Available Online: 30 October 2026 DOI: 10.54254/2753-7102/2026.38506

# Governing frontier AI under U.S.–China competition: a staged hybrid model for reducing escalation risk

***Cathy Cai Yi Yang***

Steinhardt School of Culture, Education, and Human Development, New York University, New York, USA

cathyyang0601@gmail.com

**Abstract.** Frontier Artificial Intelligence (AI) creates escalation risks under U.S.–China competition because rapid technological change, strategic secrecy, and private control of advanced systems complicate effective oversight. This paper evaluates which governance model can best reduce these risks through a qualitative comparative case-study approach supported by legal analysis. It compares binding multilateral law, expert soft law, bilateral agreements, the fragmented existing AI regime, and hybrid governance through the WTO, Tallinn Manual, SALT I and the ABM Treaty, current AI instruments, and outer-space governance. The comparison suggests that no single model adequately combines adaptability, participation, verification, and enforcement. Binding law is most useful for narrow, observable obligations; soft law supports technical clarification and experimentation; and bilateral coordination can address shared vulnerabilities between strategic rivals. The paper therefore proposes a four-stage hybrid framework: domestic oversight of frontier developers, a permanent U.S.–China commission, tested nonbinding safeguards, and a narrow binding treaty focused on nuclear decision-making and serious incident notification. Broader participation would follow as the framework becomes operational. By linking international commitments to domestic regulatory capacity and making legal obligations conditional on practical verification, this work identifies a feasible path for reducing miscalculation and technical-failure risks without assuming that governance can eliminate geopolitical competition.

**Keywords:** frontier AI, U.S.–China competition, hybrid governance, escalation risk, international law

## 1. Introduction

Frontier Artificial Intelligence is evolving at a speed that few existing models of governance can keep up with. New models and features update monthly, weekly, even daily, while international agreements can take years to negotiate and even longer to enter into force. This mismatch in speed is especially dangerous under the current U.S.–China rivalry, where technological competition rewards speed, secrecy, and strategic advantage. The central challenge is therefore how to build a framework that is strong enough to reduce escalation risks but flexible enough to adapt as the technology develops. This paper asks: What governance model, and with what features, would most effectively reduce the risks of escalation arising from geopolitical competition over frontier AI?

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

By looking at current existing models, the paper argues that no single governance model can manage the combined pressures of U.S.–China competition, rapid technological change, and private corporate control over frontier AI. The most effective response is a hybrid model that consists of different stages: binding domestic regulation of frontier AI developers, adaptable bilateral coordination between the United States and China as a foundation for broader cooperation, and ultimately a binding treaty focused on a narrow set of catastrophic risks. The framework begins with the United States and China because their rivalry currently creates the most immediate escalation risk, but its structure is designed to adapt if major players in AI change.

## 2. Literature review

As a new and constantly evolving field, scholarship concerning international AI governance remains limited and fragmented. Nevertheless, three main themes emerged in the existing literature: the difficulties involved in governing AI, the increasing competition between the United States and China for leadership in the field of AI, and the legal and institutional approaches that are available for regulating frontier AI.

### 2.1. The challenges of governing artificial intelligence

Cyphert argues that AI is particularly hard to regulate since it is global in scope, develops rapidly, involves multiple areas of law, and demands a level of technical knowledge which many lawmakers do not possess. This situation gives rise to two closely linked risks: that governments might adopt rules which they do not fully understand, or that such rules may quickly become outdated once they have been passed [1]. Tallberg and his co-authors go on to apply this issue to the international level, stating that global AI governance consists of a regime complex comprising overlapping state, international, private, and technical arrangements and lacks a central body or clear hierarchy [2]. Thus, both of these studies illustrate the difficulty of governing AI both within individual states and between states.

The fact that private companies play a central role presents another problem. Since these firms have access to a great deal of the "insider" information required for a proper assessment of risk, governments might have to rely on the very companies which they are trying to regulate.

Roberts and his co-authors see this kind of dependence as part of a wider governance issue caused by competition between states, the lack of effective international coordination, and the disagreement about which institutions should be responsible for governing AI [3]. International governance must therefore deal with both the cooperation between states and the government's capacity to keep an eye on the private companies that are developing the technology.

### 2.2. U.S.–China competition

These governance problems become more difficult under growing U.S.–China rivalry.

Cheng and Zeng show that China is not simply following rules established by Western states but is actively attempting to shape global AI governance according to its own priorities. This reflects broader differences over sovereignty, state authority, private-sector participation, and the role of multistakeholder governance [4].

Schmid and his coauthors describe the competition over AI as more than a conventional arms race. It also involves economic leadership, technological networks, security, and international influence. Landwehr-Matlé, Oppermann, and Lambach further show how official narratives reinforce this competition by presenting AI leadership as a source of national power and status [5, 6]. These narratives can make cooperation more difficult by encouraging each side to view the other's progress as a strategic loss.

The stakes rise further as AI enters military planning, intelligence, early-warning systems, and nuclear-related decision-making. Chernavskikh and Palayer argue that military AI may shorten decision timelines, produce recommendations that are difficult to explain, and increase the risk of miscalculation during a crisis. Johnson similarly shows that AI-enabled counterforce capabilities may weaken strategic stability and increase the risk of inadvertent escalation between nuclear-armed states [7, 8].

Together, this literature reveals a central tension: effective AI governance requires transparency, technical cooperation, and adaptable rules, while strategic competition encourages secrecy, speed, and resistance to external oversight.

## 2.3. Possible international governance models

The literature on international legalization provides several possible models. Abbott and his coauthors define legalization through three dimensions: obligation, precision, and delegation. Obligation concerns whether states are legally bound, precision concerns how clearly rules define expected conduct, and delegation concerns whether third parties have authority to interpret, monitor, or enforce those rules [9]. This framework shows that international governance exists on a spectrum rather than as a simple choice between binding and nonbinding law.

One option is a binding multilateral treaty. Such a treaty could establish common obligations concerning military uses of AI, incident reporting, model testing, or human control over high-risk decisions. Abbott and Snidal explain, however, that hard law creates higher negotiation costs, reduces state flexibility, and raises concerns about sovereignty, especially when future conditions remain uncertain [10]. These difficulties are especially relevant to frontier AI because states may resist permanent restrictions on a technology they consider essential to national security and economic growth.

A second option is soft law, including voluntary principles, political declarations, and codes of conduct. Soft law is generally easier to negotiate and revise, places fewer restrictions on state sovereignty, and can accommodate actors with different interests [10]. It may therefore be more realistic during the early stages of AI governance, when the technology is changing quickly and international consensus remains limited. Its weakness is that states and companies may endorse broad principles without changing their behavior when compliance conflicts with strategic or commercial interests.

A third option is a bilateral agreement between the United States and China. Writing in the trade context, Baucus argues that bilateral agreements can be negotiated more quickly, focus on the specific interests of the parties, and make progress when broader multilateral negotiations have stalled [11]. Applied to AI, a bilateral arrangement could focus on crisis communication, incident reporting, military safeguards, or shared commitments concerning nuclear-related decision-making. However, Baucus's analysis focuses on trade, and these advantages may not transfer fully to a field involving national security, private developers, and rapidly changing technology.

These models are not necessarily mutually exclusive. International AI governance could combine broad soft-law principles with narrower binding obligations, bilateral confidence-building measures, and domestic regulation of private developers. The literature, however, does not yet establish which combination of obligation, precision, and delegation would work best under the specific conditions of frontier AI and U.S.–China competition.

## 2.4. Literature gap and contribution

Existing scholarship explains why AI is difficult to regulate, how U.S.–China competition intensifies the problem, and why states choose different forms of international governance. However, these areas of research

are often discussed separately. AI-governance scholarship identifies regulatory problems without systematically comparing the available legal models. Geopolitical scholarship explains the causes of competition without determining which institutions could reduce the resulting escalation risks. Legalization theory provides tools for comparing hard and soft law, but much of it predates frontier AI and does not fully account for the central role of private companies.

This paper connects these areas of scholarship by comparing five governance models within a single analytical framework. It asks not only which model appears strongest in theory, but which combination could realistically operate under the current conditions. Its contribution is therefore to identify how different governance models can perform distinct functions and how they can be ordered so that each stage creates the conditions needed for the next.

## 3. Methodology

This paper uses a qualitative comparative case-study approach, supported by legal analysis, to evaluate five competing hypotheses about governing frontier AI (see Table 1). It draws on primary legal texts, official institutional documents, and secondary scholarship. The WTO represents binding multilateral law; the Tallinn Manual represents nonbinding expert guidance; SALT I and the ABM Treaty represent bilateral cooperation between strategic rivals; the OECD AI Principles, G7 Hiroshima AI Process, and EU AI Act represent the fragmented status quo; and outer-space governance represents a hybrid system combining treaties, narrower agreements, national regulation, and soft law (see Table 2).

**Table 1.** Hypotheses

| Code | Hypothesis |
|---|---|
| H1 | Binding Multilateral Governance: A binding multilateral legal framework can most effectively reduce frontier AI escalation risks by creating common obligations, precise rules, and delegated mechanisms for monitoring and enforcement. |
| H2 | Soft-Law Governance: Nonbinding principles, expert guidance, and confidence-building measures can most effectively reduce frontier AI escalation risks because they are easier to negotiate, revise, and adapt as the technology changes. |
| H3 | Bilateral Governance: A bilateral agreement between the United States and China can most effectively reduce frontier AI escalation risks by focusing on the two principal strategic competitors and establishing narrow, reciprocal safeguards. |
| H4 | Status Quo Governance: The current fragmented system of international principles, regional regulation, national law, and private-sector governance can effectively reduce frontier AI escalation risks without a new international legal framework. |
| H5 | Hybrid Governance: A hybrid model combining binding and nonbinding rules across multilateral, bilateral, regional, national, and private institutions can reduce frontier AI escalation risks more effectively than any single governance model. |
| Null | No governance model will significantly reduce frontier AI escalation risks because broader geopolitical, economic, military, and technological interests drive U.S.–China competition, while frontier AI develops too quickly and unpredictably for legal frameworks to remain effective. |

Each case is examined within its legal and political context, with attention to the wording of the relevant rules, how the framework operated in practice, and whether its main features could be applied to frontier AI. The comparison focuses on participation, flexibility, transparency, verification, enforcement, technical

expertise, private-sector involvement, and the ability to reduce uncertainty or miscalculation under conditions of rapid technological change and U.S.–China competition.

These cases are not treated as exact parallels to frontier AI. Instead, they are used to identify which institutional features remain useful under the conditions examined in this paper and where existing governance models may fall short.

**Table 2.** Case-study comparison

| Case Study | Model | Legal Form | Main Lesson |
|---|---|---|---|
| World Trade Organization (WTO) | H1: Binding Multilateral Governance | Binding multilateral law with delegated dispute settlement | Creates precise and enforceable obligations when conduct is observable, but depends on continued political support. |
| Tallinn Manual | H2: Soft-Law Governance | Nonbinding expert guidance | Clarifies how existing law applies to new technology and can be updated, but creates no binding obligations or enforcement mechanism. |
| SALT I: ABM Treaty and Interim Agreement | H3: Bilateral Governance | Binding bilateral treaty and interim agreement | Supports narrow and reciprocal limits between strategic rivals, but depends on verification and continued shared interest. |
| OECD AI Principles, G7 Hiroshima AI Process, and EU AI Act | H4: Status Quo Governance | International soft law, club-based guidance, and regional hard law | Provides useful standards for commercial and civilian AI, but leaves military escalation risks and international coverage uneven. |
| Outer-Space Governance | H5: Hybrid Governance | Multilateral treaties, narrower agreements, national regulation, and soft law | Addresses different risks through different layers, but requires coordination and broad participation to remain effective. |

# 4. Case studies

## 4.1. Case study 1: the WTO dispute settlement system as hard law

### *4.1.1. Introduction*

The World Trade Organization was established on January 1, 1995, after the Uruguay Round negotiations, which lasted from 1986 to 1994 and involved 123 governments. It replaced the institutional system built around the General Agreement on Tariffs and Trade, or GATT 1947. By the 1980s, the GATT dispute-settlement system was struggling because respondents could block the establishment of panels or the adoption of unfavorable reports. The WTO introduced broader trade rules and a more reliable process for resolving disputes. It now has 166 members, accounting for approximately 98 percent of world trade[1,2,3].

[1] World Trade Organization. (n.d.-b). Historic development of the WTO dispute settlement system: The system under GATT 1947 and its evolution over the years. World Trade Organization. Accessed August 3, 2026. https://www.wto.org/english/tratop_e/dispu_e/disp_settlement_cbt_e/c2s1p1_e.htm

[2] World Trade Organization. (n.d.-c). The Uruguay Round. World Trade Organization. Accessed August 3, 2026. https://www.wto.org/english/thewto_e/whatis_e/tif_e/fact5_e.htm.

[3] World Trade Organization. (n.d.-d). Who we are. World Trade Organization. Accessed August 3, 2026. https://www.wto.org/english/thewto_e/whatis_e/who_we_are_e.htm.

The WTO is one of the clearest examples of binding multilateral governance. Its members accept common legal obligations and agree to resolve disputes through an established institutional process. The Dispute Settlement Understanding, or DSU, describes this system as a "central element" in providing "security and predictability" to the multilateral trading system[4]. This case tests H1: whether binding multilateral governance can reduce uncertainty and unilateral action through clear rules, agreed procedures, and institutional oversight (see Table 1, H1).

*4.1.2. Legal analysis*

The legal foundation of WTO dispute settlement is the Understanding on Rules and Procedures Governing the Settlement of Disputes. The process begins with consultations, allowing members to resolve a dispute before formal adjudication. If consultations fail, the complaining member may request a panel. Article 6.1 requires the panel to be established unless the Dispute Settlement Body decides by consensus not to establish it[5].

The same negative-consensus rule applies to the adoption of panel reports. Under Article 16.4, a report must be adopted unless a party appeals or the Dispute Settlement Body decides by consensus not to adopt it[6]. This reversed the earlier GATT approach by preventing the losing party from blocking the panel process or report on its own.

The DSU also limits unilateral action. Article 23 requires members seeking redress to use the WTO dispute-settlement procedures rather than independently determining that another member has violated its obligations. Members must also obtain authorization from the Dispute Settlement Body before suspending trade concessions in response to noncompliance[7].

The system prioritizes compliance over punishment. Article 22 treats compensation and the suspension of concessions as temporary responses when a ruling has not been implemented. Neither is preferred to bringing the disputed measure into conformity with WTO law. Retaliation is permitted only after the dispute-settlement process has been completed and must be authorized through the WTO[8].

The original structure also included a standing Appellate Body that could uphold, modify, or reverse the legal findings of a panel. Together, consultations, panels, appellate review, and supervised remedies created a highly legalized system based on binding obligations, detailed procedures, and delegated decision-making[9].

*4.1.3. Analysis*

The record shows that members made substantial use of the WTO dispute-settlement system. Between 1995 and the end of 2024, members submitted 631 requests for consultations, and panels were established in 376 disputes. During the same period, 503 panel reports, Appellate Body reports, arbitral awards, or decisions were circulated or notified. Other disputes were resolved through consultations or settlement before a final ruling was required[10].

---

4 World Trade Organization. (1994). Understanding on rules and procedures governing the settlement of disputes. Annex 2 to the Marrakesh Agreement Establishing the World Trade Organization. https://www.wto.org/english/tratop_e/dispu_e/dsu_e.htm.

5 See note 4.

6 See note 4.

7 See note 4.

8 See note 4.

9 See note 4.

10 World Trade Organization. (n.d.-a). Dispute settlement activity—Some figures. World Trade Organization. Accessed August 3, 2026. https://www.wto.org/english/tratop_e/dispu_e/dispustats_e.htm.

The system's effectiveness depended on more than its binding legal structure. Consultations gave states space to settle disputes before adjudication, while automatic panel establishment and report adoption made it difficult for losing parties to avoid the process. WTO disputes also generally concerned observable government measures, such as tariffs, subsidies, regulations, or legislation. This made possible violations easier to identify and evaluate.

However, the system still depended on continued political support from its members. Since December 2019, the Appellate Body has been unable to hear new appeals because its vacancies have not been filled. As of December 2024, appeals in 31 proceedings remained unable to move forward. Panels and consultations have continued, but the absence of appellate review has weakened the finality of the system[11].

The WTO therefore provides partial support for H1. Binding multilateral law can create common obligations, limit unilateral action, and establish procedures for resolving disagreements. However, these advantages depend on broad participation, observable conduct, and continued support for the institutions applying the rules. These conditions are harder to achieve in frontier AI, where capabilities may be secret, technically difficult to measure, and controlled partly by private companies. A WTO-style framework would therefore be most useful for narrow and verifiable AI commitments, but it would be difficult to use as the only model for governing frontier AI.

## 4.2. Case study 2: the Tallinn Manual as expert soft law introduction

The Tallinn Manual was developed as cyber operations raised difficult questions about how existing international law applied to new technologies. The first edition, published in 2013, focused on the most serious cyber operations, including the use of force, self-defence, and armed conflict. Tallinn Manual 2.0, published in 2017, expanded the project to peacetime issues such as sovereignty, jurisdiction, state responsibility, and human rights. It contains 154 black-letter rules supported by detailed commentary [12, 13].

The Tallinn Manual is one of the clearest examples of nonbinding expert governance. The NATO Cooperative Cyber Defence Centre of Excellence describes it as a "non-legally-binding scholarly work" intended to provide an "objective restatement of international law as applied in the cyber context"[12]. This case tests H2: whether nonbinding expert guidance can reduce frontier AI escalation risks by clarifying how existing law applies to a rapidly developing technology (see Table 1, H2).

### *4.2.1. Legal analysis*

The Tallinn Manual begins from the position that cyber operations do not exist outside international law. Rule 4 of Tallinn Manual 2.0 states: "A State must not conduct cyber operations that violate the sovereignty of another State" [13]. Rule 14 provides that a state bears international responsibility for a cyber-related act when the act is attributable to that state and breaches an international legal obligation [13]. These rules apply established legal doctrines to cyber conduct rather than treating cyberspace as a separate legal system.

Each black-letter rule is followed by commentary explaining its legal basis, factual assumptions, and areas of disagreement among the experts. This structure shows not only the conclusion reached by the authors but also the reasoning behind it and the points on which no consensus exists.

The Manual's legal status, however, is deliberately limited. It represents the views of its authors rather than the legal position of NATO, the CCDCOE, participating states, or any international organization. It is therefore

[11] See note 10.

[12] NATO Cooperative Cyber Defence Centre of Excellence. (n.d.). The Tallinn Manual. NATO Cooperative Cyber Defence Centre of Excellence. https://ccdcoe.org/research/tallinn-manual/

not a treaty, an official statement of state practice, or an authoritative judicial interpretation.

This distinction matters because expert consensus is not the same as state consent. International legal rules generally derive their authority from treaties and customary international law, not from academic publication alone. The Manual may accurately identify an existing rule, but it cannot make a contested interpretation binding simply because a group of experts agrees with it.

The drafting process nevertheless seeks to make the Manual relevant to governments. The Tallinn Manual 3.0 revision process draws on state practice, official statements, international organizations, and scholarship. States are also invited to provide national perspectives, while the project commits to including "all reasonable views" on how international law applies in cyberspace[13]. The Manual therefore operates through clarification rather than command.

*4.2.2. Analysis*

The Manual's main strength is its ability to give governments, lawyers, and scholars a common framework for discussing cyber operations. Its commentary makes the reasoning behind each rule visible and identifies where interpretations differ. This turns broad legal uncertainty into more specific questions that can be debated and evaluated.

Its nonbinding form also allows it to respond more quickly than a formal treaty. Treaties require lengthy negotiation, ratification, and amendment, while expert guidance can be revised as technology, state practice, and legal interpretations change. For frontier AI, a similar process could update guidance concerning autonomous targeting, AI-assisted intelligence, cyber operations, and nuclear decision-making without reopening a treaty each time capabilities develop.

Soft-law guidance can also begin before political consensus exists. States may resist binding rules because they disagree over their meaning or fear losing strategic flexibility. An expert process can first identify competing interpretations and clarify where agreement may be possible. In AI governance, for example, states may support human oversight in principle but disagree over how much human involvement is required in military decision-making.

A comparable AI process could also bring legal and technical experts together to establish shared definitions and explain how existing law applies to complex systems. Like the Tallinn Manual, it could record the legal basis, technical assumptions, minority views, and unresolved questions behind each proposed rule instead of presenting contested interpretations as settled.

However, the Tallinn Manual operates through persuasion rather than legal authority. It cannot require states to disclose information, cooperate with investigations, accept its interpretations, or face consequences for noncompliance. States may rely on the Manual when its reasoning supports their interests and reject it when it does not. Clarification may expose disagreement without changing the strategic incentives behind it, particularly when secrecy provides an advantage.

Legitimacy presents a further challenge because states do not formally negotiate or adopt the Manual. An equivalent process for frontier AI would have limited influence if it were dominated by Western governments, companies, or experts. China and other states with significant AI capabilities would need meaningful participation for the resulting guidance to establish a credible common framework.

The case therefore supports a refined version of H2. Soft law is most useful when the central problem is uncertainty over terminology, legal application, and technical risk. It can build a common vocabulary, identify possible areas of agreement, and prepare states for later cooperation. However, it is much less effective when

[13] See note 12.

governance requires compulsory disclosure, verification, or enforcement. Expert soft law should therefore serve as an early stage of frontier AI governance rather than as a complete framework on its own.

## 4.3. Case study 3: SALT I and the ABM treaty as bilateral governance introduction

By the late 1960s, the United States and Soviet Union had spent more than two decades expanding their nuclear arsenals and improving the systems used to deliver them. The Soviet Union was increasing its intercontinental ballistic missile force to approach strategic parity with the United States, while both sides were developing anti-ballistic missile systems designed to intercept an incoming attack. Each new deployment was intended to improve national security, but it also encouraged the other side to expand or modernize its own forces[14].

Both governments therefore had a shared interest in limiting the most destabilizing parts of the arms race before it became even more expensive and dangerous. This did not mean that either side stopped pursuing strategic advantage. Rather, both recognized that continued competition could leave them less secure and increase the risk of a nuclear war with "devastating consequences for all mankind"[15]. The Strategic Arms Limitation Talks, or SALT I, began in 1969 and produced the 1972 ABM Treaty and Interim Agreement on strategic offensive arms. This case tests H3: whether bilateral governance can reduce escalation risks when two major powers share a vulnerability and can define narrow, reciprocal, and verifiable restraints (see Table 1, H3).

### *4.3.1. Legal analysis*

The ABM Treaty limited defensive systems designed to intercept strategic ballistic missiles. Article I required each party "not to deploy ABM systems for a defense of the territory of its country" or create a base for such a nationwide defense[16]. Article III permitted limited deployment in two areas: one around the national capital and one around an intercontinental ballistic missile field. These restrictions prevented either side from building a nationwide defense that could weaken the other's ability to retaliate. Article V also restricted future forms of missile defense. The parties agreed "not to develop, test, or deploy" sea-based, air-based, space-based, or mobile land-based ABM systems or components[17]. The Treaty therefore limited not only existing systems but also several ways in which either side might later expand its defensive capability.

The Interim Agreement placed narrower limits on offensive forces. Article I prohibited construction of additional fixed land-based intercontinental ballistic missile launchers after July 1, 1972. Article III limited submarine-launched ballistic missile launchers and modern ballistic missile submarines to existing and under-construction totals, while allowing agreed substitutions for older systems[18]. The accompanying protocol[19] permitted the United States up to 710 submarine-launched missile launchers on 44 submarines and the Soviet

---

[14] U.S. Department of State, Office of the Historian. (n.d.). Strategic arms limitations talks/treaty (SALT) I and II. Office of the Historian. https://history.state.gov/milestones/1969-1976/salt

[15] United States and Soviet Union. (1972c). Treaty between the United States of America and the Union of Soviet Socialist on the limitation of https://history.state.gov/historicaldocuments/frus1969-76v32/d316. Anti-Ballistic Missile Systems.

[16] See note 15.

[17] See note 15.

[18] United States and Soviet Union. (1972a). Interim agreement on certain measures with respect to the limitation of strategic offensive arms. https://history.state.gov/historicaldocuments/frus1969-76v32/d317.

[19] United States and Soviet Union. (1972b). Protocol to the interim agreement between the United States of America and the Union of Soviet Socialist Republics on certain measures with respect to the limitation of strategic offensive arms. https://history.state.gov/historicaldocuments/frus1969-76v32/d318.

Union up to 950 launchers on 62 submarines. These unequal ceilings reflected the forces each side already possessed rather than requiring exact numerical equality.

Verification relied on each country's own monitoring capabilities rather than international inspections. Article XII of the ABM Treaty and Article V of the Interim Agreement allowed both parties to use "national technical means of verification." They also agreed not to interfere with those systems or use "deliberate concealment measures" that would impede verification[20]. Each government could therefore evaluate compliance without requiring access to the other side's sensitive military facilities.

The agreements also created a process for handling uncertainty. Article XIII of the ABM Treaty established a Standing Consultative Commission to consider compliance questions, ambiguous situations, interference with verification, and possible changes to the agreement[21]. This gave both sides a regular channel for discussing concerns before treating them as deliberate violations.

The two agreements were designed differently in duration. The ABM Treaty was of unlimited duration, although either party could withdraw with six months' notice if extraordinary events threatened its "supreme interests"[22]. The Interim Agreement lasted for five years and required the parties to continue negotiations toward more complete limits on strategic offensive arms[23].

*4.3.2. Analysis*

SALT I made progress because it focused on the parts of the arms race that both governments considered especially destabilizing. The United States and Soviet Union did not need to trust each other or resolve their broader Cold War competition. They only needed to recognize that unrestricted growth in offensive missiles and missile defenses could leave both sides less secure.

Its limited scope also made agreement possible. The ABM Treaty addressed missile defense, while the Interim Agreement temporarily froze selected categories of offensive launchers. The parties did not attempt to regulate every nuclear weapon or resolve every disagreement over force structure. This allowed them to accept restraints in specific areas while continuing to negotiate over unresolved issues.

Reciprocity was important, but it did not require identical numerical limits. Each side accepted restrictions based on its existing force structure. This made the agreement politically possible while still preventing unlimited expansion. Applied to frontier AI, this suggests that the United States and China would not need identical domestic systems or technical capabilities.

They could instead accept comparable obligations, such as requiring human confirmation before nuclear use or notifying the other side of serious AI-related military incidents.

Verification was possible because the regulated systems, including launchers, submarines, radars, and test activity, produced physical signs that could be monitored from a distance.

Frontier AI is more difficult to observe because software can be developed secretly, updated quickly, and deployed through private companies. A bilateral AI framework would therefore need to rely partly on domestic regulation, developer recordkeeping, and national auditing rather than external monitoring alone.

[20] See note 15.

[21] See note 15.

[22] See note 15.

[23] See note 18.

The Standing Consultative Commission provides another useful lesson. A comparable U.S.–China mechanism could create a regular channel for discussing unexplained AI incidents, changes in military systems, possible interference with monitoring, and new technical risks. This would not eliminate disagreement, but it could prevent ambiguous conduct from immediately being interpreted as deliberate escalation.

However, SALT I limited launchers rather than directly reducing the total destructive power of either arsenal. The Interim Agreement did not restrict strategic bombers or the number of warheads carried by each missile. Both sides could therefore continue improving their forces by placing multiple warheads on existing missiles[24]. The agreements slowed selected parts of the arms race but did not stop technological competition.

Their effectiveness also depended on continued political support. The ABM Treaty lasted for almost thirty years, but the United States ultimately exercised its withdrawal right, with its withdrawal taking effect in June 2002[25]. Its precise and verifiable rules could preserve cooperation only while both sides continued to view the agreement as serving their strategic interests.

The case therefore provides partial support for H3. Bilateral governance can reduce escalation risks when two principal competitors identify a shared vulnerability and accept narrow, reciprocal, and verifiable restraints. For frontier AI, this makes U.S.–China coordination a realistic starting point, particularly for nuclear decision-making, incident communication, and military safeguards. However, AI capabilities are less observable, more rapidly updated, and controlled partly by private companies. A bilateral agreement would also exclude other states with significant AI capabilities. It should therefore serve as a foundation for broader governance rather than as a complete framework on its own.

## 4.4. Case study 4: the existing AI governance regime as the fragmented status quo introduction

Frontier AI is already governed through a growing but fragmented set of international principles, limited-group initiatives, regional law, national regulation, and company practices. The OECD AI Principles, the G7 Hiroshima AI Process, and the EU AI Act represent three different parts of this system. The OECD provides broad international guidance, the G7 develops more specific expectations for advanced AI developers, and the European Union imposes binding legal duties on systems and models connected to its market.

Together, these instruments show that the status quo is not an absence of governance. Different institutions already address safety, transparency, accountability, and risk management through different forms of law. However, they do not have the same participants, legal force, scope, or enforcement mechanisms. This case tests H4: whether the existing combination of international soft law, limited-group cooperation, and regional regulation can reduce frontier AI escalation risks without a new international legal framework (see Table 1, H4).

### *4.4.1. Legal analysis*

The OECD AI Principles provide the broadest international layer of the current system. First adopted in 2019 and updated in 2024, they promote AI that is innovative and trustworthy and that respects human rights and democratic values. The principles address transparency and explainability, robustness and safety, and accountability, while also encouraging governments to cooperate on standards and compatible policy

[24] See note 14.

[25] White House. (2002, June 13). Statement by the president. https://georgewbushwhitehouse.archives.gov/news/releases/2002/06/20020613-9.html.

frameworks. The OECD describes the Recommendation as the first intergovernmental standard on AI and records 47 adherents, including the United States and the European Union[26,27].

The principles ask AI actors to provide meaningful information about their systems and manage risks throughout the AI lifecycle. They also state that systems that create an undue risk of harm should be able to be overridden, repaired, or safely withdrawn from use. The 2024 revision added greater attention to general-purpose and generative AI, information integrity, safety, privacy, and intellectual property. This shows that the framework can be updated as technology changes[28,29].

However, the OECD Principles are nonbinding. They do not require states to adopt the same domestic laws, give an international institution authority to inspect AI developers, or create legal penalties for noncompliance. Their main function is to establish common concepts that governments can use when developing their own laws and policies.

The G7 Hiroshima AI Process moves from broad principles toward more specific expectations for organizations developing advanced AI systems. In 2023, G7 leaders endorsed International Guiding Principles and an International Code of Conduct. The documents call on developers to identify and reduce risks throughout the AI lifecycle, report serious incidents, strengthen security controls, publish transparency information, and support the development of technical standards[30,31].

The Code is voluntary and designed as a living document that can change as technology develops. This allows it to address advanced AI developers directly without waiting for a universal treaty or domestic legislation in every participating state.

In 2025, the OECD launched a voluntary reporting framework to support implementation of the Hiroshima Code. The framework allows organizations to publish comparable information concerning risk assessment, incident reporting, information sharing, security, and other governance practices. It therefore adds a monitoring and transparency mechanism to the original Code without converting it into binding law[32].

Its reach nevertheless remains uneven. Participation in the reporting framework is voluntary, and neither the Code nor the reporting process creates an inspection body or common penalty. China is not a member of the G7, and the Hiroshima Process does not establish reciprocal obligations between the United States and China. It can encourage responsible company practices, but it cannot compel every major developer or government to participate.

---

[26] OECD. (2024b). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[27] OECD. (n.d.). OECD AI principles overview. OECD.AI. https://oecd.ai/en/ai-principles.

[28] See note 26.

[29] OECD. (2024a, May 3). OECD updates AI principles to stay abreast of rapid technological developments. OECD. https://www.oecd.org/en/about/news/press-releases/2024/05/oecd-updates-ai-principles-to-stay-abreast-of-rapidtechnological-developments.html.

[30] G7. (2023a, October 30). Hiroshima process international code of conduct for organizations developing advanced AI systems. Ministry of Foreign Affairs of Japan. https://www.mofa.go.jp/files/100573473.pdf.

[31] G7. (2023b, October 30). Hiroshima process international guiding principles for organizations developing advanced AI systems. Ministry of Foreign Affairs of Japan. https://www.mofa.go.jp/files/100573471.pdf.

[32] OECD. (2025, February 7). OECD launches global framework to monitor application of G7 Hiroshima AI code of conduct. OECD. https://www.oecd.org/en/about/news/press-releases/2025/02/oecd-launches-global-framework-tomonitor-application-of-g7-hiroshima-ai-code-of-conduct.html.

The EU AI Act is the strongest binding part of the current system. Regulation (EU) 2024/1689 applies to providers placing AI systems or general-purpose AI models on the EU market, including certain providers established outside the European Union. Its market-based scope allows the law to influence global companies that want to operate in Europe[33].

The Act combines prohibited practices with detailed obligations for higher-risk systems. Articles 9–15 establish requirements concerning risk management, data governance, technical documentation, record keeping, transparency, human oversight, accuracy, robustness, and cybersecurity. The Act also creates separate obligations for providers of general-purpose AI models and additional duties for models classified as creating systemic risk[34].

The Act entered into force on August 1, 2024, but applies in stages. Prohibited practices and AI-literacy duties began applying in February 2025, while governance rules and obligations for general-purpose AI models began applying in August 2025. General enforcement powers and transparency obligations began applying on August 2, 2026. Following the July 2026 AI Omnibus, important high-risk-system rules will apply later: December 2027 for specified high-risk uses and August 2028 for systems embedded in regulated products[35].

Unlike the OECD and G7 instruments, the EU AI Act is backed by public enforcement and financial penalties. Article 99 authorizes significant administrative fines for violations. The Act therefore turns some principles of AI safety and accountability into enforceable legal obligations[36].

However, Article 2 excludes AI systems used exclusively for military, defence, or national-security purposes. The Act can regulate commercial systems and general-purpose models connected to the European market, but it does not directly govern many of the military uses most closely connected to geopolitical escalation[37].

*4.4.2. Analysis*

The fragmented status quo has several strengths. It allows different institutions to act without waiting for one universal agreement. The OECD can revise broad principles, the G7 can develop and test reporting practices for advanced developers, and the European Union can impose binding rules through its market. This division of labor allows the system to address different parts of AI governance through different legal forms.

The current framework also reaches private companies more directly than many traditional international agreements. The Hiroshima Code and reporting framework speak directly to advanced AI developers, while the EU AI Act imposes legal duties on providers and can influence companies outside Europe that want access to its market.

However, the existence of several layers does not make them a coordinated system. The three frameworks differ in membership, legal force, scope, and enforcement. China is outside the G7 process, the OECD Principles do not create reciprocal obligations, and the EU AI Act does not establish shared rules for strategic

---

[33] European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[34] See note 33.

[35] European Commission. (2026). "AI Act." strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Space Accessed August 3, 2026. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[36] See note 33.

[37] See note 33.

conduct between the United States and China. None of the three institutions has authority to resolve conflicts or close gaps between the separate frameworks.

The largest gap concerns military escalation. Existing instruments focus mainly on commercial safety, individual rights, transparency, and responsible development. They do not create U.S.–China crisis-communication channels, reciprocal incident-notification duties, verification of military safeguards, or common limits on AI use in nuclear and early-warning systems. The EU AI Act also excludes systems used exclusively for military, defence, or national-security purposes.

The case therefore rejects H4 as stated. The fragmented status quo makes an important contribution to civilian and commercial AI governance, particularly by establishing shared principles and regulating private developers. However, it cannot by itself reduce the reciprocal military escalation risks at the center of U.S.–China frontier AI competition. Additional bilateral safeguards and broader international coordination are still required.

## 4.5. Case study 5: outer-space governance as a hybrid model introduction

Outer-space governance did not develop through one comprehensive treaty or institution. As space activity expanded during the Cold War, states first established broad principles and later added more specific rules concerning liability, registration, international cooperation, private operators, and newer forms of exploration. The resulting system combines binding multilateral treaties, program-specific agreements, domestic regulation, and nonbinding principles.

The 1967 Outer Space Treaty remains the legal foundation of this system and had 118 parties as of January 1, 2026[38]. Later agreements address narrower problems, while national law and initiatives such as the Artemis Accords adapt these principles to new actors and activities. This case tests H5: whether different legal instruments can perform complementary functions within one broader system more effectively than any single governance model (see Table 1, H5).

### *4.5.1. Legal analysis*

The Outer Space Treaty establishes the system's broad multilateral principles. Article I provides that outer space is free for exploration and use by all states, while Article II prohibits national appropriation through sovereignty, occupation, or other means. Article IV prohibits placing nuclear weapons or other weapons of mass destruction in orbit or on celestial bodies and reserves the Moon and other celestial bodies for peaceful purposes[39]. These provisions establish durable limits without attempting to regulate every possible space activity.

Article VI connects private activity to state responsibility. States "shall bear international responsibility" for national space activities, including those conducted by private entities.

Non-governmental activity also requires state "authorization and continuing supervision"[40]. Private companies may therefore conduct space activities, but states remain responsible for ensuring that those activities comply with international law.

---

[38] United Nations Office for Outer Space Affairs. (2026). Status of international agreements relating to activities in outer as at January 1, 2026. A/AC.105/C.2/2026/CRP.9/Rev.1. https://www.unoosa.org/res/oosadoc/data/documents/2026/aac_105c_22026crp/aac_105c_22026crp_9rev_1_0_html/AC105_C2_2026_CRP09Rev01E.pdf.

[39] United Nations. (1967). Treaty on principles governing the activities of states in the exploration and use of outer space, the Moon and other celestial bodies. https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/outerspacetreaty.html. 610 UNTS 205.

[40] See note 39.

Article IX addresses the risk that one state's activities may interfere with another's. It requires states to act with "due regard" for the interests of others and to undertake consultations when a planned activity may cause "potentially harmful interference"[41]. The Treaty does not establish a central regulator or automatic dispute-settlement process, but it creates an obligation to communicate before a foreseeable risk develops into a larger dispute.

Later treaties translated some of these broad principles into more specific rules. The Liability Convention distinguishes between different forms of damage. A launching state is absolutely liable for damage caused by its space object on Earth or to aircraft in flight, while liability for damage occurring elsewhere depends on fault[42]. This gives greater precision to the general liability principle established by Article VII of the Outer Space Treaty.

The Registration Convention adds a transparency mechanism. It requires each state of registry to maintain a national register and provide the United Nations with identifying and orbital information about launched objects, including their launch date, basic orbital parameters, and general function[43]. Registration makes objects and the states connected to them more observable, although the system still depends on states providing accurate information.

The Moon Agreement attempted to create more detailed rules for activities on the Moon, particularly the future exploitation of natural resources. Article XI describes the Moon and its natural resources as the "common heritage of mankind" and calls for an international regime to govern exploitation when it becomes feasible[44].

However, the Moon Agreement has only 17 parties and has not been joined by the United States, China, or Russia[45]. Its limited participation shows that greater legal precision does not necessarily create more effective governance. Detailed hard law has little practical influence when the states most capable of conducting the regulated activity do not accept its obligations.

Narrower cooperative agreements provide another layer. The International Space Station framework combines a binding intergovernmental agreement with agency-level memoranda and implementing arrangements. The intergovernmental agreement establishes the overall legal framework, while the more detailed arrangements allocate operational and technical responsibilities among the participating agencies[46]. This structure allows states to cooperate on a defined activity without negotiating a universal treaty for every operational question.

---

41 See note 39.

42 United Nations. (1972). Convention on international liability for damage caused by space objects. 961 UNTS 187. https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/liability-convention.html.

43 United Nations. (1975). Convention on registration of objects launched into outer space. 1023 UNTS 15. https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/registration-convention.html.

44 United Nations. (1979). Agreement governing the activities of states on the moon and other celestial bodies. 1363 UNTS 3. https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/moon-agreement.html.

45 United Nations Treaty Collection. (n.d.). Agreement governing the activities of states on the moon and other celestial bodies. Retrieved August 3, 2026, from https://treaties.un.org/pages/ViewDetails.aspx?chapter=24&clang=_en&mtdsg_no=XXIV-2&src=TREATY.

46 International Space Station Intergovernmental Agreement. 1998. Agreement among the Government of Canada, Governments of Member States of the European Space Agency, the Government of Japan, the Government of the Russian Federation, and the Government of the United States of America concerning Cooperation on the Civil International Space Station. https://www.state.gov/wp-content/uploads/2019/02/12927-Multilateral-Space-Space-Station-1.29.1998.pdf.

National regulation implements the international rules governing private actors. Because Article VI requires authorization and continuing supervision, states use domestic laws and licensing systems to regulate commercial launches and other private space activities.

International law establishes the responsibility, but national institutions determine the more detailed licensing, oversight, and enforcement requirements.

The Artemis Accords add a nonbinding and limited-group layer. NASA and the U.S. Department of State established the Accords with seven other initial signatories in 2020. By July 2026, the initiative had expanded to 70 signatory states[47]. The Accords reaffirm the Outer Space Treaty and develop more operational expectations concerning transparency, interoperability, registration, scientific data, space resources, harmful interference, safety zones, and orbital debris[48].

The Accords are particularly important because they translate Article IX's general consultation principle into more specific practices. Signatories intend to disclose the location and general nature of their operations and coordinate through temporary safety zones to avoid harmful interference. These commitments can develop without formally amending the Outer Space Treaty.

However, the Accords remain political commitments rather than binding international law. They apply among participating states and do not represent a universally accepted interpretation of space law. Their rapid expansion demonstrates the flexibility of limited-group governance, but the absence of some major space powers also creates concerns about legitimacy and competing legal interpretations.

*4.5.2. Analysis*

Outer-space governance shows how different legal forms can perform different functions within one broader system. The Outer Space Treaty establishes stable principles, the Liability and Registration Conventions regulate narrower and more observable problems, cooperative agreements manage particular relationships, national law reaches private companies, and the Artemis Accords develop more adaptable operational practices.

This division of labor combines stability with flexibility. Rules concerning sovereignty, weapons of mass destruction, and state responsibility can remain durable, while registration practices, technical standards, and operational coordination can develop as technology and commercial activity change. No single institution is expected to regulate every actor or anticipate every future activity.

The regime also provides a practical method for connecting private actors to international obligations. Article VI places responsibility on states, while domestic licensing allows national governments to impose more detailed requirements on companies operating under their jurisdiction. This preserves a clear line of legal responsibility even as commercial actors play a larger role in space activity.

The differences between the Outer Space Treaty, Moon Agreement, and Artemis Accords also reveal a persistent tradeoff. Broad principles can attract extensive participation but may leave important questions unresolved. More precise treaties may fail to attract the states whose conduct matters most. Nonbinding initiatives can expand and adapt more quickly, but they may produce competing coalitions and interpretations.

---

[47] NASA. 2026. “NASA Welcomes Mauritius as 70th Artemis Accords Signatory.” July 17, 2026. https://www.nasa.gov/organizations/oiir/artemis-accords/nasa-welcomes-mauritius-as-70th-artemis-accords-signatory/.

[48] Artemis Accords. 2020. The Artemis Accords: Principles for Cooperation in the Civil Exploration and Use of the Moon, Mars, Comets, and Asteroids for Peaceful Purposes. https://www.nasa.gov/wp-content/uploads/2022/11/Artemis-Accords-signed-13Oct2020.pdf.

Hybrid governance therefore does not eliminate fragmentation automatically. Different instruments may use inconsistent definitions, include different participants, or leave responsibility unclear. National implementation may also vary, while overlapping institutions may create disputes over which rule or interpretation should control.

This is the main distinction between H4 and H5. The fragmented status quo contains multiple layers that often operate independently. A functioning hybrid regime requires those layers to have clearly allocated and complementary roles. The International Space Station framework illustrates this relationship by placing technical arrangements beneath a common legal agreement, while the Artemis Accords expressly connect their operational principles to existing space treaties.

The case therefore supports H5, subject to an important condition. Hybrid governance is more effective than a single model when multilateral law provides common principles, narrower agreements address specific risks, national law reaches private actors, and soft law supplies adaptable technical detail. The layers must also reinforce one another and include mechanisms for identifying gaps or conflicting rules.

For frontier AI, the lesson is not simply to create more institutions. A successful hybrid framework would need to specify which risks belong at the multilateral, bilateral, national, and technical levels and how those rules interact. Without that coordination, a hybrid model could reproduce the same fragmentation identified under H4.

## 4.6. Cross-cutting analysis

The five cases do not identify one legal model that performs best across every dimension of frontier AI governance. Instead, they show that the problem contains several distinct tasks: securing participation from the states whose conduct matters most, adapting rules as the technology changes, supervising private developers, verifying compliance, and preventing miscalculation between strategic rivals. Historical regimes usually perform some of these tasks well and others poorly. The central question is therefore not which legal form is strongest in the abstract, but which institution and instrument is best suited to each task and what conditions must exist before governance can become more binding.

First, broad participation and legal strength are not substitutes for one another. The WTO shows that binding multilateral law can create common obligations, delegated review, and enforceable procedures when major states accept the system. Its appellate crisis also shows that legal design cannot preserve cooperation when political support disappears. The Moon Agreement demonstrates the opposite problem: detailed legal rules have little practical effect when the states with the greatest relevant capabilities remain outside the agreement. For frontier AI, a framework that excludes either the United States or China would fail to govern the strategic relationship currently creating the most immediate escalation risk. However, a permanent two-state arrangement would also lack broader legitimacy and leave other capable states outside the system.

Second, the precision of a rule must correspond to the observability of the conduct it regulates. WTO obligations often concern public laws or trade measures, while SALT I regulated launchers, submarines, radars, and testing activity that produced physical evidence. Frontier AI capabilities are less visible. They can change through software updates, new training methods, access decisions, or modifications inside private companies without producing clear external signs. A treaty attempting to limit model intelligence or total national AI capability would therefore be difficult to verify. More credible rules would initially focus on observable conduct and procedures, including human authorization, incident notification, crisis communication, domestic reporting, recordkeeping, and noninterference with agreed monitoring.

Third, stable legal commitments should be separated from technical rules that require frequent revision. The Tallinn Manual, OECD AI Principles, and Hiroshima AI Process show that nonbinding instruments can

clarify terminology, record competing interpretations, and update technical expectations as knowledge develops. Their flexibility makes them useful for identifying and testing possible safeguards. However, they cannot compel disclosure, verification, or compliance when cooperation conflicts with strategic or commercial interests. Soft law is therefore most useful for technical definitions, evaluation methods, incident categories, and emerging practices, while binding law should be reserved for durable obligations whose underlying purpose is unlikely to change with each new generation of technology.

Fourth, international commitments depend on domestic regulatory capacity. Outer-space governance connects international responsibility to national authorization and continuing supervision of private operators, while the EU AI Act imposes direct obligations on providers operating within or entering its market. Frontier AI requires a similar connection because private developers control the models, computing infrastructure, technical expertise, and much of the information governments would need to evaluate risk. A state cannot make a credible international commitment if it cannot identify relevant developers, require evaluations, investigate serious incidents, protect sensitive evidence, or punish false disclosure within its own jurisdiction. Companies should contribute technical knowledge and comply with domestic rules, but they should not determine the public standards by which their own conduct is judged.

Fifth, bilateral governance is most useful when strategic competitors share a specific vulnerability but do not share broader political trust. SALT I did not end the Cold War or eliminate technological competition. It succeeded because the United States and Soviet Union isolated a limited set of risks, accepted reciprocal restraints, and created a continuing process for consultation. A comparable U.S.–China process could address crisis communication, serious AI incidents, human control over strategic decisions, and risks involving early-warning or nuclear systems without requiring either side to resolve its wider economic and military competition.

Bilateral governance cannot serve as the complete framework, but it can establish safeguards in the relationship where miscalculation is currently most dangerous and later provide a foundation for broader participation.

The evidence therefore produces a qualified judgment across the five hypotheses. H1 is supported only for narrow obligations that are sufficiently clear and observable; binding multilateral law is not a realistic starting point for governing frontier AI as a whole. H2 is strongly supported for clarification, experimentation, and adaptation, but not for compulsory disclosure or enforcement. H3 receives partial support because bilateral governance can address immediate U.S.–China escalation risks when commitments are narrow, reciprocal, and supported by permanent consultation. H4 is rejected as an adequate response to military escalation, although the existing regime remains important for civilian and commercial governance. H5 receives the strongest support, but only when the different layers have clearly allocated and complementary functions. The cases also reject the null hypothesis as stated. Governance cannot eliminate the strategic competition driving AI development, but it can reduce specific pathways through which uncertainty, technical failure, or misinterpretation could lead to escalation.

These findings support a staged hybrid model rather than an unstructured combination of existing institutions. Domestic law must first give governments reliable information and authority over frontier developers. Bilateral institutions can then establish communication channels, shared definitions, and procedures for managing incidents. Nonbinding commitments can test safeguards while technology and technical knowledge continue to change. Binding law should follow only when an obligation has become sufficiently clear, politically acceptable, and observable, with later participation open to other states possessing significant AI capabilities.

The next section develops how these layers should be ordered and what conditions should determine movement from one stage to the next.

## 5. Recommendations

The evidence supports a four-stage hybrid model. Its long-term objective should be a binding U.S.–China baseline for the most dangerous military uses of frontier AI, especially those involving nuclear weapons, early-warning systems, and strategic decision-making. Such an agreement cannot be the starting point. Neither government has complete information about every frontier system developed within its jurisdiction. Mutual trust is limited, and the technology continues to change. Governance should therefore begin with domestic oversight, move through permanent bilateral consultation and nonbinding safeguards, and become legally binding only when specific obligations are clear and observable.

### 5.1. Stage one: build domestic oversight

Before either government can negotiate meaningfully with the other, it needs reliable information about the systems being developed at home. Major developers, such as OpenAI and Anthropic in the United States and DeepSeek and ByteDance in China, should be required to report model capabilities, safety testing, serious incidents, cybersecurity risks, and data-collection methods to national regulators.

In the United States, Congress should authorize the Department of Commerce to enforce these requirements. The Center for AI Standards and Innovation, or CAISI, could design common evaluations and test whether companies' results are accurate[49]. In China, the State Council could assign the Cyberspace Administration of China, working with the Ministry of Industry and Information Technology, to perform a similar role. China could expand its existing safety-assessment and algorithm-filing system to cover frontier models even when they are not directly offered to the public[50].

The rules should cover models that exceed a defined level of computing power or demonstrate advanced cyber, biological, autonomous, or military capabilities. This threshold should be reviewed every year so that companies cannot avoid oversight simply because the technology has changed.

Developers should submit a report before releasing a covered model and notify regulators within 72 hours of discovering a serious incident. This could include the theft of model weights, a model bypassing its safety controls, or evidence that it provided meaningful assistance for an advanced cyber or biological attack. Companies should also retain test results, internal incident reports, access records, and major deployment decisions so that regulators can check their claims.

Regulators should audit each covered developer at least once a year and after the release of a major new model. They should be able to conduct secure evaluations, inspect relevant records, and penalize false or incomplete reporting. Companies would not need to publish sensitive technical information, but they could not withhold it from the government responsible for regulating them.

This stage would be complete when both governments can identify their major frontier developers, receive reports in a standard format, independently evaluate important capabilities, and respond to noncompliance.

---

[49] National Institute of Standards and Technology. 2026. “CAISI Evaluation of DeepSeek V4 Pro.” May 1, 2026. https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro.

[50] Cyberspace Administration of China. 2023. Seven Departments, Including the Cyberspace Administration of China, Jointly Issue the Interim Measures for the Management of Generative Artificial Intelligence Services.” July 13, 2023. https://www.cac.gov.cn/2023-07/13/c_1690898326795531.htm.

### 5.2. Stage two: create a permanent U.S.–China Commission

The U.S. Department of State and China's Ministry of Foreign Affairs should establish a Standing U.S.–China Frontier AI Commission through a bilateral memorandum. The Commission should have permanent staff, named emergency contacts, and working groups that meet throughout the year. It should not depend on occasional presidential summits.

The Commission should include diplomats, national AI regulators, defense officials, military experts, and government scientists from both countries. Companies such as OpenAI, Anthropic, DeepSeek, and ByteDance could advise the technical working groups, but they should not set the rules or decide whether they have complied with them.

The Commission should have separate commercial and military tracks. The commercial group should compare how the two countries evaluate model capabilities, classify serious incidents, protect model weights, and audit company reports. The goal would not be to make U.S. and Chinese law identical. It would be to ensure that both governments understand how the other side supervises its companies.

The military group should focus on situations that could lead to escalation. It should examine how AI is used in early-warning systems, nuclear command structures, intelligence analysis, and military targeting. It should also identify failures that could be mistaken for hostile action, such as a false missile warning produced by an AI system or the corruption of data used in a strategic assessment.

During its first year, the Commission should produce a shared glossary, a list of reportable incidents, a secure notification form, and a procedure for emergency contact. Both sides should exchange only the information needed to explain an incident. They should not be required to disclose complete model designs, intelligence sources, or classified military capabilities.

This stage would be complete when the Commission has permanent contact points, a secure communication channel, shared definitions, and a working procedure for reporting and discussing incidents.

### 5.3. Stage three: test nonbinding safeguards

The Commission should then draft nonbinding protocols for commercial and military AI risks. The U.S. Department of State and China's Ministry of Foreign Affairs should approve the protocols, while domestic regulators and defense authorities should carry them out.

The commercial protocol should create common categories for serious incidents and a confidential process for notifying the other government when an incident could have cross-border effects. For example, notification may be necessary if model weights are stolen by a foreign actor, a widely used model develops a serious cybersecurity weakness, or a company discovers a capability that could create an unexpected national-security risk. The first notice could be sent within 24 hours, followed by a fuller report after the domestic investigation.

The military protocol should include four initial safeguards. First, an AI system should never independently authorize the use of nuclear weapons. Second, an identified human official should make the final launch decision. Third, neither government should rely only on AI-generated warning or targeting information when making that decision. Fourth, each side should notify the other of an AI failure or cyber incident that could reasonably be misunderstood as preparation for an attack.

These rules should be tested through exercises every six months. One exercise could simulate a false missile warning generated by corrupted data. Another could test how the two sides would communicate after the theft of a powerful model. Each exercise should end with a joint review identifying unclear definitions, slow communication, or conflicting interpretations. The Commission should then revise the protocol.

The protocols should remain nonbinding during this stage because their technical details may need to change. However, implementation should still be measured. Each government should report whether it has

created the required contact points, reporting procedures, and human-authorization rules. Publication of another general principle would not count as progress. Movement to the final stage should occur only after both sides have completed at least two exercises, used the same rules through a full review cycle, and shown that the safeguards can be enforced through their domestic systems.

### 5.4. Stage four: adopt a narrow binding treaty

Once the safeguards have been tested, the two governments should convert the clearest military rules into a binding treaty. The U.S. Department of State and China's Ministry of Foreign Affairs should lead the negotiations and submit the final agreement for the domestic approval required in each country.

The treaty should prohibit an AI system from independently authorizing nuclear-weapons use. It should require an authorized human official to make and record the final decision. It should also prohibit either government from relying only on AI-generated warning or targeting information and require notification of serious incidents involving nuclear command, early-warning systems, or strategic communications.

Verification should focus on whether these procedures exist and are followed, not on trying to measure the total strength of either country's AI systems. Each government should place the treaty rules into its written command procedures, record relevant AI recommendations and human decisions, retain those records securely, and complete an annual compliance review.

If a serious incident occurs, a national review team outside the military unit involved should examine the records. The government should then provide the Commission with an agreed set of facts, while protecting classified information. Neither side would receive automatic access to the other's source code or military facilities. However, destroying required records or interfering with agreed monitoring should count as a treaty violation.

The Standing Commission should become the treaty's consultative body. Either government should be able to request an emergency meeting when it suspects a violation or cannot explain an incident. This would give both sides a chance to examine technical failure, human error, and deliberate misconduct before treating the event as hostile action.

The main obligations should remain in the treaty. Technical definitions, notification forms, and reporting thresholds should appear in an annex reviewed every two years. The Commission could update the annex with the approval of both governments without reopening the entire treaty.

The treaty should also allow other states with significant frontier AI or nuclear capabilities to join after the U.S.–China system has operated successfully. Bilateral cooperation would therefore serve as the starting point for broader governance, not its final form.

The four stages can overlap, but they cannot be reversed. Domestic oversight gives governments reliable information. The Commission creates regular communication. Nonbinding protocols test whether the proposed safeguards work. Binding law preserves the rules that have become clear, practical, and verifiable.

## 6. Conclusion

This paper acknowledges its limited access to the most advanced AI systems. Much of the relevant information on model capabilities, military uses, safety testing, and government–company cooperation remains confidential or classified. Because the analysis relies mainly on public sources, it may not fully reflect the actual frontier of AI development or the risks already emerging behind closed doors.

The proposed staged approach also risks moving more slowly than the technology itself. Domestic regulation, bilateral consultation, and treaty negotiations may take years, while frontier AI can change within

months. However, a faster binding agreement is unlikely under current U.S.–China relations. Low trust, strategic competition, and uncertainty about future capabilities make an immediate comprehensive treaty politically unrealistic. The staged approach is therefore not ideal, but it may be the most serious option available under present conditions.

This paper treats U.S.–China cooperation as a starting point, not a permanent two-country system. Because their competition currently creates the most immediate escalation risk, testing safeguards between them could create a solid foundation for broader cooperation. Other states with major AI, semiconductor, cloud-computing, or nuclear capabilities could later join technical discussions, adopt compatible domestic rules, and accept the same military safeguards. If the major players change, participation should change with them. The framework is based on capabilities and risks rather than a fixed list of countries, so its central argument would remain relevant even if the United States and China no longer lead frontier AI development.

Further research depends on better access to information. AI companies should provide more reliable data on model capabilities, safety testing, and serious incidents to regulators.

Universities and independent researchers should also build structured partnerships with developers to study advanced systems under controlled conditions. These partnerships could protect commercial and national-security information while still allowing outside experts to test whether proposed safeguards are technically realistic. Without stronger evidence, future governance will continue to be designed around an incomplete picture.

## 7. Artificial intelligence use disclosure

This paper was completed with the assistance of ChatGPT. AI was used for citation checking and grammar correction for the following sections: Introduction, Literature Review, Methodology, Hypotheses, Case Studies, Cross-Cutting Analysis, Recommendations, and Conclusion. All AI-generated materials have been checked and verified by the author. The author assumes responsibility for any errors and mistakes generated by AI.

## References

[1] Cyphert, A. B. (2025). Confronting the challenges of regulating artificial intelligence. *FIU Law Review, 20*(1), 82–117. https://doi.org/10.25148/lawrev.20.1.5

[2] Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). The global governance of artificial intelligence: Next steps for empirical and normative research. *International Studies Review, 25*(3), Article viad040. https://doi.org/10.1093/isr/viad040

[3] Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. *International Affairs, 100*(3), 1275–1286. https://doi.org/10.1093/ia/iiae073

[4] Cheng, J., & Zeng, J. (2023). Shaping AI's future? China in global AI governance. *Journal of Contemporary China, 32*(143), 794–810. https://doi.org/10.1080/10670564.2022.2107391

[5] Schmid, S., Lambach, D., Diehl, C., & Reuter, C. (2025). Arms race or innovation race? Geopolitical AI development. *Geopolitics, 30*(4), 1907–1936. https://doi.org/10.1080/14650045.2025.2456019

[6] Landwehr-Matlé, J., Oppermann, K., & Lambach, D. (2026). Great power competition for global leadership in artificial intelligence (AI): Reconstructing AI narratives of the United States, China, and the European Union. *Review of International Studies, 1*–19. https://doi.org/10.1017/S0260210526101776

[7] Chernavskikh, V., & Palayer, J. (2025). *Impact of military artificial intelligence on nuclear escalation risk* (SIPRI Insights on Peace and Security No. 2025/6). Stockholm International Peace Research Institute.

https://doi.org/10.55163/FZIW8544

[8] Johnson, J. (2023). Inadvertent escalation: A new model for nuclear risk. In *AI and the bomb: Nuclear strategy in the digital age* (pp. 111–153). Oxford University Press. https://doi.org/10.1093/oso/9780192858184.003.0004

[9] Abbott, K. W., Keohane, R. O., Moravcsik, A., Slaughter, A.-M., & Snidal, D. (2000). The concept of legalization. *International Organization, 54*(3), 401–419. https://doi.org/10.1162/002081800551271

[10] Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International Organization, 54*(3), 421–456. https://doi.org/10.1162/002081800551280

[11] Baucus, M. (1989). A new trade strategy: The case for bilateral agreements. *Cornell International Law Journal, 22*(1), 1–24. https://scholarship.law.cornell.edu/cilj/vol22/iss1/1

[12] Schmitt, M. N. (Ed.). (2013). *Tallinn manual on the international law applicable to cyber warfare*. Cambridge University Press. https://doi.org/10.1017/CBO9781139169288

[13] Schmitt, M. N. (Ed.). (2017). *Tallinn manual 2.0 on the international law applicable to cyber operations*. Cambridge University Press. https://doi.org/10.1017/9781316822524