapons, nine states are assessed to possess them rather than the fifteen to twenty-five that President Kennedy feared, and the weapons have not been used in war since 1945.[1] That record has more than one cause, deterrence and diplomacy among them. Verification is the part of it that rivals could build together, and it is the part we can build again.

My position is that artificial superintelligence should be viewed, treated, handled, and governed as a strategic weapon of the same class. Not because it is only a weapon; its potential to lift humanity out of scarcity is real, and I want us to reach it. But because a technology that can confer decisive and possibly permanent advantage on whoever reaches it first will be pursued as a weapon whether or not we call it one, and history offers only one model for keeping such a technology under control among rivals who do not trust each other.

The deeper concern is that reaching superintelligence first does not mean being able to control it. At this stage, reliable human control over artificial superintelligence remains unproven, regardless of which laboratory or country develops it. Good intentions and national allegiance do not resolve that uncertainty. I believe the potential consequences of losing control are too serious for humanity to cross that threshold on the promise that we will figure out the safeguards afterward. The nuclear experience shows how rivals who distrust one another can still verify, restrain, and cooperate. For superintelligence, those arrangements must do something the nuclear ones never had to: govern whether, and under what conditions, we cross the threshold at all.

WHAT THIS PAPER PROPOSES

The Threshold Compact is a binding charter among the world's frontier AI laboratories. The Threshold Authority is the standing body it creates: led and operated by private-sector professionals, because that is where the knowledge lives, and overseen by governments through an Oversight Council and national law, because that is where legitimacy lives. It is co-chaired by American and Chinese

[1] SIPRI Yearbook 2026, "World nuclear forces"; John F. Kennedy, News Conference 52, March 21, 1963, on the possibility of "fifteen or twenty or twenty-five" nuclear powers by the 1970s.

> industry members. Every organization capable of training a frontier model holds a seat and bears the same obligations: to declare frontier training runs, submit models to independent evaluation, pause at defined capability tripwires, admit inspectors, register compute, pool safety research, and fund the enterprise. The Authority reports on a fixed cadence to every member government and to the public, with the same core findings in both. A graduated ladder of penalties ends in lawful restrictions on the chips, services, and capital that frontier development requires. None of this exists to stop the technology. It exists to make sure we reach what the technology promises: cures for diseases that have killed our children for all of history, the end of material scarcity, and, through an Abundance Program written into the charter, a share of those gains for every nation rather than a prize for whoever wins the race. Verification is the price. Abundance is the prize. And because the technology is moving faster than legislatures, the paper recommends a fast path that stands the institution up in twelve to eighteen months on executive action and existing authorities, with statute and treaty following rather than leading.

The paper is organized in five parts. Part One describes the problem and why the current approach is failing. Part Two reasons from first principles to the nuclear frame, including where the analogy breaks down, and states the control problem plainly. Part Three sets out the design of the Compact and the Authority in detail: governance, oversight, membership, verification, reporting, funding, and enforcement, and is candid about what the charter cannot do by itself. Part Four explains why each of the necessary parties has reason to join and lays out two paths for standing the institution up, one fast and one deliberate. Part Five answers the strongest objections. Appendices contain an outline of the charter, the reasoning behind the name, and the sources.

I have tried to write this so that a head of state, a legislator, a chief executive, and an ordinary citizen can each read it in one sitting and know exactly what is being asked of them. The ask is at the end.

Part One: A Race No One Can Win by Running Faster

Where we are

As I write this in September 2026, reporting describes preparations for a new round of talks between the United States and China on AI safety, ahead of a planned meeting of the two heads of state later this month. The two governments held an earlier intergovernmental dialogue on AI risk and safety in Geneva in May 2024.[2] The talks are a good sign. They are also a small one. On August 31, commentary from a Chinese state-media-affiliated account laid out two conditions: first, that the definition of "AI safety" be set jointly rather than by Washington alone, and second, that American AI companies be subject to the same safety and audit requirements sought for Chinese firms. The commentary singled out a leading American laboratory by name and accused the United States of trying to export its own safety boundaries as global norms.[3] State-affiliated commentary is evidence of an argument circulating in Beijing, not an official negotiating position. It is still worth reading closely.

It is easy to read those conditions cynically. Read another way, they are an admission. Beijing is saying that it will not accept an arrangement it did not help design and that does not apply equally. That is the same thing every American negotiator has said about every arms control proposal since 1946. It is the sound of two rivals describing, from opposite sides, the shape of the only agreement either would ever sign.

Meanwhile the institutional landscape is fragmenting rather than converging. The summer of 2026 saw a G7 summit that seated AI executives alongside heads of government, the first session of a United Nations Global Dialogue on AI Governance open to all 193 member states, and, in Shanghai, the signing by 29 countries of an agreement to establish a World Artificial Intelligence Cooperation Organization under Chinese leadership. The United States has pursued an adoption-first posture, with executive actions favoring voluntary pre-release access and a coalition effort to secure semiconductor supply chains among allies.[4] The former network of national AI safety institutes now operates as the International Network for Advanced AI Measurement, Evaluation and Science; the Bletchley Declaration of 2023 was signed by both Washington and Beijing; and an International AI Safety Report, chaired by Yoshua Bengio and written by more than one hundred experts with an advisory panel nominated by more than thirty countries and international organizations, was published in February.[5] Not one of these arrangements can compel a single laboratory to disclose a

[2] Reuters, "U.S., China gear up for mid-September AI safety talks," September 4, 2026; Foreign Policy, August 31, 2026; Reuters, "U.S., China hold AI risk and safety talks, White House says," May 15, 2024. September events are reported plans as of the research cutoff.

[3] Tech Times, "China Tells US: Agree on What AI Safety Means or September Talks Cannot Proceed," September 3, 2026, reporting commentary published August 31, 2026 by Yuyuantantian, an account affiliated with China Central Television.

[4] Tanner, Kerry, Tabassi, Renda, and Wyckoff, "A summer of AI summits reveals a widening US-China divide," Brookings Institution, August 11, 2026; Elysee, "G7 Summit in Evian: Day Three," June 17, 2026; Xinhua, July 16, 2026, on the WAICO signing; The White House, executive order on advanced AI innovation and security, June 2026.

single training run. None amounts to an inspectorate with access to both American and Chinese frontier labs.

Two numbers presented at the launch of the United Nations scientific panel's preliminary report show how concentrated the infrastructure is. In the dataset of the five hundred largest known public and private AI compute clusters, roughly 75 percent of the capacity is located in the United States and roughly 15 percent in China.[6] Everyone else shares the remaining tenth. Those are shares of a surveyed set of large clusters, not a census of all AI compute, and location is not the same as ownership or government control. They are still the clearest picture we have, and they describe a two-power problem with a long tail, which is precisely what the nuclear problem was in 1950.

The builders already know

It would be one thing if the case for international control came only from outsiders. It does not. It comes, in their own words, from the people building the systems.

In May 2023 the three leaders of OpenAI wrote that "we are likely to eventually need something like an IAEA for superintelligence efforts; any effort above a certain capability (or resources like compute) threshold will need to be subject to an international authority that can inspect systems, require audits, test for compliance with safety standards, place restrictions on degrees of deployment and levels of security, etc." They added that "tracking compute and energy usage could go a long way, and give us some hope this idea could actually be implementable."[7] That same month, the chief executives of OpenAI, Anthropic, and Google DeepMind joined hundreds of researchers in signing a one-sentence statement: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[8]

On July 14, 2026, Google DeepMind chief executive Demis Hassabis published a proposal for a Frontier AI Standards Body modeled on FINRA, the industry-funded regulator of American securities brokers that operates under the oversight of the Securities and Exchange Commission. Under his proposal, "frontier-class" models would at first be shared voluntarily before release, evaluation protocols would be updated as often as quarterly, and assessments could later become mandatory before deployment in the United States.[9] In September, reporting described working-level discussions among OpenAI, Google, and Anthropic about forming such a body.[10]

---

[5] UK Department for Science, Innovation and Technology, December 9, 2025, on the renamed International Network for Advanced AI Measurement, Evaluation and Science; UK Government, "The Bletchley Declaration," November 1, 2023; Bengio et al., International AI Safety Report 2026, February 3, 2026, description of contributors.

[6] United Nations, transcript of the launch of the Independent International Scientific Panel on AI preliminary report, 2026; Brookings, August 11, 2026. The figures concern the 500 largest known public and private clusters, not all AI compute.

[7] Sam Altman, Greg Brockman, and Ilya Sutskever, "Governance of Superintelligence," OpenAI, May 22, 2023.

[8] Center for AI Safety, "Statement on AI Risk," May 30, 2023, with signatories.

[9] Demis Hassabis, "A Framework for Frontier AI and the Dawning of a New Age," July 14, 2026. On the model, see FINRA, "About FINRA": a private self-regulatory organization funded by member fees and supervised by the Securities and Exchange Commission.

[10] The Information, September 2026, with follow-on reporting by Reuters; discussions were reported to have begun in July 2026. Reporting of discussions is not evidence of an agreement.

Most telling of all is what happened when one company tried to hold the line alone. In 2023 Anthropic tied the training and deployment of more capable models to specified safeguards. On February 24, 2026, it replaced that structure with a revised Responsible Scaling Policy that separates what the company will do on its own from what it recommends for the industry as a whole. Its reasoning was blunt. Some of the strongest safeguards, it wrote, "might prove outright impossible to implement without collective action," and "if one AI developer paused development to implement safety measures while others moved forward training and deploying AI systems without strong mitigations, that could result in a world that is less safe."[11] That is not a company abandoning safety. It is a company discovering that safety cannot be achieved unilaterally, and saying so.

Its chief executive has now said the same thing in his own voice. In January, Dario Amodei warned that a combination of intelligence, agency, coherence, and poor controllability could create existential danger.[12] In September he went further, in an essay titled "We Must Pace the Frontier": "We must slow the pace at which we improve the capabilities of AI models." He committed his company to embedded third-party evaluators, and proposed that frontier labs in democracies agree common limits on the rate of capability advancement, with governments mediating, and that democratic governments then attempt coordination with authoritarian ones. He also stated the constraint that makes such coordination necessary: slow down too far alone, and rival projects "will pull ahead, creating significant national security risk."[13]

On September 8, 2026, a researcher named Jacob Coxon announced his resignation from Anthropic in a post that was viewed tens of millions of times within a day. He had spent three years in pretraining research at OpenAI and Anthropic, and he wrote that neither company was acting responsibly: "They are racing straight to self-improving superintelligence and gambling with our lives." In interviews he argued that coordination among laboratories, backed by pacing agreements and legislation, was the way out.[14] His post did not begin AI safety diplomacy; the talks and institutions described above predate it. Nor does a viral post establish when superintelligence will arrive or how likely catastrophe is. What matters here is the problem he named from the inside: competitive pressure pushes even safety-minded companies toward risks they would rather not take.

Read those statements together. The builders believe the danger is real. They have said, in writing, that an international authority with inspection powers will be needed. They have begun, on their own, to build a standards body. Their chief executives are now calling for pacing agreements and outside evaluators. And they have demonstrated, in the most credible way possible, that unilateral restraint collapses under competitive pressure. Their positions differ, and none of them has endorsed this paper's design. But

[11] Anthropic, "Responsible Scaling Policy, Version 3.0," effective February 24, 2026, and Anthropic's accompanying explanation of the revision, February 24, 2026. Both quotations are from Anthropic's own text.

[12] Dario Amodei, "The Adolescence of Technology," January 2026, section on autonomy risk; reported in Euronews, January 28, 2026.

[13] Dario Amodei, "We Must Pace the Frontier," September 2026. Quotations are from the essay.

[14] Jacob Coxon, X, September 8, 2026; Kaitlyn Huamani, Associated Press, September 9, 2026; TechCrunch, September 9, 2026; TIME, September 9, 2026. View counts were reported at publication and are not a count of unique readers.

what the industry is reaching for is missing exactly two things: the other superpower, and consequences. This paper supplies both.

The argument I accept

The strongest case against any pause, slowdown, or restraint in frontier AI development runs like this. Superintelligence, if it can be built, will be built. The nation or coalition that builds it first will hold an advantage in economics, intelligence, cyber operations, and warfare that may never be overcome. The United States cannot verify what China is doing inside its laboratories, and China cannot verify what the United States is doing inside its own. Therefore any unilateral restraint by the United States is not caution; it is surrender of the decisive advantage to a strategic competitor whose values are not our own. Better to win the race and then set the terms.

I accept every premise of that argument except the conclusion, and I accept them because I have seen how the world actually works from inside the institutions that exist to watch adversaries. I have sat in coalition headquarters where allies fighting the same enemy would not share a target list without a written agreement, the proper clearance, and a way to confirm it was honored. Nations do not take one another's word. They should not. A leader who slowed his country's most important strategic program on the strength of a rival's promise would be failing in his first duty. If restraint requires trust, restraint is impossible.

But look at what the argument proves. It proves that restraint by agreement is impossible. It does not prove that restraint by verification is impossible, and those are entirely different things. No one trusted the Soviet Union in 1970, when the Non-Proliferation Treaty entered into force, and no one trusted the United States either. The treaty's inspections fell principally on states without weapons; the superpowers verified each other through separate arms control agreements with their own inspection arrangements.[15] The lesson is narrower than the popular version and still decisive: rivals who trusted one another in nothing accepted defined, limited checks on the things that mattered most. They did not agree to trust one another. They agreed to let one another look.

Why running faster does not end the race

The "win the race" strategy has a second flaw that the intelligence world understands better than most. Winning presumes a finish line. There is none. If the United States reaches superintelligence first, China does not stop; it accelerates, with every resource of the state behind it, and in a world where model weights can be exfiltrated on a hard drive and capabilities can be distilled from a rival's outputs, the durability of any lead is measured in months. The first mover's "decisive advantage" would have to be used, immediately and coercively, to prevent the second mover from catching up. I do not believe the American people want their government to be in that position, and I know they do not want to be on the receiving end of it. Every serious planner on both sides knows this, which is why both sides are racing: not toward victory, but away from defeat.

[15] Treaty on the Non-Proliferation of Nuclear Weapons, opened for signature July 1, 1968, entered into force March 5, 1970, Article III; IAEA, "Safeguards agreements." Nuclear-weapon states' voluntary-offer agreements cover selected peaceful facilities, not their arsenals.

That is a security dilemma, and security dilemmas do not resolve on their own. They ease when the parties build something that lets each of them see enough of the other to stop running. Verification does not remove conflicting interests. It removes the blindness that turns conflicting interests into a sprint. The remainder of this paper is about what that something should be.

Part Two: The Right Frame

Starting from first principles

Before reaching for any historical analogy, I want to build the argument from first principles, because an analogy is only as good as the reasoning underneath it, and because the people this paper is addressed to are entitled to see that reasoning rather than take it on faith. Five propositions follow, each from the one before.

First, a technology that confers decisive advantage will be pursued by every actor capable of pursuing it. This is not cynicism; it is the first duty of any government to its people and of any company to its shareholders, and no appeal to restraint has ever suspended it for long.

Second, when several rivals pursue such a technology at once, each one's fear of the others sets the pace. No participant can slow down unless it knows the others have, so the race runs at the speed of the most fearful participant, not the most prudent.

Third, knowledge that the others have slowed cannot come from their word. Rivals do not trust one another, and they are right not to, because the payoff for cheating grows with the stakes. Agreement without verification is therefore worth least exactly when it matters most.

Fourth, verification requires evidence that can be checked independently of the party being verified. A promise cannot be verified. A file cannot easily be verified. A physical input that is scarce, traceable, and necessary to the enterprise can be, and so can records, evaluations, and technical audits when inspectors have access to them. No single method guarantees that a concealed activity will be found. Several together raise the cost of concealment until it is no longer worth paying.

Fifth, it follows that the only governance of a decisive technology that can work among rivals is one built on the independent observation of its scarce physical inputs and its declared activities, backed by consequences severe enough to outweigh the gain from cheating, and designed so that the benefits of participation exceed the benefits of racing.

Everything in this paper is an application of those five propositions. The nuclear order is the one time in history that humanity has applied them at scale, which is why the comparison is worth making. Where the comparison holds, we should borrow. Where it breaks, we should go back to the propositions and reason forward again.

Why superintelligence is a strategic weapon

I use the word "weapon" deliberately, and I want to be exact about why. A technology belongs in the class of strategic weapons when four conditions hold. It confers decisive advantage on whoever possesses it. It cannot be un-invented. Its development requires concentrated, scarce, traceable inputs. And its misuse, whether deliberate or accidental, can cause harm at a scale that no single nation can absorb or contain. Nuclear fission met all four. Advanced artificial intelligence meets all four today, and the degree to which future systems will meet them is the reason the comparison deserves attention rather than a reason to defer it.

The decisive advantage is not speculative; it is the stated reason both superpowers are racing. Irreversibility is obvious: the mathematics of training large models is published, and the knowledge cannot be recalled. The concentrated inputs are the most important condition for our purposes and the least discussed. The largest training programs use very large clusters of the most advanced semiconductors on earth, produced by a handful of firms in a handful of countries, installed in facilities that draw power on the scale of a small city and are visible from orbit. Compute is to AI what fissile material was to the bomb: the one ingredient that is hard to make, hard to hide, and possible to count. The analogy has limits, and I will state them: chips can be reused for many tasks, and a count of hardware does not tell you what a model can do. But the largest training runs likely passed ten to the twenty-sixth floating-point operations in 2025,[16] and a body of scholarship now argues that compute is the most governable feature of the whole enterprise, because it is detectable, excludable, and quantifiable in a way that algorithms and data are not.[17] The chief executive of Anthropic put it more bluntly: "Chips and chip-making tools are the single greatest bottleneck to powerful AI."[18] Finally, the scale of harm is the subject of the International AI Safety Report, of a 2024 paper in Science by twenty-five of the field's most senior researchers warning of risks including "large-scale loss of life" and "the marginalization or extinction of humanity,"[19] and of the published safety frameworks of the frontier labs themselves, which now evaluate their own models for biological, cyber, autonomy, and loss-of-control risks.[20] When the builders of a thing describe it in those terms, the burden of proof shifts to anyone who says it should be governed like ordinary software.

The problem no one has solved

There is a difference between the nuclear problem and this one that I want to state before the historical comparison, because it changes what the institution must do.

A nuclear weapon does what its owner tells it to do. The danger was always in the owner. Artificial superintelligence, if it is built, is a system whose capabilities would by definition exceed our own across most domains and whose behavior its creators cannot fully predict. Whether such a system would reliably do what its owners intend is an open scientific question. The International AI Safety Report describes scenarios in which advanced systems come to "operate outside of anyone's control," notes early signs of such behaviors in current systems, and records that expert views on their likelihood vary widely.[21] The chief executive of a leading lab has described poor controllability as an

[16] International AI Safety Report 2026, section 1.3 and Figure 1.6. The statement concerns estimated training compute, not a safety or superintelligence threshold.

[17] Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, et al., "Computing Power and the Governance of Artificial Intelligence," February 13, 2024.

[18] Dario Amodei, "The Adolescence of Technology," January 2026, discussion of chip exports. The bottleneck statement is the author's assessment.

[19] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Stuart Russell, et al., "Managing extreme AI risks amid rapid progress," Science 384, no. 6698 (May 2024). The cited harms are risks, not established outcomes.

[20] Anthropic, "Responsible Scaling Policy"; OpenAI, "Our updated Preparedness Framework," April 15, 2025; Google DeepMind, "Strengthening our Frontier Safety Framework," September 22, 2025, updated April 17, 2026. Their categories and commitments differ.

[21] International AI Safety Report 2026, discussion of loss-of-control scenarios, which records early signs of such behaviors in current systems at low capability levels and wide variation in expert views on likelihood.

ingredient of existential danger.[22] A researcher who left that lab this month described the industry as "gambling with our lives."[23] Those are not the words of outsiders.

So the deeper concern is this: reaching superintelligence first does not mean being able to control it. Reliable human control over artificial superintelligence remains unproven, regardless of which laboratory or country develops it. Good intentions and national allegiance do not resolve that uncertainty. An American superintelligence that no one controls is not an American victory. I believe the potential consequences of losing control are too serious for humanity to cross that threshold on the promise that we will figure out the safeguards afterward.

This has two consequences for the design. First, the Compact cannot only govern who crosses the threshold and how openly. It must govern whether, and under what conditions, anyone crosses it at all. That is why the design below includes mandatory pauses at defined capability tripwires, a Board review before deployment, and a Safety Research Commons in which the hardest technical problems are worked jointly rather than in secret. Second, the Compact must be honest about what it is not. It is not a solution to the control problem. No institution can legislate a scientific result. What it can do is buy time, pool the effort, and make sure that if the threshold is approached, it is approached deliberately, together, and in the open, with the evidence of control examined by people who do not work for the developer.

Where the analogy breaks, and what that means

I do not want to oversell the comparison. The differences matter, and every one of them argues for a stronger framework built earlier, not a weaker one built later.

First, there is no mutual assured destruction with superintelligence. The nuclear peace held in part because each side knew that a first strike meant its own annihilation. A decisive AI advantage may carry no such penalty for the one who holds it and, as the previous section argued, may not be an advantage at all. Deterrence, the backstop that made nuclear arms control survivable when it failed, may not exist here. Some serious strategists have proposed building an AI deterrence doctrine anyway, under which states would threaten to sabotage any rival's destabilizing AI project.[24] I take the proposal seriously, and I note that it depends on each side being able to see the other's projects clearly enough to know what to threaten. That is a verification problem, and it is the problem this paper addresses. Deterrence without observation is a bluff. Observation without deterrence is at least a warning. We need both, and observation comes first.

Second, software copies itself. A warhead is a physical object that can be counted and dismantled. A model is a file. Once weights leave a secure environment they can be everywhere. This means that the point of control must be upstream, at the training run and the hardware, and that it must be paired with model security, deployment controls,

[22] Dario Amodei, "The Adolescence of Technology," January 2026, section on autonomy risk; reported in Euronews, January 28, 2026.

[23] Jacob Coxon, X, September 8, 2026; Kaitlyn Huamani, Associated Press, September 9, 2026; TechCrunch, September 9, 2026; TIME, September 9, 2026. View counts were reported at publication and are not a count of unique readers.

[24] Dan Hendrycks, Eric Schmidt, and Alexandr Wang, "Superintelligence Strategy: Expert Version," March 2025, proposing Mutual Assured AI Malfunction. The observation requirement is this paper's analysis.

and oversight of later modifications. Training declarations alone cannot govern copies that have already escaped.

Third, the benefits are of a different order. Nuclear energy is useful; superintelligence could end material scarcity, cure diseases that have killed our children for all of history, and lift billions of people into dignity. Governance that merely suppresses the technology would be a moral failure as grave as governance that ignores its dangers. The framework must be designed to accelerate the shared benefits while controlling the shared risks, and that dual purpose must be written into its charter, not bolted on afterward.

Fourth, the actors are not only states. The frontier is being built by private companies, some of which command more relevant expertise and capital than most governments. Any arrangement that seats only governments will be negotiating with the wrong people. Any arrangement that seats only companies will lack the legitimacy and the legal force to make its decisions stick. It has to be both, in the right roles.

What the nuclear era teaches

Between 1945 and 1970 the world ran an unplanned experiment in governing a decisive technology among rivals who despised one another. The results are on the record, and they point in one direction.

The Acheson-Lilienthal Report of 1946 proposed placing the mining and use of fissile material, and every facility capable of producing weapons, under an international authority. Bernard Baruch's version, presented to the United Nations in June 1946, added enforcement provisions and demanded that the Security Council veto not apply to sanctions for violations. It failed.[25] The dispute ran through the sequencing of disarmament and inspection, the American monopoly, and the Soviet refusal to surrender its veto. Three years later the Soviets tested their own bomb. The lesson is not that international control is naive. The lesson is that a proposal framed as one side's surrender to the other's rules will be refused, and the refusing party will simply build. Any AI arrangement that reads as American rules for Chinese laboratories, or the reverse, will meet the Baruch Plan's fate. Both conditions in the August commentary from Beijing are, in effect, a demand that we not repeat 1946.

The International Atomic Energy Agency, established in 1957 after President Eisenhower's "Atoms for Peace" address, succeeded where Baruch failed because it changed the deal. Instead of demanding that nations forgo the technology, it offered them the benefits of the technology in exchange for accepting safeguards and inspections. Access was the incentive; verification was the price.[26] The Non-Proliferation Treaty later combined non-proliferation obligations, peaceful cooperation, and a commitment to pursue disarmament. It did not abolish nuclear weapons and it did not impose identical obligations on states with and without them.[27] Within those limits it is the most successful arms control arrangement in history, and it worked because the thing being inspected was physical and countable.

[25] U.S. Department of State, Office of the Historian, "The Acheson-Lilienthal and Baruch Plans, 1946." Acheson-Lilienthal Report: March 16, 1946; Baruch presentation: June 14, 1946; first Soviet nuclear test: August 29, 1949.

[26] Dwight D. Eisenhower, "Atoms for Peace," December 8, 1953; Statute of the International Atomic Energy Agency, adopted October 23, 1956, entered into force July 29, 1957, especially Articles II and III.

[27] Treaty on the Non-Proliferation of Nuclear Weapons, especially Articles III, IV, and VI; International Atomic Energy Agency, "Safeguards agreements."

CERN, founded by convention in 1954, proved that rival nations can do frontier science together when the science is too expensive and too important for any of them to do alone. ITER, the international fusion project agreed in 2006, proved the same among the United States, China, Russia, the European Union, India, Japan, and South Korea, even as those governments fell out on nearly every other subject. Both are intergovernmental projects with professional scientific staffs rather than regulators, and neither is a precedent for sharing weapons technology.[28] What they show is narrower and still essential: scientists and engineers from adversary nations have worked side by side on the most sensitive technologies of their age for seventy years, under agreements that governments ratified but professionals ran.

Finally, the Nunn-Lugar Cooperative Threat Reduction program of the 1990s showed that even a collapsing adversary's most dangerous materials can be secured when the incentives are right and the money is on the table.[29] Practical cooperation on containment is possible between rivals who agree on nothing else.

Taken together: share the benefit, verify the inputs, seat the professionals, and give governments oversight and the final word on enforcement. That is the template.

A word about what we owe one another

I hold this position for reasons that go beyond strategy, and I think it is dishonest to hide them. I believe every human being possesses a dignity that does not depend on his usefulness, his intelligence, or his nation. I believe that those who hold great power hold it as stewards rather than owners, and that the measure of a society is how it treats the people who have the least say in the decisions that shape their lives. A technology that could remake the conditions of human life is being developed by a few thousand people in a few dozen buildings, and the eight billion people who will live with the result have no seat, no vote, and no window. That is not a stable arrangement, and it is not a just one.

The moral tradition I come from also teaches prudence, which is not timidity but the discipline of matching means to ends and refusing to gamble what one has no right to lose. It teaches that decisions should be made at the lowest level competent to make them well, with higher authority stepping in only where lower levels cannot. Both principles shaped what follows. The Authority I propose is run by the people who actually understand the technology, because they are the competent level. It answers to governments and to the public, because the stakes exceed what any private body has the right to decide alone. And it is built to share the fruits of this technology widely, because a world in which superintelligence enriches a few nations and leaves the rest as spectators would be a betrayal of what the technology could be.

[28] CERN Convention, entered into force September 29, 1954; ITER Agreement, signed November 21, 2006, entered into force October 24, 2007. ITER's members are China, the European Union, India, Japan, Korea, Russia, and the United States.

[29] Soviet Nuclear Threat Reduction Act of 1991, Title II of Public Law 102-228, December 12, 1991; National Research Council, Global Security Engagement: A New Model for Cooperative Threat Reduction, 2009.

Part Three: The Threshold Compact and the Threshold Authority

The name

I call the agreement the Threshold Compact because the word names the thing we are actually worried about. There is a line, or a series of lines, past which artificial systems stop being tools and start being agents whose capabilities exceed our ability to predict, correct, or contain them. No one knows exactly where those lines are, and the word does not claim a scientific consensus about a fixed boundary. Everyone who builds these systems agrees the lines exist. The purpose of the Compact is to ensure that no one crosses them alone, in secret, or without the rest of the world knowing it has happened, and that no one crosses them at all until the evidence of control has been examined by people who do not work for the developer. The word also carries the sense of a doorway: on the other side is either the greatest expansion of human flourishing in history or the end of human agency, and we would like to walk through it deliberately. "Compact" rather than "treaty" because the signatories are companies before they are states, and the word carries the sense of mutual obligation freely undertaken. The body the Compact establishes is the Threshold Authority.

Eight founding principles

Everything in the design follows from eight principles, which are themselves derived from the first principles set out at the start of Part Two rather than borrowed from any existing institution. They are the non-negotiables; the mechanisms that implement them can and should be argued over. The organs, thresholds, terms of office, budget, and timetable that follow are design proposals, not descriptions of an existing body. Two definitions govern the rest: superintelligence here means a system substantially exceeding human capability across most cognitive tasks; a frontier model is one that meets the Compact's published compute or capability criteria.

- **Led by the private sector, overseen by governments.** The Authority is staffed, governed, and funded by the organizations that build frontier systems and the professionals they employ, because that is where the knowledge lives and because they can move at the speed the problem demands. Governments do not run it. They oversee it through a standing Oversight Council, give it force through national law, receive its reports, and hold the final word on enforcement. This is the division of labor securities regulation has used for decades: the industry body inspects and disciplines its members; the public authority supervises the industry body.
- **Co-chaired by American and Chinese members.** The two nations that host nine-tenths of the world's large AI clusters supply the two permanent Co-Chairs, elected by their member delegations. Neither can outvote the other on matters of charter. An arrangement led by one would be refused by the other; one led by neither would be ignored by both.
- **Universal at the frontier.** Every laboratory on earth capable of training a frontier model holds a seat and bears the same obligations. There are no observers at this

table and no exemptions for national champions, state-owned enterprises, or defense programs.

- **Verification over trust.** No obligation under the Compact rests on a member's word. Every obligation is attached to a mechanism by which the Inspectorate can assess compliance, anchored in the physical inputs to frontier development and in access to records and models where hardware alone cannot answer the question.
- **Transparency by default.** The Authority publishes what it knows on a fixed schedule. Information is withheld from the public only where disclosure would itself create danger or where law requires protection, and even then the withholding is disclosed and the material is delivered to governments under the access rules set out below.
- **Shared safety, proprietary capability.** Members pool their work on alignment, evaluation, security, and containment under agreed security and confidentiality rules. They keep their commercial products, architectures, and training data to themselves, subject only to the secure access needed for verification. The line between the two is drawn in the charter and policed by the Inspectorate.
- **Control before crossing.** No member deploys, and no member continues training, a system that has tripped a defined capability tripwire until the Board has reviewed the developer's evidence that the system remains under reliable human control. The burden is on the developer. The default is the pause.
- **Consequences that bite.** Violation of the Compact leads, by a published and graduated ladder, to penalties that end in lawful restrictions on access to the compute, capital, and markets that frontier development requires. The ladder is enforced by members first and by anchor governments last.

Structure of the Authority

The Authority has eight organs. I describe each briefly; a fuller treatment belongs in the charter.

The Board of Principals is the governing body. Each member organization seats one Principal, who must be its chief executive or a direct report with authority to bind the organization. The Board meets quarterly in person, alternating between host cities in the United States and China with periodic sessions elsewhere. It adopts standards, approves the annual report, and votes on enforcement actions. Ordinary decisions pass by two-thirds of eligible voting Principals, except where the enforcement ladder sets another threshold. Charter amendments require two-thirds plus the assent of both Co-Chairs.

The Co-Chairs are two individuals, one drawn from the American members and one from the Chinese members, elected by their respective delegations for staggered three-year terms. They set the agenda, represent the Authority to governments and the public, and must each assent to charter amendments and to the appointment of the Inspector General. They hold no veto over enforcement votes, so that neither can shield its own nationals.

The Oversight Council is the government tier. Each anchor government designates one senior official to sit on it. The Council does not vote on standards, membership, or enforcement; those belong to the Board. It receives every report and every enforcement

referral, may compel the Board, the Inspector General, or any member to answer questions on the record, commissions independent audits of the Inspectorate, may order a review of any Board decision, and may recommend, by a vote of its members, that a matter be taken up by national authorities. Its proceedings are published except where law requires otherwise. The Council is what makes "overseen by governments" a fact rather than a phrase. It is also the body through which governments can satisfy themselves, without running the institution, that the institution is running as promised.

The Independent Inspectorate is the heart of the institution. It is led by an Inspector General appointed by the Board with the assent of both Co-Chairs for a single seven-year term and removable only for cause by three-quarters vote, with notice to the Oversight Council. Its staff are technical professionals employed directly by the Authority, drawn from member nations in rough proportion to their compute share, security-vetted by a process both Co-Chairs accept, and barred from returning to member employment for three years after service. The Inspectorate conducts declared and unannounced inspections, audits compute declarations, runs independent evaluations of models that cross reporting thresholds, receives whistleblower reports, and refers violations to the Board.

The Technical Secretariat is the permanent staff. It maintains the standards, operates the Compute Registry described below, coordinates the Safety Research Commons, drafts the reports, and administers the budget. It is led by a Director General appointed for a five-year term.

The Safety Research Commons is the collaborative program. Every member commits a fixed share of its frontier research capacity, in people and in compute, to work on alignment, interpretability, evaluation, model security, and containment whose results are shared promptly among all members under the Compact's security and confidentiality rules and published to the world on a delay set by the Board. The Commons is the CERN inside the IAEA: the place where rivals do the science that protects all of them, and the place where the control problem described in Part Two is worked in the open.

The National Liaison Offices are the interface with governments below the Council level. Each member government designates one body, whether a ministry, an agency, or a legislative committee, to receive the Authority's restricted annex and to serve as the point of contact for enforcement referrals. The reporting schedule and core format are fixed by the charter; implementing agreements establish the legal authority, confidentiality protections, and access rights that reporting requires.

The Council on Human Dignity is an advisory body of twelve to fifteen persons who are not employed by any member and have no financial interest in the industry: ethicists, jurists, physicians, and representatives of the world's major religious and philosophical traditions. It reviews the Authority's standards and reports for their effect on human persons and human communities, and its dissents are published in full alongside the reports. It cannot block a decision. It can make sure no decision is taken without the question being asked.

Who sits at the table

Membership is defined by capability rather than nationality or ownership. Any organization that has trained, or has the demonstrated capacity to train, a model meeting a published compute or capability threshold set by the Board is a Frontier Lab with a seat, a vote, and the full set of obligations. Cloud providers and semiconductor firms whose capacity is necessary to frontier training are Infrastructure Members with a seat, a vote on standards affecting them, and the obligations of the Compute Registry. Organizations approaching the threshold are Associate Members with reporting obligations and a non-voting seat. Governments participate through the Oversight Council and the anchor statutes, not as voting members. The table below shows the proposed composition.

Category	Who	Obligations	Vote
Frontier Labs	Every organization, public or private, state-owned or independent, that meets the published compute or capability criteria. At founding this is roughly a dozen organizations in the United States and China and a handful in Europe, the Gulf, and Asia; the founding roster would be set by verified assessment.	Full: declarations, inspections, Commons contribution, levy, incident reporting, pre-deployment evaluation, tripwire pause	Full
Infrastructure Members	The semiconductor designers and fabricators, lithography suppliers, and cloud operators whose capacity is necessary to frontier training.	Compute Registry participation, know-your-customer rules, lawful implementation of compute restrictions when ordered	On standards affecting infrastructure
Associate Members	Organizations within one order of magnitude of the frontier threshold.	Declarations and incident reporting; reduced levy	Non-voting seat
Anchor Governments	Governments of the states in which Frontier Labs and Infrastructure Members are domiciled.	Seat the Oversight Council; enact anchor statutes; receive the restricted annex; act on enforcement referrals	Oversight Council only; none on the Board

State-owned and defense-affiliated laboratories are Frontier Labs if they meet the capability test. I know this is the hardest point in the proposal, and I address it directly in Part Five. I will say here only that an arrangement which exempts military programs is an arrangement that governs the part of the problem that was never the danger.

What must be reported

The Compact establishes three kinds of reporting, each tied to a measurable trigger.

Training declarations. Before beginning any training run above the frontier compute threshold, a member files a declaration with the Secretariat stating the compute to be used, the facilities, the intended capabilities, and the safety plan. The threshold is set by

the Board, is public, and is reviewed as efficiency and capabilities change. Declarations are not requests for permission. They are notice, so that no frontier run happens in the dark. A compute threshold alone is not enough: the charter pairs it with capability-based triggers, so that a more efficient model, a fine-tuned derivative, or a system given very large inference resources does not fall outside the framework because its original training run was small.

Capability evaluations. Before deploying any model above the threshold, and at defined checkpoints during training, a member submits the model to a standardized evaluation suite maintained by the Secretariat and administered by the Inspectorate, covering the risk categories the International AI Safety Report and the frontier labs' own frameworks already recognize: biological and chemical uplift, offensive cyber capability, autonomous replication and recursive self-improvement (RSI), deception of overseers, and persuasion at scale. Results are reported to the Board and summarized in the public report. Models that trip defined capability tripwires are subject to a mandatory pause and a Board review of the developer's evidence of control before training continues or deployment proceeds. Passing an evaluation does not certify that a model is safe in every setting; the reports must state what was tested, what access the evaluators had, and what remains uncertain.[30]

Incident reporting. Any member that observes a safety-relevant failure, a security breach involving model weights, an attempted exfiltration, a model behaving in ways its developers did not intend and could not readily correct, or evidence of a non-member approaching the threshold, reports it to the Inspectorate within seventy-two hours. Incidents are shared among members promptly under the Compact's security and confidentiality rules and published, in aggregate and with sensitive details removed, in the quarterly report.

Open weights. A model whose weights are released publicly cannot be recalled, paused, or inspected after the fact. The Compact therefore treats the public release of weights for any model above the threshold as a deployment of the highest consequence: it requires the full evaluation suite, a Board review, and a finding that the model's capabilities fall below every defined tripwire. Below the threshold, the Compact does not restrict open research and says so plainly, because the openness of ordinary science is one of the things this arrangement exists to protect.

How verification actually works

This is the section that answers the argument from Part One, so I want to be concrete. The Compact does not ask anyone to trust anyone. It builds on the fact that frontier AI, like fissile material, has physical inputs that can be counted, and it adds the access to records and models that hardware alone cannot supply. I learned the design principle behind this section in intelligence operations: a source's report is worth little until it is corroborated by something the source does not control, and partners will share what can be checked and withhold what cannot. So the arrangement must be built around the checkable things. In frontier AI, the most checkable thing is the hardware. This is the fourth of the first principles from Part Two, applied.

[30] International AI Safety Report 2026, discussion of evaluations and risk management; Bengio et al., "Managing extreme AI risks," 2024. Testing can miss capabilities and does not establish universal safety.

The Compute Registry is the foundation. Every advanced AI accelerator above a performance threshold set by the Board is registered at the point of manufacture, with a unique identifier, and its location and owner are reported by Infrastructure Members throughout its life, borrowing the discipline of nuclear material accounting without pretending that chips and uranium behave alike. Several stages of the advanced-chip supply chain are concentrated in a handful of firms, every one of them domiciled in a nation with an interest in this framework. That concentration is what makes verification possible, and it will not last forever, which is why the Registry must be built now. Legacy hardware, resale, smuggling, and non-member production will all have to be accounted for; the charter treats them as known gaps to be measured and narrowed, not as reasons to do nothing.

Facility declarations and inspections come second. Any facility housing more than a threshold quantity of registered compute is declared, and the Inspectorate may inspect it, announced or unannounced, to confirm that the chips present match the Registry and to examine evidence that the runs in progress match the declarations. Routine hardware checks do not require access to a member's code, data, or model weights. Verifying what a cluster is actually doing sometimes will, and the charter provides for secure access to records, technical interfaces, or the model itself under managed access procedures adapted from chemical weapons and nuclear inspection practice, which protect legitimate secrets while answering the inspector's question.[31] Inspection rights rest on member consent under the charter and, once enacted, on the anchor statutes.

Independent measurement comes third. Frontier facilities draw power on a scale that leaves evidence in utility records and satellite imagery. The Inspectorate maintains its own capacity to analyze lawfully obtained energy data and imagery at declared and suspected sites and to reconcile them against declarations. These observations corroborate; they do not by themselves identify a training run or its purpose, and commercial monitoring is not a substitute for a government's national technical means. They are the thing that makes a concealed facility hard to hide, and that is enough to earn their place.

Hardware-enabled safeguards come fourth and grow in importance over time. Researchers at the Center for a New American Security and elsewhere have laid out how accelerators could be built to attest to their location and workload, to report their participation in a cluster, and to require periodic authorization to continue operating.[32] The Compact directs Infrastructure Members to build such mechanisms into successive generations of hardware on a published schedule, so that within a decade the Registry is enforced by the silicon itself. That is an engineering program, not a finished product: it requires security testing, resistance to tampering, and agreement on who holds the authorization keys, and the charter says so.

Protected disclosure comes last and matters most. The people most likely to know that a member is cheating are that member's own engineers. The Compact establishes a protected channel to the Inspectorate, with legal protection guaranteed by anchor statute in every participating state, financial rewards funded by the levy, and a standing

[31] Chemical Weapons Convention, Verification Annex, Part X, challenge inspections and managed access. AI-specific procedures, consent, and implementing law would need to be negotiated.

[32] Onni Aarne, Tim Fist, and Caleb Withers, "Secure, Governable Chips," Center for a New American Security, January 8, 2024; Sastry et al., 2024.

offer of employment and relocation for those whose disclosures are substantiated. The nuclear arrangements never had this. They would have been stronger if they had. A cross-border promise of protection is only as credible as the laws behind it, which is one more reason the anchor statutes cannot wait.

Reporting to governments and to the public

The Authority speaks on a schedule that no member can alter. Each quarter it publishes a Compliance Report listing every declared training run, every evaluation result in summary, every incident in aggregate, and every enforcement action, with the vote of each Principal recorded by name. Each year it publishes a State of the Frontier report assessing where the technology stands, what risks have grown or receded, what the evidence on control shows, and what the Authority recommends to governments.

Each report is issued in two versions. The public version is released to the world on the same day for everyone, in the working languages of the Authority. Its redactions are narrow and explained: dangerous technical detail, protected personal or commercial information, and material restricted by law. The government version carries the same core compliance findings and a restricted annex, delivered simultaneously to every Oversight Council member and National Liaison Office. Access to the annex follows the security and information-sharing arrangements written into the implementing agreements; the Compact cannot promise every government unrestricted access to another country's classified information, and it does not. What it promises is symmetry: the same standards, the same core findings, and a stated reason for every exception. Where a limit on access prevents the Inspectorate from reaching a firm conclusion, the report says so. That symmetry is what makes the arrangement acceptable to rivals, and the public release is what makes it accountable to the people.

How it is paid for

The Authority is funded by its members, not by governments, and this is a matter of principle rather than convenience. A body that depends on appropriations from national legislatures will be captured by national interests within one budget cycle. Industry funding carries its own risk of capture, which is why the Oversight Council, the Inspector General's single fixed term, the post-employment bar, and the published votes exist.

Each Frontier Lab and Infrastructure Member pays an annual levy assessed on its registered frontier compute, so that those who build the most contribute the most and the burden tracks the risk. The Board sets the rate. At founding I would expect an operating budget on the order of one to two billion dollars a year, a planning assumption rather than a costed estimate, and one that is less than any single member's frontier training budget and a rounding error against the value of the industry the Compact protects. A portion of each year's levy is placed in a permanent endowment with the goal of accumulating, within a decade, reserves sufficient for several years of Inspectorate operations, so that the Inspectorate cannot be starved into silence. The levy rate, staffing plan, and endowment assumptions are published.

Governments may fund the Safety Research Commons with grants and may contribute to the Abundance Program described in Part Four. They may not fund the Inspectorate or

the Secretariat, directly or through grants. State-owned members pay the same levy as everyone else.

The charter

The Compact is a signed instrument creating contractual obligations among members, enforceable as contracts are enforceable. The anchor statutes supply what a contract cannot: public powers, legal recognition, and the authority to act on findings. Its articles are outlined in Appendix A. Its core commitments fit on one page: to declare, to submit to evaluation, to pause at the tripwires, to report incidents, to admit inspectors, to register compute, to contribute to the Commons, to pay the levy, to protect those who disclose, to implement lawful compute restrictions when ordered, and to accept the enforcement ladder. Competition law, export controls, confidentiality, review rights, and the limits on delegating public power differ by jurisdiction and are addressed in each anchor statute before coercive measures take effect.[33]

Enforcement

A charter without consequences is a press release. The Compact's enforcement ladder is public, graduated, and automatic at its lower rungs, so that no Principal has to choose between enforcing and offending.

Rung	Finding	Consequence	Decided by
1	Late or incomplete declaration, minor procedural lapse	Public notice; corrective plan within 30 days	Inspector General
2	Repeated procedural violation; failure to correct	Contractual penalty in addition to the levy; forfeiture of posted compliance bond	Board, simple majority
3	Material misstatement; obstruction of an inspection; continuing past a tripwire without review	Suspension from the Safety Research Commons; loss of vote; public censure of named executives	Board, two-thirds
4	Undeclared frontier training run; deployment past a tripwire without review	Compute restrictions: Infrastructure Members limit new supply, service, and support where lawful under the anchor statutes; anchor governments and the Oversight Council receive referrals	Board, two-thirds
5	Concealed weapons-relevant capability; unauthorized exfiltration or transfer of frontier weights; sustained defiance	Expulsion; referral to all anchor governments for sanctions on the organization and personal sanctions on responsible executives; publication of the evidentiary basis, subject to the reporting protections	Board, two-thirds; Co-Chairs may not veto

[33] U.S. Federal Trade Commission, "Group Boycotts," Guide to Antitrust Laws. A private charter creates no antitrust exemption; the anchor statutes must.

Three features make this ladder credible. First, the decisive penalty, restriction of compute, is executed by members rather than governments, because members control the chips and the clouds. Members can restrict what they actually control, new supply and continuing service, and they do so under the anchor statutes, which is what turns a coordinated refusal to deal from an antitrust problem into a lawful sanction. A violator does not need to be prosecuted. It needs to be cut off. Second, the anchor statutes give the ladder legal force at home: each member government commits, by domestic law, to treat an Authority finding at rungs four and five as grounds for export control action, market access restriction, and personal liability for the executives responsible, under that country's own evidentiary and procedural rules. An Authority finding supports a referral; it does not replace the legal determination each country must make. Third, Principals vote by name and the votes are published, so that a Principal who shields a violator does so in front of his own government, his own customers, and the Oversight Council. The charter specifies notice, an opportunity to respond, recusal for conflicts, an independent appeal, and narrowly defined emergency measures subject to prompt review.

Every member posts a compliance bond at accession, sized to its compute, held by the Authority, and forfeited in whole or in part at rung two and above. Money that has already been paid concentrates the mind in a way that money merely threatened does not.

What the charter cannot do by itself

I have described the institution as I believe it should be built. I owe the reader an equally plain account of what a charter among companies cannot do on its own, because the people who can build this will ask, and because an argument that hides its weak points does not deserve to win.

A private charter cannot create public power. It cannot waive an export control, grant immunity from liability, authorize competitors to restrict supply to a third party, or let anyone unplug hardware they do not own. Every coercive element of the ladder depends on the anchor statutes, which is why they are the second act of the founding sequence in Part Four rather than an afterthought. Until they exist, the Compact is a contract with bonds, censure, and publication as its teeth. Those are real teeth, and the nuclear arrangements began with less. They are not the whole animal.

Verification cannot be perfect. Hardware inventories, power records, and satellite imagery can establish that a facility exists and roughly what it is doing, not exactly what a model is being trained to do. Legacy chips, resale, smuggling, non-member supply, and improving efficiency all open gaps, and the on-chip safeguards that would narrow them are a research program today. The Compact's answer is to measure its own detection limits and publish them, and to pair hardware verification with managed access to records and models. That is how arms control has always worked: not certainty, but a cost of cheating high enough and a chance of detection real enough that cheating stops being worth it.

Independence cannot be assumed. An industry-funded body inspecting its own funders faces the risk of capture. The fixed term, the post-employment bar, the published votes,

and the Oversight Council are the best safeguards I know of from institutions that have faced the same problem. They will need to be tested.

Feasibility cannot be proven on paper. Whether Beijing would accept co-chairmanship on these terms, whether Washington would accept identical obligations for its own labs, whether the labs would accept inspectors in their data centers, and whether legislatures would pass the statutes in time are questions only the attempt can answer. The nuclear arrangements were dismissed as naive for most of the decade before they were built.

And the charter cannot solve the control problem. It can pause, review, pool the research, and make the evidence public. It cannot legislate a scientific result. If the Commons and the field cannot demonstrate reliable control of systems approaching the threshold, the Compact's honest function is to keep the pause in place and say why, in public, to everyone, until they can.

Part Four: Why This Can Actually Happen

Proposals like this one usually die of a single question: why would anyone sign? I have tried to design the Compact so that each necessary party gains more by joining than by staying out, and so that the parties who join first can make membership the price of admission for everyone else.

Why the frontier labs would join

The frontier labs have four reasons. First, they are already doing much of this. The leading laboratories have published safety frameworks, run capability evaluations, formed an industry forum in 2023 to share safety research, and are now, as Part One described, discussing a standards body with pre-release testing and calling publicly for pacing agreements and embedded evaluators.[34] They have done all of it voluntarily, without verification, and without the ability to know whether their rivals, above all their rivals in China, are doing the same. The American negotiating team heading into September's talks is reported to be carrying a proposal that American and Chinese laboratories "police themselves" and share threat information.[35] That proposal is the Compact without the verification and without the consequences. It is a good instinct that needs a skeleton. Second, liability. A lab that follows Authority standards and passes Authority evaluation should, under anchor statute, receive a safe harbor against liability for harms it could not reasonably have foreseen. Only legislation can grant that, and it should never excuse concealment or misconduct; but it is a benefit worth billions that only a recognized standards body can unlock. Third, regulatory coherence. Frontier labs today face divergent and multiplying national rules. A single set of standards recognized by every anchor government is cheaper than fifty, even if implementation still differs by legal system. Fourth, the thing they now say in public as well as in private: they would slow down if they could be sure their competitors had too. The Compact is the mechanism that lets them be sure.

Why the United States would anchor it

The United States hosts roughly three-quarters of the capacity in the large-cluster dataset discussed in Part One. That is a snapshot of today's infrastructure, not a guaranteed future share, and it will not hold forever. The moment to write the rules is the moment one holds the strongest hand, and an arrangement anchored in compute verification locks in the advantage of the party that controls the chips. American negotiators should understand the Compact as the conversion of a temporary lead into a durable structure, much as the NPT converted a temporary American nuclear lead into a framework that has served American interests for half a century, while imposing obligations on the United States as well. The Compact also answers Beijing's second

[34] Microsoft, "Anthropic, Google, Microsoft and OpenAI launch Frontier Model Forum," July 26, 2023; Amodei, "We Must Pace the Frontier," September 2026.

[35] U.S.-China Economic and Security Review Commission news summary, "US, China to discuss AI safety risks, cooperation on monitoring AI-directed cyberattacks," September 2026, and Reuters, September 4, 2026. A reported U.S. proposal, not an agreed bilateral arrangement.

stated condition, equal audit requirements for American firms, on terms the United States helps write and verifies rather than terms imposed from outside.

Why China would co-lead it

The conditions in the August commentary from Beijing are a co-equal voice in defining what safety means and identical obligations for American firms. The Compact grants both. It also offers China something it cannot obtain by racing: a basis for negotiating more predictable, lawful access to the frontier hardware supply chain, which is presently the object of escalating export controls, and a seat at the table where the standards that will govern the global market are written. Membership would not by itself lift an export restriction; that would remain a decision for governments, and the Compact gives them a reason to make it.[36] A China outside the Compact faces restrictions from every participating supplier, softened by domestic production and existing stocks but real, and the prospect of an American-led framework that treats it as the object of governance rather than a partner in it. A China inside the Compact is a founder of the most important institution of the century. I do not expect Beijing to join out of goodwill. I expect it to join because the alternative is worse, which is the only reason any state has ever joined anything. Whether that bargain outweighs its concerns about sovereignty and access is a question only the negotiation can answer.

Why everyone else would join

For the nations and organizations that do not hold frontier compute, the Compact must deliver the Atoms for Peace bargain: benefit in exchange for participation. The charter commits members to an Abundance Program under which the safe, evaluated fruits of frontier development, in medicine, agriculture, education, and energy, are made available to member and non-member nations on terms that do not require them to build their own frontier capacity. A nation that can obtain the benefits of superintelligence without racing for it has far less reason to race, and every reason to support a framework that keeps the race from ending in catastrophe. This is not charity. It is the same non-proliferation logic that kept the nuclear club small, and it is the means by which the Compact grows from a two-power arrangement into a global one.

Sequencing: how to make membership the price of admission

The Compact does not need every party at once. It needs the right parties in the right order.

It begins with a founding convening of the American and Chinese frontier labs, hosted by a neutral institution and attended by the Infrastructure Members whose chips and fabs sit in the supply chain. If the labs that control the great majority of frontier compute and the firms that supply it agree that compute will flow only to organizations that accept the Compact's obligations, membership ceases to be a choice for anyone else. This is how the arrangement bootstraps: not by persuading every actor, but by aligning the handful of actors who control the inputs. A private agreement among competitors is not, by itself, a

[36] Bureau of Industry and Security, U.S. Department of Commerce, "Department of Commerce Revises License Review Policy for Semiconductors Exported to China," January 2026. A licensing framework; Compact membership would not by itself authorize an export.

lawful basis for excluding other firms, which is why governments must be in the room from the first day, not to draft the charter but to say what the law will and will not permit and to begin writing the statutes that will permit more.

Governments enter formally second, through the Oversight Council and the anchor statutes that recognize the Authority's standards, protect its whistleblowers, and define action on its referrals. The two co-leading governments move first and, if they move together, others follow, because a company domiciled in a country without an anchor statute is a company its competitors can lawfully decline to supply.

Formal treaty codification, with a permanent international legal personality for the Authority, comes last, once the institution has operated long enough to have proven it works. The IAEA preceded the NPT by more than a decade. The IAEA itself was created by an intergovernmental statute, so the analogy is to the sequence, not the legal form: build the working institution first, and let the treaty ratify what already exists.

Two paths to a standing institution

Three years is the timetable a careful institution-builder would write, and it is roughly the interval between the Acheson-Lilienthal Report and the first Soviet test. I no longer believe we have three years to spare. The rate of capability advancement described by the labs themselves, the September warnings from inside them, and the possibility that the compute chokepoint erodes before the Registry is built all argue for standing the institution up faster and letting the law catch up. So I present two paths, and I recommend the first.

The fast path stands up a working Authority in twelve to eighteen months by building on executive action and authorities that already exist, and by accepting that the first version will be provisional, contractual, and incomplete. It runs as follows.

Phase	Timing	Milestones
Ignition	Days 0 to 30	The two presidents agree the one sentence in the ask below. Founding convening of American and Chinese Frontier Labs and Infrastructure Members within thirty days, hosted by a neutral institution. Interim Secretariat funded by the founding labs. Each government designates its Oversight Council member and National Liaison Office by executive action.
Founding Declaration	Days 30 to 90	Founding members sign an interim compact with immediate effect: training declarations above a provisional threshold, seventy-two hour incident reporting, exchange of evaluation results, and posting of bonds. Governments direct existing export-control and customs data on advanced accelerators to the Interim Secretariat as a provisional Registry. The International Network for Advanced AI Measurement, Evaluation and Science is engaged as provisional evaluation partner. First Compliance Report published at day ninety.

Phase	Timing	Milestones
Standing up	Months 3 to 9	Charter signed. Co-Chairs elected. Inspector General appointed and Inspectorate stood up with directly employed staff supplemented by personnel seconded from members under strict conflict rules. Compute Registry live for new production through Infrastructure Member participation. First announced inspections. Anchor statutes introduced in both legislatures; interim whistleblower protections by executive action where law allows. Safety Research Commons begins pooled work on control and evaluation.
Full operation	Months 9 to 18	First unannounced inspections. Anchor statutes enacted, activating rungs four and five of the ladder. Hardware safeguards roadmap agreed with chip designers. Abundance Program pilot. Endowment funding begins. Oversight Council's first public audit of the Inspectorate.
Codification	Month 18 onward	Remaining anchor statutes enacted in other member states. Treaty negotiation to grant the Authority permanent international legal personality. Thresholds reviewed against capability and efficiency evidence.

The fast path has costs, and I want to name them. For its first year the Compact rests on contract and executive action, both of which can be reversed by a change of mind or a change of administration. Inspections are consensual until statutes make them mandatory. Seconded staff raise a fair question about independence, which the conflict rules and the Oversight Council are designed to answer but cannot dissolve. The coercive rungs of the ladder wait for legislation, so the first year's teeth are bonds, censure, suspension, and publication. And a Registry built from export-control data will be incomplete. None of that is an argument against the fast path. It is a description of what a real institution looks like at twelve months instead of thirty-six. The nuclear framework was provisional for a decade, and it held.

The legislative path is the same institution built in the more familiar order, with statutes preceding coercive powers and a fuller Registry preceding inspections. It is the path to take if the two presidents cannot agree the sentence, or if the legislatures move first.

Phase	Timing	Milestones
Founding	Months 0 to 6	Founding convening of American and Chinese Frontier Labs and Infrastructure Members. Agreement on the eight principles; government consultations on legal feasibility. Interim Secretariat stood up. Charter drafting committee, co-chaired, begins work. Public statement of intent by founding members.
Charter	Months 6 to 18	Charter signed. Co-Chairs elected. Inspector General appointed. Oversight Council seated. Compute Registry pilot for participating new production. First training declarations filed. First quarterly Compliance Report published. Compliance bonds posted.

Phase	Timing	Milestones
Anchoring	Months 18 to 36	Anchor statutes enacted in the United States and China, then in other member states. Inspectorate at full strength; first unannounced inspections. Standardized evaluation suite in force. Safety Research Commons operating with pooled compute. Abundance Program launched. Whistleblower protections enacted and coverage published.
Codification	Year 3 onward	Validated hardware safeguards phased into new accelerator generations. Endowment accumulation continues toward the reserve goal. Treaty negotiation for permanent international legal personality. Thresholds reviewed against capability and efficiency evidence.

THE FIRST NINETY DAYS

Whichever path is taken, the Authority should prove itself with a single deliverable before anyone is asked to trust its design: a first Compliance Report, published at day ninety, listing every frontier training run declared by founding members since the Founding Declaration, the compute and facilities involved, the evaluations completed, the incidents reported, and the gaps in what the Interim Secretariat could see. It will be thin. It will also be the first document in history in which American and Chinese frontier laboratories disclose, side by side and to the public, what they are building. That is the proof of concept. Everything after it is scale.

Part Five: Objections

I have tried to state each objection in its strongest form.

"China will cheat."

Probably someone will, at some point. The question is not whether cheating is possible but whether it is detectable and costly, and the Compact is designed so that both are true. A concealed frontier training run requires tens of thousands of chips diverted from declared facilities or obtained outside the Registry, a power draw that leaves evidence in utility records and from orbit, a facility that inspectors cannot enter, and the silence of every engineer who works there. None of those signals is conclusive alone; unregistered hardware, domestic supply, and more efficient methods can each open a gap, and the Compact must test its detection methods and publish their limits. But the nuclear arrangements caught cheaters with far less, and the consequence of being caught here is not a strongly worded letter; it is the loss of the chips. A framework in which cheating is hard, visible, and expensive is a framework in which cheating is rare, which is all any arms control arrangement has ever achieved and all this one needs to.

"The United States would be giving away its lead."

The Compact asks the United States to give away nothing that it can keep. Model weights, architectures, and training data remain proprietary. What it shares is safety research, which it benefits from sharing, and what it accepts is inspection of hardware it already knows it possesses, under obligations that fall on its rivals identically. In exchange it obtains independently checked visibility into Chinese frontier development that no intelligence program can reliably provide, and it locks in a compute-based verification arrangement while it controls the compute.

"Private companies cannot be trusted to govern themselves."

They cannot, and the Compact does not ask them to. It asks them to govern one another, under an independent Inspectorate they fund but do not control, beneath an Oversight Council of governments that can compel answers and order reviews, with every vote published, every report delivered to governments, and every serious finding backed by the force of national law. This is not self-regulation. It is industry regulation under public oversight, which is how the safety of securities markets actually functions and how much of aviation and nuclear safety functions in practice. The risk of capture is real and I have said so. Governments set the consequences and supervise the supervisors. The professionals do the inspecting, because they are the only ones who know what to look for.

"Governments will never cede this to the private sector."

They are not asked to cede anything. Every government retains full sovereign authority over the companies within its borders, and the anchor statutes expand that authority rather than shrink it. Governments gain a standing seat on the Oversight Council, a

source of independently checked information about foreign frontier development that they presently lack, delivered on a schedule they do not have to negotiate, in a format their rivals receive too. Whether a given government finds that exchange acceptable is a negotiating question. No head of state has ever refused a reliable intelligence source because it came from an inconvenient direction.

"Military and intelligence AI programs will never be included."

This is the hardest objection and I will not pretend otherwise. Two responses. First, the Compact governs the training of frontier models and their highest-risk capabilities, not their every application. A government may apply a declared, evaluated model to any lawful national security purpose it chooses, subject to the agreed safeguards; what it may not do is train a frontier model in secret or change its capabilities materially without review. Second, the compute chokepoint applies to defense programs as to everyone else. A classified laboratory still needs chips, and the chips are registered. The framework does not need to inspect the classified program; it needs to know the program's compute exists and where it is. That is a partial answer, and I acknowledge it. Hardware accounting cannot establish the purpose or capability of every classified program, and the NPT, which never gave the IAEA a full account of the weapon states' arsenals, is not a precedent for claiming otherwise. Full coverage would require specifically negotiated national security arrangements. A partial answer that governs the hardware is better than a complete answer that governs nothing, and it is a reason to negotiate the hard part in the open rather than pretend it is solved.

"This will slow down the benefits."

It will slow some things down, and it should. A model that trips a capability tripwire will wait for a review it would not otherwise face, and the paper should say that plainly. Everything else points the other way. Today's frontier labs duplicate one another's safety work, hoard evaluation results, and deploy under legal uncertainty. The Commons reduces the duplication, the shared evaluation suite reduces the hoarding, and the safe harbor reduces the uncertainty. More importantly, the Abundance Program is the only mechanism on offer that would deliver the benefits of this technology to the majority of humanity that will otherwise wait decades for them. Safety and abundance are not in tension. Unverified racing is the enemy of both.

"It is too late; the technology is moving too fast."

It is later than it should be. It is not too late, because the compute chokepoint still exists. The number of firms that can manufacture frontier accelerators or the tools that make them is still small enough to count, and every one is domiciled in a nation with a stake in this outcome. That will change. Manufacturing will diffuse, efficiency will improve, and the day will come when a frontier run can be assembled from hardware no registry tracks. The window in which verification is physically possible is open now and closing. That is the reason for the fast path in Part Four, and the reason to act this year rather than to study the question for another five.

What I Am Asking

To the President of the United States and to the President of the People's Republic of China: when you meet, I ask you to agree on one sentence. That the frontier labs of your two nations should convene, within thirty days, with the firms who supply their hardware, to draft a compact for the verified and collaborative development of advanced artificial intelligence, co-chaired by both nations' industries, overseen by both governments, and open to all who meet its obligations. You need not agree on what safety means. You need only agree that your professionals should sit down and work it out under rules that bind both sides equally, which is what your own negotiators are already saying to each other in different words.

To the Congress of the United States and to legislators in every nation that hosts a frontier lab: I ask you to begin drafting the anchor statute now, in parallel, so that when the Compact is signed there is law waiting to give it force. The statute needs to do four things: recognize Authority standards, protect those who report violations, provide a defined safe harbor for those who comply, and establish lawful procedures for acting on Authority referrals. None of these requires you to wait for the diplomats.

To the chief executives of the frontier labs: you have said, most of you, that the technology you are building could be the most dangerous in human history, and you have said that you cannot slow down alone because your competitors will not. Some of you have now called for pacing agreements and outside evaluators. I am offering you the mechanism. I ask you to say publicly that you would join such a compact if your peers did and to name one person on your leadership team to attend the founding convening. That is all it takes to begin. Everything else in this paper is detail.

To everyone else: this is being built in your name and with your future. You have every right to ask to see it. Ask.

Conclusion

The argument that we cannot slow down because our adversaries will not is correct. It is also the strongest argument for the Threshold Compact that I know of, because it describes with perfect precision a problem that trust cannot solve and verification can. Reason from first principles and you arrive at the same place: rivals race, rivals do not trust, verification requires independently checkable evidence, and frontier AI has exactly such a thing in its hardware. The design follows. We faced this problem once before, with a technology that could end the world, among rivals who hated each other more than any rivals do today. We built institutions. We counted the material. We let one another look. It did not remove the danger. It worked well enough that I am alive to write this and you are alive to read it.

This time the problem is harder in one respect that changes everything: the weapon may not obey its owner. Reliable human control over superintelligence is unproven, and no flag on the building changes that. So the institution we build must do what the nuclear one never had to. It must govern not only who crosses the threshold and how openly, but whether and on what evidence anyone crosses it at all. That is a heavier task. It is also the only version of the task worth doing.

We can do this, and we can do it better than our grandparents did, because this time we know the template and because this time the prize on the other side of the threshold is not merely the absence of catastrophe but the presence of abundance beyond anything they imagined. The only thing we cannot do is keep running and hope. I have spent my life in institutions that were built because someone refused to hope when they could act. I am asking the people who can act to do so.

Appendix A: Outline of the Threshold Compact

The charter is organized in fifteen articles. This outline states the substance of each; drafting language is for the charter committee.

- **Article I.** Purpose. To ensure that advanced artificial intelligence is developed under conditions of mutual verification, collaborative safety research, government oversight, and public transparency, and that its benefits are shared widely, so that no organization or nation crosses the threshold to superintelligence in secret, alone, or without evidence of control adequate to the risk.
- **Article II.** Definitions. Compute and capability thresholds, frontier model, superintelligence, registered compute, declared facility, incident, tripwire capability, member categories, anchor government, anchor statute.
- **Article III.** Membership. Capability-based admission; categories and obligations; accession procedure; compliance bond; withdrawal with twelve months' notice and continuing obligations for previously declared systems.
- **Article IV.** Governance. Board of Principals; Co-Chairs; voting thresholds; conflict of interest and recusal rules; publication of votes; the Council on Human Dignity.
- **Article V.** Government Oversight. The Oversight Council: composition, powers to compel answers, commission audits, order reviews, and refer matters; National Liaison Offices; limits on Council authority over Board decisions.
- **Article VI.** Declarations, Evaluation, and the Tripwire Pause. Pre-training declarations; capability-based triggers; checkpoint and pre-deployment evaluation; mandatory pause and Board review of evidence of control; treatment of open-weight release; the standardized evaluation suite and its revision.
- **Article VII.** Incident Reporting. Seventy-two-hour reporting; categories; sharing under security and confidentiality rules; public aggregate disclosure.
- **Article VIII.** The Compute Registry. Registration at manufacture; lifetime tracking and treatment of legacy or unregistered hardware; facility declarations; know-your-customer obligations of Infrastructure Members; hardware-enabled safeguards schedule.
- **Article IX.** The Inspectorate. Independence; appointment and removal of the Inspector General; staffing, secondment rules, and vetting; declared and unannounced inspection; managed access to records and models; independent measurement; protected disclosure channel and rewards.
- **Article X.** The Safety Research Commons. Contribution obligations in personnel and compute; priority of work on control and evaluation; sharing under security and confidentiality rules; delayed public release; the line between shared safety and proprietary capability.
- **Article XI.** Reporting. Quarterly Compliance Report; annual State of the Frontier; public and government versions; simultaneous delivery of common core findings;

lawful access to the restricted annex; withholding standard and disclosure of withholding.

- **Article XII.** Finance. Compute-scaled levy; endowment; prohibition on government funding of the Inspectorate and Secretariat; publication of assumptions; audit.
- **Article XIII.** Enforcement and Anchor Statutes. The five-rung ladder; lawful compute restrictions; notice, response, independent appeal, and emergency review; referral; the model anchor statute members undertake to seek in their home jurisdictions, including whistleblower protection, defined safe harbor, executive liability subject to national law, and competition-law compatibility.
- **Article XIV.** The Abundance Program. Shared-benefit obligations toward non-frontier nations; governance; funding.
- **Article XV.** Amendment, Interim Operation, and Dispute Resolution. Two-thirds plus Co-Chair assent for amendment; provisions governing the interim compact and the transition to full operation; arbitration for disputes among members; publication of awards.

Appendix B: On the Name

I considered many names and rejected most for the same reason: they described the institution rather than the problem. "International Artificial Intelligence Authority" is accurate and forgettable, and it invites the immediate and correct objection that it sounds like a copy of the nuclear agency, which would concede the argument that this is a bureaucratic exercise rather than a response to a threshold in human history.

Threshold does three things. It names the thing everyone fears and no one can yet locate: the line past which these systems exceed our ability to govern them. It carries the sense of a doorway, which keeps the promise of abundance in view alongside the danger. And it is a plain English word that a head of state, an engineer, and a citizen understand the same way without a briefing. The institution exists to make sure the threshold is crossed, if it is crossed, deliberately, together, and in the open.

Compact rather than treaty because the first signatories are organizations, not states, and because the word describes a mutual obligation freely entered rather than a settlement imposed. The Mayflower Compact and the interstate compacts of the American constitutional tradition carry exactly this sense.

Authority rather than agency, council, or forum because the body must be able to find, to censure, and to trigger consequences. Agencies advise. Forums discuss. An authority decides. The label grants no power by itself; the charter and the anchor statutes define what this body may decide.

I hold no attachment to any name. I hold a great attachment to the institution.

Appendix C: Sources

Sources are cited in footnotes throughout the paper and gathered here by type. Research cutoff: September 14, 2026. Statements by developers and researchers establish what

their authors said, not that their predictions are certain or that they endorse this proposal. Reported negotiations are identified as reporting. The Compact's institutional design, budget, and timetables are proposals.

STATEMENTS AND PUBLICATIONS BY FRONTIER AI DEVELOPERS AND RESEARCHERS

- **Altman, Sam, Greg Brockman, and Ilya Sutskever.** "Governance of Superintelligence." OpenAI, May 22, 2023. openai.com/index/governance-of-superintelligence
- **Amodei, Dario.** "The Adolescence of Technology." January 2026. darioamodei.com/essay/the-adolescence-of-technology
- **Amodei, Dario.** "We Must Pace the Frontier." September 2026. darioamodei.com/post/we-must-pace-the-frontier
- **Anthropic.** "Responsible Scaling Policy, Version 3.0." Effective February 24, 2026. anthropic.com/responsible-scaling-policy/rsp-v3-0. Anthropic. "Responsible Scaling Policy" collection. anthropic.com/responsible-scaling-policy
- Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Stuart Russell, et al. "Managing extreme AI risks amid rapid progress." Science 384, no. 6698 (May 2024): 842 to 845. doi.org/10.1126/science.adn0117
- **Bengio, Yoshua (Chair), et al.** International AI Safety Report 2026. February 3, 2026. internationalaisafetyreport.org/publication/international-ai-safety-report-2026
- **Center for AI Safety.** "Statement on AI Risk." May 30, 2023. safe.ai/work/statement-on-ai-extinction-risk
- **Coxon, Jacob.** Resignation statement. X, September 8, 2026. x.com/hilbertspaess/status/2097476196791709843. Reported in Huamani, Kaitlyn, Associated Press, September 9, 2026; TechCrunch, "'Gambling with our lives': Anthropic researcher quits, warns against self-improving AI," September 9, 2026; TIME, September 9, 2026.
- **Google DeepMind.** "Strengthening our Frontier Safety Framework." September 22, 2025, updated April 17, 2026. deepmind.google/blog/strengthening-our-frontier-safety-framework. OpenAI. "Our updated Preparedness Framework." April 15, 2025. openai.com/index/updating-our-preparedness-framework
- **Hassabis, Demis.** "A Framework for Frontier AI and the Dawning of a New Age." July 14, 2026. demishassabis.substack.com/p/a-framework-for-frontier-ai-and-the-dawning-of-a-new-age
- **Hendrycks, Dan, Eric Schmidt, and Alexandr Wang.** "Superintelligence Strategy: Expert Version." March 7, 2025, revised April 14, 2025. arxiv.org/abs/2503.05628
- **Microsoft.** "Anthropic, Google, Microsoft and OpenAI launch Frontier Model Forum." July 26, 2023. blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum
- The Information, September 2026, on standards-body discussions among OpenAI, Google, and Anthropic, with follow-on reporting by Reuters and summaries

including AI Weekly, aiweekly.co/alerts/anthropic-openai-google-meet-since-july-on-ai-standards-body

COMPUTE GOVERNANCE AND VERIFICATION

- **Aarne, Onni, Tim Fist, and Caleb Withers.** "Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI and Advanced Computing." Center for a New American Security, January 8, 2024. cnas.org/publications/reports/secure-governable-chips
- Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, et al. "Computing Power and the Governance of Artificial Intelligence." February 13, 2024. arxiv.org/abs/2402.08797
- **Organisation for the Prohibition of Chemical Weapons.** Chemical Weapons Convention, Verification Annex, Part X, challenge inspections and managed access. opcw.org/chemical-weapons-convention/annexes/verification-annex

CONTEMPORARY REPORTING AND ANALYSIS, 2026

- **Tanner, Brooke, Cameron F.** Kerry, Elham Tabassi, Andrea Renda, and Andrew W. Wyckoff. "A summer of AI summits reveals a widening US-China divide." Brookings Institution, August 11, 2026. brookings.edu/articles/a-summer-of-ai-summits-reveals-a-widening-us-china-divide
- **United Nations.** Transcript of the launch of the Independent International Scientific Panel on AI preliminary report, 2026, presenting the dataset of the 500 largest known AI compute clusters. transcripts.un.org/en/asset/k1v/k1v0ss6l5a
- **Reuters.** "U.S., China hold AI risk and safety talks, White House says." May 15, 2024. Reuters. "U.S., China gear up for mid-September AI safety talks." September 4, 2026, republished by CNBC September 5, 2026.
- **Belmonte, Kyle.** "China Tells US: Agree on What AI Safety Means or September Talks Cannot Proceed." Tech Times, September 3, 2026, reporting commentary published by Yuyuantantian on August 31, 2026.
- U.S.-China Economic and Security Review Commission (news summary). "US, China to discuss AI safety risks, cooperation on monitoring AI-directed cyberattacks." September 2026.
- **UK Department for Science, Innovation and Technology.** Announcement of the International Network for Advanced AI Measurement, Evaluation and Science. December 9, 2025. UK Government. "The Bletchley Declaration by Countries Attending the AI Safety Summit." November 1, 2023.
- **Presidency of the French Republic.** "G7 Summit in Evian: Day Three." June 17, 2026. Xinhua. "29 countries sign agreement on establishing World AI Cooperation Organization." July 16, 2026. The White House. "Promoting Advanced Artificial Intelligence Innovation and Security." Executive order, June 2026.
- Bureau of Industry and Security, U.S. Department of Commerce. "Department of Commerce Revises License Review Policy for Semiconductors Exported to China." January 2026.

- **Financial Industry Regulatory Authority.** "About FINRA." finra.org/about. U.S. Federal Trade Commission. "Group Boycotts." Guide to Antitrust Laws. ftc.gov/advice-guidance/competition-guidance/guide-antitrust-laws/dealings-competitors/group-boycotts

HISTORICAL SOURCES ON THE NUCLEAR ORDER

- U.S. Department of State, Office of the Historian. "The Acheson-Lilienthal and Baruch Plans, 1946." Milestones in the History of U.S. Foreign Relations. history.state.gov/milestones/1945-1952/baruch-plans
- **Eisenhower, Dwight D.** "Atoms for Peace." Address to the United Nations General Assembly, December 8, 1953. Statute of the International Atomic Energy Agency, adopted October 23, 1956, entered into force July 29, 1957. iaea.org/about/statute
- **Treaty on the Non-Proliferation of Nuclear Weapons.** Opened for signature July 1, 1968; entered into force March 5, 1970. International Atomic Energy Agency, "Safeguards agreements." iaea.org/topics/safeguards-agreements
- CERN Convention, signed July 1, 1953, entered into force September 29, 1954. ITER Agreement, signed November 21, 2006, entered into force October 24, 2007.
- Soviet Nuclear Threat Reduction Act of 1991, Title II of Public Law 102-228, December 12, 1991. National Research Council. Global Security Engagement: A New Model for Cooperative Threat Reduction. National Academies Press, 2009.
- **Kennedy, John F.** News Conference 52, March 21, 1963. John F. Kennedy Presidential Library. Stockholm International Peace Research Institute. SIPRI Yearbook 2026: World Nuclear Forces.

About the Author

Logan Reed works on technology modernization and the responsible adoption of AI. He currently works at a professional services firm serving the federal government. His experience includes leading a Responsible and Ethical AI subgroup and facilitating an AI strategy workshop with senior executives. He is also Co-Founder and President of Volo Match. He is a former Army Captain with eleven years of military experience, including air and missile defense, NATO exercise planning in Europe, and coalition operations in Iraq, Syria, and Kuwait in support of Operation Inherent Resolve, and he later returned to Iraq as an intelligence contractor supporting United States and partner special operations forces with intelligence collection and analysis. His civilian career includes operations and technical leadership roles at Amazon and federal strategy consulting at Deloitte for defense and veterans' health organizations. He holds a Bachelor of Arts in Philosophy from Loyola University Maryland and has held a Top Secret clearance with access to Sensitive Compartmented Information. He lives with his family in Boiling Springs, North Carolina. He can be reached at hello@thresholdcompact.org.