International Agreements to Limit Frontier AI: Objectives and Exit

Lennart Finke [1 2 3]

Abstract

An international agreement to limit AI development could be crucial to mitigate risks from AI. However, it remains unclear which conditions should determine when the limiting measures are relaxed. We survey existing international agreements, outline what properties appropriate conditions should satisfy, list possible conditions, and finally give a recommendation in an example scenario. We recommend a fixed time period after which a new organization established at the start of the period specifies conditions that address when AI development can be safely conducted, with a possibility of withdrawal in extraordinary circumstances. We hope to illustrate the considerations that would likely go into an international agreement to limit AI.

1. Introduction

Both large risks and large gains are expected to stem from the development of increasingly powerful AI. To lower the risks from future AI systems, some have put forward proposals to limit the development of frontier AI, such as Scher et al. (2025); Miotti & Wasil (2023); Alaga & Schuett (2023); Scholefield et al. (2025). Public support for limiting AI development appears plausible in countries where data is available, such as the United States and the United Kingdom (Elsey & Moss, 2023; Perrigo, 2025; Bryson Gillette Insights, 2025), though data in most countries is sparse. However, broad restrictions of AI development would forgo potential benefits of AI development, which leads most proposals to limit AI development to envision an end of the agreement, as in Aschenbrenner & Trammell (2024). Here we investigate how the signatories that enter into the agreement will likely think about the path towards the end of the agreement.

[1]ETH Zürich, Zurich, Switzerland [2]Harvard University, Cambridge, United States [3]MATS Research, Berkeley, United States. Correspondence to: Lennart Finke <lfinke@ethz.ch>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author.

The mechanism to weaken or end the agreement has received little attention, as initial works have naturally focused on the restrictions themselves. Nonetheless, it would be a vital consideration for the parties entering the agreement, and some see this lack of consensus as a barrier to a potential agreement.

Since the design of an end phase is tied to the context of the agreement, we must specify which methods of limiting AI development we assume. For simplicity, unless stated otherwise, we assume a similar method to Scher et al. (2025), namely an upper limit on the amount of computational power deployed on training an AI. This method will not pause AI development even if the upper limit on computational power is lower than what is commonly used, because other factors like training data quality and algorithmic improvements will lead to further improvements overall.

We further assume that the agreement is signed by the NATO member states and China, but that there may remain non-signatory states with much less but non-negligible capability to develop frontier AI. Note that we understand the important risks from AI development to put all of humanity at risk, not only the states that conduct AI development, and that we therefore expect signatories to exert pressure on other states to sign as well. A likely scenario would therefore be that all states that could house frontier AI development become signatories. We still assume a more pessimistic scenario to show that the agreement discussed is robust to it.

Apart from the text of the agreement itself, it is important to consider the context in which its signatories enter into it. One crucial aspect is the expectation of countermeasures should a signatory violate the agreement. The existing literature on AI development agreements envisions either a scenario close to all other international agreements, where violations might lead to sanctions or escalation to the United Nations Security Council, or a more favorable scenario where signatories collectively hold the keys to each other's AI infrastructure and can unilaterally halt AI development by withholding their key (Aarne & Petrie, 2025; Scher & Thiergart, 2025). As above, while we hope for the implementation of infrastructure that enables enforcement of an agreement, we assume the more pessimistic case throughout, in which signatories do not have the capacity

to unilaterally halt AI development. The great powers that enter into the agreement would then be bound by only minor material factors, and instead the main reason to comply is the reputational cost of a violation. Reputation is especially valuable if there are continued gains from trade between states with and without advanced AI, which is uncertain.

The setting, an upper limit on computational power used for training an AI, offers a straightforward model of relaxing the restrictions imposed by the agreement: Once a certain condition is met, the limit is increased. Raising the limit to a sufficiently large amount of computational power would be equivalent to terminating the agreement entirely. We envision a series of gradual increases. Though the agreement will not be exited completely, we call a condition to increase the limit a *relaxation condition*, because such an increase would mark the exit of one phase of the agreement to the next. In contrast, we call a condition that definitively ends the agreement a *termination condition*. We think of termination conditions as a subset of relaxation conditions.

While we will treat relaxation conditions as relaxing all of the agreement, or the most important restriction of the agreement, we also draw attention to the existing notion of treaty reservation provisions (United Nations, 2011), which cover any kind of possibility of modification of the treaty after its entering into force. Relaxation conditions can be understood as a subcategory of reservation provisions.

The decision to exit from an agreement can occur in one of two scenarios. First, the purpose of the agreement has been achieved, or second, a signatory wishes to exit even though the original purpose has not been achieved. The agreement should be set up in such a way that the first scenario becomes more likely, and the second less likely.

In this work, then, we survey existing proposals and related international agreements for termination conditions and their implementations, lay out criteria for appropriate relaxation conditions, put forward candidates for conditions, and finally grade the conditions against the criteria.

2. Termination Conditions in Comparable Agreements

Previous international agreements have attempted to take countermeasures against other large risks stemming from coordination failures, namely risks from nuclear weapons, biological weapons, and environmental catastrophe. The risk profile of AI development combines aspects of each: A substantial probability of human extinction given unrestricted continued development (Grace et al., 2025; Center for AI Safety, 2023), as is the case for nuclear and biological weapons, and possible forgone short-term economic benefits if restrictions are in place (Sytsma, 2025), as is the case for environmental protection regulation.

In Section A, we analyze the termination conditions of the Strategic Arms Reduction Treaty, the Washington Naval Treaty, the Treaty on the Non-Proliferation of Nuclear Weapons, the Treaty on the Prohibition of Nuclear Weapons, the Anti-Ballistic Missile Treaty, the Biological Weapons Convention, the Paris Agreement, and the Vienna Convention for the Protection of the Ozone Layer and Montreal Protocol.

None of the surveyed treaties specify particular exit criteria under which the treaty is voided entirely. Instead, they mainly specify how the case is to be treated where a state wishes to exit the treaty without the goal having been satisfied.

3. Desiderata for Relaxation Conditions

Systematizing the selection of relaxation conditions, we propose a set of desiderata that a condition might satisfy. We intend, but do not suppose, to be exhaustive in these criteria.

3.1. Accuracy

An ideal relaxation condition should capture whether the risks of developing advanced AI have been adequately addressed in the time since the agreement came into effect. This might be difficult to determine, as it requires the signatories to not only agree on the goal of the agreement (to limit AI development), but also on the underlying reason, which is currently viewed as multifaceted. Specific risks that the signatories might want to reduce include risks from autonomous AI entities up to including human extinction, military AI and failure of mutual deterrence, catastrophic misuse like engineered pandemics, or simply economic issues such as rising unemployment.

Given that many non-state entities have an economic stake in the development of frontier AI, it would be expected that those entities would direct substantial efforts toward satisfying the relaxation condition irrespective of the intent behind it. Conversely, an actor may dedicate resources to keep the agreement in place, even if safe and beneficial AI development becomes possible. A relaxation condition ideally captures the true level of risk even under such adversarial conditions.

3.2. Verifiability

It should be unambiguous when the relaxation condition has been fulfilled. This property generally trades off against accuracy, because progress in research or societal resilience are difficult to determine unambiguously.

Even when a condition is unambiguous in its definition, it may still be difficult to produce adequate evidence trusted

by all signatories. With regard to the verification of the restrictions themselves, see Scher & Thiergart (2025) and Baker (2023), though the mechanisms necessary to verify relaxation conditions might not overlap much.

3.3. Preservation of Sovereignty

While international agreements generally are not understood to supplant national sovereignty, the "complete autonomy and self-determination over its internal and external affairs" (EBSCO Information Services, 2026), in practice international agreements constrain how states can exercise their sovereignty. Especially in the context of AI development, where a signatory's assessment of AI risks and benefits likely changes more quickly than in other policy domains, signatories likely prefer agreements that leave open a plausible path to exit the agreement in case their assessment changes.

If not enough sovereignty is preserved within the agreement, it becomes more likely that a signatory violates the agreement. This threshold is low in our assumed scenario without chip infrastructure to halt other signatory's development.

For a full discussion of withdrawal conditions as an incentive to participate in a treaty, see Koremenos & Nau (2010).

3.4. Precedent

Signatory states may place more confidence in a condition with successful precedent in previous international agreements, and may disfavor previously unsuccessful conditions.

3.5. Prioritization of Desiderata

Insofar as the above desiderata trade off against each other, it is important to determine whether certain desiderata carry more weight than others. This problem is adequately addressed by modeling the objectives of the signatories. One approach to doing so might be to treat the most important nations as rational actors, assume quantifiable beliefs about AI risk and potential benefits, the likelihood of developing a certain level of AI capabilities before the other actors, and so on, and then assign utility to each desideratum. For the sake of simplicity, we are content with qualitatively laying out the considerations that would go into such a model.

The incentives of signatories at the time of entering on the agreement may depend sensitively on the gap in AI capabilities between the leading and next closest state. We assume that the existing trends continue, with the United States as the leading state producing AI models that are around 7 months ahead of the models by the next closest state, China (Emberson, 2026). A larger gap allows for more accuracy at the expense of precedent and verifiability, assuming the leading state is most interested in addressing risks from AI.

The signatory states may have stronger or weaker beliefs in the risks of AI development. In the case of nuclear weapons, many actors signing international treaties can be assumed to have a stronger belief that nuclear war would pose an unacceptable risk to humanity or their citizens in the absence of international agreements. From an address to the UN General Assembly by then U.S. president John F. Kennedy: "Mankind must put an end to war—or war will put an end to mankind" (Kennedy, 1961). Separately, Kennedy is quoted as estimating at least a one-in-three chance that the Cuban Missile Crisis would have led to nuclear war (John F. Kennedy Presidential Library and Museum, 2002). As a result of this strong belief, signatories to treaties like the NPT have acceded to verification mechanisms that might otherwise be perceived as expensive or in conflict with national sovereignty, such as inspections by the International Atomic Energy Agency. This level of control by an international body as a result of a voluntary agreement had been unprecedented.

We assume that, given a stronger belief in large-scale AI risk, the desiderata preservation of sovereignty, verifiability, and precedent are considered less important.

Signatories may further have a higher or lower degree of belief in the benefits of AI development. If high, accuracy carries more weight, as the safe relaxation of the agreement is expected to confer larger benefits.

We assume that the signatories treat risk from AI development as comparable to risk from nuclear warfare in the 1960s. This assumption seems justified as a plurality of leading figures in AI from both the United States and China have already signed a statement to this effect (Center for AI Safety, 2023), and that conditional on states being willing to negotiate an agreement to limit AI development, they would be likely to share this conviction. We further assume that the signatories have similar beliefs about the economic benefits of unrestricted AI development, which we take to be a substantial, sustained increase in GDP growth rates.

Based on the above considerations, the desiderata might be weighed as: Accuracy > Preservation of Sovereignty > Verifiability > Precedent.

4. Types of Relaxation Conditions

4.1. Absence of a Relaxation Condition

Some authors have proposed that it is advisable to never develop certain levels of AI capability (Aguirre, 2023). This stance implies that an agreement should contain no relaxation and in particular no termination condition.

The option to terminate is generally present in international

agreements similar to the one we consider here. Nonetheless, agreements without termination conditions are possible. An example of an agreement without termination conditions is the UN Charter (United Nations, 1945). The Vienna Convention on the Law of Treaties (not to be confused with the treaty concerning ozone depletion) (United Nations, 1969), Article 54, reads:

> "The termination of a treaty or the withdrawal of a party may take place: (a) in conformity with the provisions of the treaty; or (b) at any time by consent of all the parties after consultation with the other contracting States."

Article 56, paragraph 1, reads:

> "A treaty which contains no provision regarding its termination and which does not provide for denunciation or withdrawal is not subject to denunciation or withdrawal unless: (a) it is established that the parties intended to admit the possibility of denunciation or withdrawal; or (b) a right of denunciation or withdrawal may be implied by the nature of the treaty."

And Article 62, paragraph 1, reads:

> A fundamental change of circumstances which has occurred with regard to those existing at the time of the conclusion of a treaty, and which was not foreseen by the parties, may not be invoked as a ground for terminating or withdrawing from the treaty unless: (a) the existence of those circumstances constituted an essential basis of the consent of the parties to be bound by the treaty; and (b) the effect of the change is radically to transform the extent of obligations still to be performed under the treaty.

While it is naturally difficult to anticipate what the most likely scenarios to satisfy Article 62, paragraph 1(a) and (b) are, we believe one likely scenario would be a non-signatory state gaining a realistic chance to develop more advanced AI than the signatory states.

For the absence of a condition, accuracy is medium. Risk from AI development is fully addressed, at least if consent of all parties adequately reflects that a reduction in risk from AI development has indeed occurred during the period of the agreement. On the other hand, the agreement may continue even if the risks of AI are addressed. A more formal codification of the conditions under which the risk would be reduced could provide yet higher accuracy. Precedent and preservation of sovereignty are low, while verifiability is by construction high.

4.2. Fixed Time Limit

A fixed time-limit expiry ending with a conference where the signatories vote on an extension may be a simple solution. Its main advantages are trivially high verifiability, feasibility, and good precedent. Furthermore, it could be used as a stopgap measure while more nuanced relaxation conditions are developed.

Assuming AI developers are confident that the agreement will not be renewed, they could further use the time before expiry to prepare for increased development efforts, compromising accuracy. The limiting measures of the agreement might consequently have to be amended to cover this effect, increasing the difficulty of enforcing the agreement. Further, the end of the agreement might present a coordination point for entities to influence the states' renewal decision.

4.3. Individual Withdrawal / Extraordinary Events

A condition allowing signatory states to withdraw from the agreement is included in many international treaties to preserve the signatories' sovereignty. Such a withdrawal can be made conditional on the withdrawing signatory disclosing some extraordinary events that caused it to withdraw, as is done in the NPT and ABM treaties. While such a condition has been included in existing agreements, precedent cannot unequivocally be said to be positive, as there are prominent cases of signatories disagreeing about whether another signatory's withdrawal is legitimate, creating tension. For example, the United Nations does not recognize North Korea's withdrawal from the NPT as legitimate. How a withdrawal deemed inconsistent with the treaty by other signatories would be handled is beyond the scope of this document.

Withdrawal clauses usually contain a notice period, and so a natural question is how long this should be for the agreement in question. We imagine the situation under which such a clause would be triggered. While this is challenging, we find the following scenario to be representative: a non-signatory state that was previously significantly behind acquires the chips necessary to conduct a training run at the limit specified in the agreement, thus putting it on roughly equal footing with the leading signatories. It is currently starting to build the facilities. The signatories should at least have enough time to stay ahead of the non-signatory in this scenario. Assuming the computation limit is roughly the contribution of a single U.S. data center for a current frontier training run, the fastest planned build time for a gigawatt-scale frontier data center is 12 months, and the fastest buildout of a frontier data center facility to date was completed in around 4 months (Epoch AI, 2026). Based on this, we recommend a notice period of 4 months. Note that this assumes that the signatories do not have a reliable way to prevent or predict compute capacity increases of non-

signatory states, which might well be the case if large-scale compute monitoring is implemented. It also assumes that frontier data center buildout times will not change; in particular, buildout times could become shorter if it becomes easier to build a data center at the smallest scale that the agreement restricts. A notice period of 4 months is the 28th percentile of notice periods among 86 international agreements analyzed in Koremenos & Nau (2010).

4.4. Progress in Preparedness

We summarize the totality of approaches to mitigate risks from AI in the term *preparedness*. Preparedness includes AI safety research, such as alignment and interpretability, as well as societal resilience to effects of AI development, such as increased cyberattack capabilities and economic upheaval.

As progress in preparedness continues, the risk from AI development might be sufficiently lowered to continue development. How to best determine this is an open problem, however.

4.4.1. Expert Consensus

During the time of the agreement, the scientific community may come to the conclusion that the risk from AI development has been sufficiently lowered, serving as an indication that the agreement may be safely terminated. As one example of several ways to reduce risk, Scher et al. (2025) propose an "established consensus about the efficacy of ASI-alignment methods" as a condition. The signatories may appoint a council of experts in the field of AI safety and condition relaxation on the council's approval after regular meetings.

What remains to specify is what, exactly, the signatories would task the expert council to answer. One might task it with an end-to-end prediction of the likelihood of catastrophic harm from AI (say, the likelihood of human extinction) from unrestricted AI development over the coming years (say, 25 years), and terminate the agreement if this likelihood is substantially lower with the agreement in place than with the agreement terminated. However, there are several reasons why a signatory might disagree with the expert council on AI risk once it is established. First, each signatory and the council might have different views on the contributions of different risk factors, for instance weighing the prospect of large-scale unemployment against more catastrophic harms like risk from rogue AI systems. Second, the signatories might also collectively be less or more conservative in their assessment of AI risk than the council. These possibilities would be perceived as a limit on the sovereignty of signatories, rendering them less likely to enter into the agreement.

Alternatively, the council might be tasked with judging solutions to more specific questions that the signatories believe to be central to the safe development of AI. One could imagine drawing from the field of pure mathematics, specifying a list of problems akin to the Millennium Prize Problems that, once solved, provide a foundation for safe AI development. One could imagine, for instance, that the expert council asks for progress in AI with provable guarantees.

We present Tegmark & Omohundro (2023) as a representative effort to characterize what constitutes sufficient progress in AI with provable guarantees. They outline an agenda to develop AI systems that provably satisfy safety requirements by extracting algorithms from neural network weights using interpretability techniques, then verifying the adherence of the extracted algorithm to requirements by formulating the requirements as proof-carrying code. Such proofs could conceivably rule out that an AI produces harmful output, or mathematically bound the probability with which an AI produces harmful output. The agenda also identifies interventions at several other layers of AI development to prevent execution of unsafe AI systems. A relaxation condition based on these proposals might require the development of all of the above for a class of AI systems that can be expected to yield a significant fraction of economic benefit as compared to what unrestricted AI systems would yield. A more comprehensive overview of approaches to provably safe AI is given in Dalrymple et al. (2024).

Given that many actors in AI development today explicitly or implicitly treat provably safe AI systems as intractable, we present two other representative types of conditions that have recently emerged out of practice.

With regard to risk of catastrophic AI misuse, Zou et al. (2025) give a practical definition for robustness against adversarial attacks. They let well-incentivized testers spend a fixed amount of time trying to find an input that makes their system output harmful content. While they made significant advancements, and could therefore report that no tester elicited harmful output on all 10 test cases, the testers did elicit harmful output in many cases. Further, testing was done for a cumulative 3,000 hours across all testers, which is still below what the model is exposed to in deployment. When exposing the system to testers through an openly accessible challenge, multiple testers found strategies that produced harmful input on all 10 test cases within the span of 5 days. A relaxation condition based on this line of research might ask for development of a system that performs above a given capability threshold, while producing no harmful output in a test like the above.

Largely aiming to address risks from rogue AI, Olah (2023) and Elhage et al. (2022) outline an agenda to simplify the problem of fully characterizing likely outputs from an AI. They claim that this may be possible by enumerating some

distinct concepts, or "features", that make up the computation that leads to the output. If those features are exhaustively enumerated and characterized by when they arise, and in turn satisfactorily explain the output of the system, this would mark significant progress in mitigating many types of AI risk. While there is no agreed-upon precise operationalization of the above argument that could readily serve as a condition, most researchers in interpretability today work within this paradigm (Sharkey et al., 2025). A relaxation condition based on this line of research might ask for an AI model that performs above a given capability threshold, with outputs explainable by an exhaustive set of intelligible features.

Progress in preparedness for the development of advanced AI encompasses more than AI research alone. For example, sufficiently advanced AI could cause rapid job displacement (Occhipinti et al., 2025; Korinek & Suh, 2024), or disruption of international relations (Pavel et al., 2025). We therefore believe it would be beneficial for states to develop contingency plans for advanced AI as it relates to their internal affairs, as well as publish plans on how they will act internationally once advanced AI is developed. A relaxation condition based on this might necessitate submission of a report in a specified format by each signatory.

We do not recommend specific conditions here out of those sketched above. None of the conditions presented here are to be understood as completely sufficient conditions to nullify risks from AI development in isolation.

4.5. Deference to a Later Specification

Given the difficulty of accurately specifying an adequate level of preparedness, one can imagine a two-stage agreement, the first stage of which terminates when such a specification has been produced.

Since the creation of such a specification is sufficiently difficult to warrant coordinated research efforts, this could be understood analogously to the Vienna Convention, in which the signatories were not sufficiently certain to commit to specific measures, and instead decided to commit to researching appropriate measures instead. The crucial difference between AI development and ozone depletion, besides the much larger scale of risk from AI development, is the high speed and exponential nature of AI development compared to the low speed and approximately linear nature of ozone depletion. Therefore, for the agreement we treat here, we could require that restrictions are already in place for the first stage of the agreement, whereas the Vienna Convention did not impose any binding restrictions to lower ozone-depleting emissions. Instead of deferring the restrictions themselves to the second stage, we defer the relaxation conditions.

Observing a period in which the restrictions of the agreement are in place might enable a more accurate specification of a relaxation condition, as this allows gathering more data on how preparedness research is impacted by the restrictions.

4.6. Absence of AI-Related Incidents

While the development of frontier AI systems would be restricted under the agreement, it is likely that the deployment of AI systems at least as powerful as those available today remains unrestricted. Therefore, incidents of harm from AI can provide an indication that relaxation would not be appropriate, and conversely, absence of such incidents provides weak evidence that relaxation is more appropriate than previously thought. Relatedly, Scher et al. (2025) propose "strong misuse and proliferation controls" as a relaxation condition, which might be empirically evaluated by their efficacy in preventing incidents with current technology. Damages from AI-related incidents can be measured with existing liability infrastructure, though it might be unclear what incident is attributable to an AI system as opposed to its user, owner, or manufacturer (Smith et al., 2024).

On the other hand, this approach is vulnerable to manipulation by an entity wishing to prevent relaxation of the agreement via purposefully instigating AI-related incidents. Also, and most importantly, it is far from clear that the absence of AI-related incidents under the agreement restrictions provides sufficient evidence that development without the restrictions would continue to be safe.

One example of an entity already tracking AI-related incidents is the AI Risk Repository, which counted 370 incidents in 2025, of which 211 were attributed to malicious actors (Slattery et al., 2024).

4.7. Creation of a Global AI Development Project

Some of the risks stemming from AI development are mediated by the fact that current frontier AI development efforts are undertaken by several distinct companies, which treat each other as competitors within and beyond state borders. The current situation is adequately described as a race of multiple actors who individually wish, or claim to wish, to slow development and spend significantly more resources on safe development, but cannot, because they would be overtaken by the other actors, losing their upside. A unified AI development project would allow these actors to balance AI development and safety research to a level deemed more appropriate, enable a more democratic AI development process, and allow the signatories to continue to impose restrictions on the unified project as needed. Scher et al. (2025) propose the "creation of a globally-monitored AI development project that proceeds with due caution" as one of the necessary conditions to terminate the agreement. We as-

sume that actors besides the international project would still be restricted by the original conditions.

The creation of a global AI development project only indirectly addresses risks from AI development, by enabling resources to be adequately allocated to preparedness research. As in Section 4.5, the project would still need to determine under what conditions additional development restrictions would need to be initiated and terminated, albeit under significantly more favorable conditions.

One downside of this approach is that a global AI development project would require a perhaps unprecedented level of international coordination of state actors and, once successful, could become a powerful entity in comparison to a single signatory state. Therefore, the creation of such an entity would be an unusual policy goal for any one party, and certainly for most states taken together, to pursue, which in turn might mean that the relaxation condition would be unduly unlikely to be satisfied. If the condition includes a requirement that a certain state must be part of the development project for the agreement to terminate, that state would have considerable influence over keeping the agreement in force.

An overview of our judgment for each relaxation condition type on each desideratum is given in Table 1.

5. Future Work

As current efforts to establish an international agreement are still in very early stages, we are tempted to recommend various lines of future work. We limit ourselves to mentioning only those that we expect to be immediately useful given the unprecedented speed of AI development (Kwa et al., 2025).

A consensus on what constitutes sufficient conditions to continue AI development could be reached in the near future. One simple step towards this could be a survey of AI experts in the United States and China on what they believe AI risks are and which methods to address each risk are most promising. While we believe that such a survey would currently yield many contradictory responses and no clear majority, it could also uncover unexpected commonalities within and across state borders. We think that beliefs around AI risk will converge over time as risks become more apparent. Beyond a passive assessment of researchers' stances on AI risk, dialogue between experts would likely accelerate consensus.

In a dialogue during the 2026 World Economic Forum, the CEOs of Anthropic and Google DeepMind agreed that both were in favor of a slower pace of development, but showed concern about the feasibility of an international agreement (Miailhe, 2026). We encourage frontier AI companies to develop their existing necessary conditions for release of a current model generation into sufficient conditions that would be suitable as relaxation conditions in an international agreement. We believe that engagement from AI companies on questions of international coordination would be highly promising in general.

6. Conclusion

We give a recommendation for relaxation conditions based on our analysis.

The agreement should end after 5 years, during which time the organization determining satisfaction of the relaxation conditions should be established. This organization proposes the relaxation conditions, and if the parties cannot agree to them after 5 years, they may agree to extend the agreement by another 2 years, or terminate it. The accepted relaxation conditions must address preparedness for the possibility of rogue AI, catastrophic misuse, and catastrophic societal developments. Individual withdrawal based on extraordinary events is possible with 4 months' notice.

We briefly give our reasoning for each element of the agreement. The creation of an organization that specifies relaxation conditions is based on the construction from the Vienna Convention (see Section 4.5) and might cease to be necessary once a consensus on adequate preparedness emerges among experts. Requiring parties to later agree on the terms, as opposed to a unilateral decision by the newly created organization, prevents conditions that are either too lax or too strict: we expect some signatories to have stronger, and some weaker, beliefs regarding AI risk. In case it is deemed too unlikely that all signatories agree, a possible variation would be to designate a subset of signatories whose agreement is sufficient for the agreement to proceed to the next phase. We include the withdrawal clause largely based on the assumption that a non-signatory state has a realistic chance to develop frontier AI. If such a non-signatory state were not to exist, we would not recommend a withdrawal clause, based on the relative success of the Washington Naval Treaty and controversial uses in other agreements, like the ABM treaty and NPT.

We stress again that the above is based on one specific proposal for restrictions on AI development, a compute limit. Though an unsolved problem, we believe it to also be possible to impose restrictions that do not need relaxation conditions, because they accurately capture the risks of advanced AI without the need for a proxy. For instance, one could imagine international restrictions of AI systems based on their capability; for example an AI system that can autonomously self-replicate, that can autonomously perform a large fraction of cognitive labor of today, or that can autonomously perform cyber attacks at a certain level of capability. This type of restriction would be more similar to

frontier safety policies already implemented at companies (METR, 2025). Because this approach comes with many challenges of its own, such as the lack of an adversarially robust specification of levels of capability, or disagreement about what capabilities are economically valuable enough to keep unrestricted, we find relaxation conditions in conjunction to more verifiable restrictions to be a promising avenue.

We look forward to continued, increasingly practical progress in international cooperation on AI. We believe that an international agreement as discussed would adequately address the single largest risk that each potential signatory state, and humanity as a whole, faces today.

Acknowledgements

We greatly appreciate comments by Aaron Scher, David Krueger, Mauricio Baker, Oliver Guest, Matthijs Maas, Thomas Larsen, and Nate Soares, which significantly improved this work. All opinions expressed are the authors', and in no way reflect those of the acknowledged. All errors are the authors'.

Impact Statement

We write this work in hope of having a positive impact on the trajectory of AI development. Some potential negative side effects of our proposal, such as forgone short-term economic benefits, are discussed in the main text. We believe that international agreement to limit AI development would be highly beneficial overall.

AI Usage Statement

The text in this article was written manually by the authors. Language models were used for spell-checking, some copy-editing, literature search, and formatting, though all resulting edits were applied manually.

A. Comparable Treaties

A.1. Strategic Arms Reduction Treaty

The Strategic Arms Reduction Treaty (START) (United States & Union of Soviet Socialist Republics, 1991) is perhaps the closest international agreement to what we treat here. In it, the United States and Soviet Union coordinate to reduce their capabilities to launch nuclear weapons. The agreement contains detailed technical requirements and sophisticated mechanisms of verifying compliance. Its Article XVII specifies the termination conditions. It includes a 15-year time limit after which a 5-year extension is to be agreed on, and a right to withdraw from the treaty based on "extraordinary events, related to the subject matter of this Treaty" that "have jeopardized the supreme interests" of the signatory country, with 6 months' notice. Before the 15 years ran out, the Soviet Union had collapsed, and a new agreement was reached between the United States and Russia, the Strategic Offensive Reductions Treaty. This was in turn superseded by the Measures for the Further Reduction and Limitation of Strategic Offensive Arms Treaty (New START) (United States Department of State, 2010). The latter has identical termination conditions to START, except with 3 months' notice and a 10-year renewal period and a renewal of "no more than five years". One month before expiry, the agreement was renewed by 5 years, though Russia later announced that it was suspending its participation before expiry, alleging non-compliance by the United States (Agence France-Presse, 2023). The agreement has

Table 1. Qualitative assessment of relaxation condition types against the desiderata.

	Desideratum			
Relaxation Condition	**Accuracy**	**Verifiability**	**Preservation of Sovereignty**	**Precedent**
Absence of a Relaxation Condition	Medium	High	Low	Medium
Fixed Time Limit	Low	High	Medium	High
Extraordinary Events	Low	Low	High	Medium
Absence of AI-Related Incidents	Medium	High	Medium	Low
Creation of a Global AI Development Project	Medium	High	Low	Low
Progress in Preparedness	High	Medium	Medium	Low
Deference to a Later Specification	N/A	N/A	High	Medium

now expired.

A.2. Washington Naval Treaty

The Washington Naval Treaty (Washington Naval Conference, 1922) coordinated the slowing of warship development between the Allies after World War I. While old, it is relevant to the agreement we treat because it restricts as of then undeveloped technology in a quantitative manner, i.e., it restricts displacement volume of warships, barrel diameters, and similar characteristics. Its Article XXIII treats termination. After a minimum duration of around 14 years, a signatory could give unilateral notice of termination without providing a reason for doing so, subject to a two-year notice period. The treaty would then terminate within two years of receipt of the notice. After several parties violated the agreement either only in spirit or fully, Japan gave notice of termination at the end of the 24-year period.

A.3. Treaty on the Non-Proliferation of Nuclear Weapons

The Treaty on the Non-Proliferation of Nuclear Weapons (NPT) coordinates an effort to stop the spread of nuclear weapons among 190 states (United Nations, 1968). Notably, India, Pakistan, and Israel did not sign the treaty. India and Pakistan first demonstrated functional nuclear weapons after NPT went into force, whereas Israel is commonly believed to have developed functional nuclear weapons since before NPT was signed. The treaty's Article X specifies the termination conditions, which are similar to START except for a 25-year initial duration and a 6 months' notice of withdrawal. Notably, the latter clause was exercised by North Korea in 1993, then reversed before the end of the notice period. It withdrew once more in 2003, citing what it perceived as hostile U.S. foreign policy and an IAEA resolution (Government of the Democratic People's Republic of Korea, 2003; Ministry of Foreign Affairs of Japan, 2003). As the withdrawal is not recognized by many other signatories, it is considered by some states, including the United States, to be in violation of the treaty (U.S. Department of State, 2025). No other signatory state has exercised the article.

A.4. Treaty on the Prohibition of Nuclear Weapons

The Treaty on the Prohibition of Nuclear Weapons expands on the provisions of the NPT by coordinating further, and eventually complete, nuclear disarmament of 95 signatory states (United Nations, 2017). No state in possession of nuclear weapons has signed the treaty. Its Article 17 specifies the termination conditions. The duration is explicitly declared as unlimited. Withdrawal is possible under similar conditions as in the NPT, except with a longer 12 months' notice. The withdrawal is considered invalid if the withdrawing signatory is party to an armed conflict, and the signatory remains bound by the treaty until the end of that conflict. No signatory state has exercised the article.

A.5. Anti-Ballistic Missile Treaty

The Anti-Ballistic Missile (ABM) Treaty was a bilateral agreement between the United States and the Soviet Union to restrict the development of defense systems against intercontinental ballistic missiles (United States & Union of Soviet Socialist Republics, 1972). It is similar to the agreement we assume here in that it voluntarily limits the scale of both existing and future technology, forgoing a potential military advantage over the counterparty. Its Article XV specifies the termination clause. While the duration is explicitly specified as unlimited, the signatories reserve a right to withdraw, again based on "extraordinary events, related to the subject matter of this Treaty" that "have jeopardized the supreme interests" of the signatory country. In 2002, the United States withdrew. Former president George W. Bush issued a public statement, distinct from the notification to Russia, citing that the agreement "hinders our government's ability to develop ways to protect our people from future terrorist or rogue-state missile attacks" (Arms Control Association, 2001) and a shifting geopolitical situation after the collapse of the Soviet Union.

A.6. Biological Weapons Convention

The Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction coordinates the

non-proliferation of biological weapons between 189 parties (United Nations, 1972; United Nations Office for Disarmament Affairs, 2026). Its Article XIII specifies termination conditions. The agreement is explicitly stated as being unlimited in duration. Similarly to the ABM and NPT treaties, parties can withdraw based on "extraordinary events", with a 3 months' notice and a statement to the United Nations Security Council. No signatory has exercised this article.

There have been several cases of parties to the treaty alleging violation of the treaty by other parties. The United States has alleged violation of the treaty by Russia and North Korea, for instance in its 2025 Adherence to and Compliance with Arms Control, Nonproliferation, and Disarmament Agreements and Commitments Report (U.S. Department of State, 2025). Conversely, Russia has alleged violation by the United States, chiefly in its Outcome Report of the Parliamentary Commission on Investigation into the Circumstances Related to Creation of Biological Laboratories by U.S. Specialists on the Territory of Ukraine (Ministry of Foreign Affairs of the Russian Federation, 2023). An instance of a party to the treaty admitting to violation is given by former Chairman of the Supreme Soviet of the Russian SFSR Boris Yeltsin, then president of Russia (Cold Spring Harbor Laboratory Archives, 2026).

A.7. Paris Agreement

The Paris Agreement coordinates the lowering of greenhouse gas emissions among 194 parties (United Nations, 2015). Its only termination condition is Article 28, which concerns withdrawal from the agreement. Withdrawal is stated as possible starting 3 years after the agreement entered into force, with a 1 year notice period. The United States exercised the article shortly after the agreement entered into force, and once the 3 year period was over, withdrew. It was readmitted into the agreement soon after, but exercised the article again four years later. No other signatory state has exercised the article.

However, many parties are widely understood to be in violation of different requirements of the agreement. For instance, the agreement requires submission of nationally determined contribution reports on emission targets and efforts to reach them. Though submission was due in February 2025, by September 2025 only 57 parties out of 194 had submitted such a report (United Nations Environment Programme, 2025).

A.8. Vienna Convention for the Protection of the Ozone Layer and Montreal Protocol

The Vienna Convention for the Protection of the Ozone Layer coordinates the lowering of emissions that deplete atmospheric ozone among 198 ratifying states (United Nations, 1985). Its Article 19 specifies termination conditions, and is structured identically to the Paris Agreement, except for allowing withdrawal only after 4 instead of 3 years since going into force. No ratifying state has exercised the article.

Importantly, the convention largely leaves the implementation of what precisely is to be restricted to later-determined protocols, mainly focusing on commitment to research efforts to determine appropriate restrictions. This led to the creation of the Montreal Protocol two years later, specifying the corresponding binding restrictions.

The Montreal Protocol on Substances That Deplete the Ozone Layer coordinates 197 ratifying states toward the same goal as the Vienna Convention (United Nations, 1987). Its Article 19 specifies termination conditions, and inherits the conditions from the Vienna Convention's Article 19. No ratifying state has exercised the article.

References

Aarne, O. and Petrie, J. Part iii: International security applications of flexible hardware-enabled guarantees, April 2025.

Agence France-Presse. Putin says moscow suspending participation in new start nuclear treaty, February 2023. URL https://www.barrons.com/news/putin-says-moscow-suspending-participation-in-new-start-nuclear-treaty-d307fa0f.

Aguirre, A. Keep the future human: Why and how we should close the gates to agi and superintelligence, and what we should build instead, 2023. URL https://arxiv.org/abs/2311.09452.

Alaga, J. and Schuett, J. Coordinated pausing: An evaluation-based coordination scheme for frontier ai developers, 2023. URL https://arxiv.org/abs/2310.00374.

Arms Control Association. U.s. withdrawal from the abm treaty: President bush's remarks and u.s. diplomatic notes. Arms Control Association, December 2001. URL https://www.armscontrol.org/act/2002-01/us-withdrawal-abm-treaty-president-bushs-remarks-and-us-diplomatic-notes.

Aschenbrenner, L. and Trammell, P. Existential risk and growth. Technical Report 13-2024, Global Priorities Institute, 2024. URL https://globalprioritiesinstitute.org/existential-risk-and-growth/.

Baker, M. Nuclear arms control verification and lessons for ai treaties, 2023. URL https://arxiv.org/abs/2304.04123.

Bryson Gillette Insights. Opinions on advanced ai development: Findings from a national survey of adults, September 2025. URL https://www.brysongillette.com/opinions-on-advanced-ai-development.

Center for AI Safety. Statement on AI risk. https://aistatement.com/, 2023. Accessed: 2026-06-12.

Cold Spring Harbor Laboratory Archives. Matthew meselson collection, 2026. URL https://archivesspace.cshl.edu/repositories/2/resources/94. Accessed 2026-04-22.

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J., and Tenenbaum, J. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems, 2024. URL https://arxiv.org/abs/2405.06624.

EBSCO Information Services. Sovereign state. Research Starters, 2026. URL https://www.ebsco.com/research-starters/law/sovereign-state. Accessed 2026-04-22.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. Toy models of superposition. Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html#strategic.

Elsey, J. and Moss, D. Us public opinion of ai policy and risk, 2023. URL https://www.safe.ai/work/us-public-opinion-of-ai-policy-and-risk.

Emberson, L. Chinese AI models have lagged the US frontier by 7 months on average since 2023, 2026. URL https://epoch.ai/data-insights/us-vs-china-eci. Accessed: 2026-06-13.

Epoch AI. Frontier data centers, 2026. URL https://epoch.ai/data/data-centers. Published online at epoch.ai. Accessed 2026-04-21.

Government of the Democratic People's Republic of Korea. Statement of the dprk government on withdrawal from the npt. Government statement, 2003. URL https://www.atomicarchive.com/resources/documents/deterrence/dprk-npt-statement.html.

Grace, K., Sandkühler, J. F., Stewart, H., Weinstein-Raun, B., Thomas, S., Stein-Perlman, Z., Salvatier, J., Brauner, J., and Korzekwa, R. C. Thousands of ai authors on the future of ai. *Journal of Artificial Intelligence Research*, 84, 2025.

John F. Kennedy Presidential Library and Museum. Cuban missile crisis: A historical perspective. Transcript, October 2002. URL https://www.jfklibrary.org/events-and-awards/kennedy-library-forums/browse-all-forums/transcripts/cuban-missile-crisis-a-historical-perspective.

Kennedy, J. F. Address in new york city before the general assembly of the united nations. The American Presidency Project, September 1961. URL https://www.presidency.ucsb.edu/documents/address-new-york-city-before-the-general-assembly-the-united-nations.

Koremenos, B. and Nau, A. Exit, no exit. *Duke J. Comp. & Int'l L.*, 21:81, 2010.

Korinek, A. and Suh, D. Scenarios for the transition to agi. Technical Report 32255, National Bureau of Economic Research, 2024. URL https://www.nber.org/papers/w32255.

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., Rein, D., Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and Chan, L. Measuring ai ability to complete long software tasks, 2025. URL https://arxiv.org/abs/2503.14499.

METR. Common elements of frontier ai safety policies (december 2025 update). https://metr.org/blog/2025-12-09-common-elements-of-frontier-ai-safety-policies/, 12 2025.

Miailhe, N. The lab leaders want to slow down on ai. someone needs to help them. The Business Times, February 2026. URL https://www.businesstimes.com.sg/opinion-features/lab-leaders-want-slow-down-ai-someone-needs-help-them.

Ministry of Foreign Affairs of Japan. Iaea board of governors adopts resolution on safeguards in north korea. Press release, March 2003. URL https://www.mofa.go.jp/region/asia-paci/n_korea/iaea0301.html.

Ministry of Foreign Affairs of the Russian Federation. The outcome report of the parliamentary commission on investigation into the circumstances related to creation of biological laboratories by u.s. specialists on the territory of ukraine. Official statement, June 2023. URL https://www.mid.ru/en/foreign_policy/international_safety/1873584/.

Miotti, A. and Wasil, A. An international treaty to implement a global compute cap for advanced artificial intelligence, 2023. URL https://arxiv.org/abs/2311.10748.

Occhipinti, J.-A., Hynes, W., Prodan, A., Eyre, H., Green, R., Burrow, S., Tanner, M., Buchanan, J., Ujdur, G., Destrebecq, F., Song, C., Carnevale, S., Hickie, I. B., and Heffernan, M. Generative ai may create a socioeconomic tipping point through labour displacement. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-025-08498-x. URL https://www.nature.com/articles/s41598-025-08498-x.

Olah, C. Interpretability dreams. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/interpretability-dreams/index.html#end-goals.

Pavel, B., Ke, I., Smith, G., Brown-Heidenreich, S., Sabbag, L., Acharya, A., and Mahmood, Y. How artificial general intelligence could affect the rise and fall of nations: Visions for potential agi futures. Technical report, RAND, 2025. URL https://www.rand.org/pubs/research_reports/RRA3239-1.html.

Perrigo, B. Exclusive: Brits want to ban "smarter than human" ai. Time, February 2025. URL https://time.com/7213096/uk-public-ai-law-poll/.

Scher, A. and Thiergart, L. Mechanisms to verify international agreements about ai development, 2025. URL https://arxiv.org/abs/2506.15867.

Scher, A., Abecassis, D., Barnett, P., and Abeyta, B. An international agreement to prevent the premature creation of artificial superintelligence, 2025. URL https://arxiv.org/abs/2511.10783.

Scholefield, R., Martin, S., and Barten, O. International agreements on ai safety: Review and recommendations for a conditional ai safety treaty, 2025. URL https://arxiv.org/abs/2503.18956.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., and McGrath, T. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence, 2024. URL https://arxiv.org/abs/2408.12622.

Smith, G., Stanley, K. D., Marcinek, K., Cormarie, P., and Gunashekar, S. Liability for harms from ai systems: The application of u.s. tort law and liability to harms from artificial intelligence systems. Technical report, RAND, 2024. URL https://www.rand.org/pubs/research_reports/RRA3243-4.html.

Sytsma, T. Quantifying ai's economic potential: Growth differentials between assistive and autonomous development scenarios. Technical Report RR-A4220-1, RAND Corporation, 2025. URL https://www.rand.org/pubs/research_reports/RRA4220-1.html.

Tegmark, M. and Omohundro, S. Provably safe systems: The only path to controllable agi, 2023. URL https://arxiv.org/abs/2309.01933.

United Nations. Charter of the united nations and statute of the international court of justice. Treaty text, 1945. URL https://treaties.un.org/doc/Publication/CTC/uncharter.pdf.

United Nations. Treaty on the non-proliferation of nuclear weapons. Treaty text, 1968. URL https://www.iaea.org/sites/default/files/publications/magazines/bulletin/bull10-4/10403501117.pdf.

United Nations. Vienna convention on the law of treaties. Treaty text, 1969. URL https://legal.un.org/ilc/texts/instruments/english/conventions/1_1_1969.pdf.

United Nations. Convention on the prohibition of the development, production and stockpiling of bacteriological (biological) and toxin weapons and on their destruction. Treaty text, 1972. URL https://www.state.gov/wp-content/uploads/2022/06/BWC-Full-Text.pdf.

United Nations. Vienna convention for the protection of the ozone layer. Treaty text, 1985. URL https://ozone.unep.org/sites/default/files/2019-04/VC-Text%202016-English_0.pdf.

United Nations. Montreal protocol on substances that deplete the ozone layer. United Nations Treaty Series, 1987. URL https://treaties.un.org/doc/Publication/UNTS/Volume%201522/volume-1522-I-26369-English.pdf.

United Nations. Guide to practice on reservations to treaties (2011), 2011. URL https://legal.un.org/ilc/texts/instruments/english/draft_articles/1_8_2011.pdf.

United Nations. Paris agreement. Treaty text, 2015. URL https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf.

United Nations. Treaty on the prohibition of nuclear weapons. Treaty text, 2017. URL https://treaties.un.org/doc/Publication/UNTS/No%20Volume/56487/Part/I-56487-080000028 04c2398.pdf.

United Nations Environment Programme. Emissions gap report 2025: Off target. Technical report, UNEP, 2025. URL https://www.unep.org/resources/emissions-gap-report-2025.

United Nations Office for Disarmament Affairs. Unoda treaties database, 2026. URL https://treaties.unoda.org. Accessed 2026-04-22.

United States and Union of Soviet Socialist Republics. Treaty between the united states of america and the union of soviet socialist republics on the limitation of anti-ballistic missile systems. Treaty text, 1972. URL https://media.nti.org/documents/abm_treaty.pdf.

United States and Union of Soviet Socialist Republics. Treaty between the united states of america and the union of soviet socialist republics on the reduction and limitation of strategic offensive arms. Treaty text, 1991. URL https://2009-2017.state.gov/t/avc/trty/146007.htm.

United States Department of State. New strategic arms reduction treaty (new start). Treaty text, 2010. URL https://2009-2017.state.gov/t/avc/newstart/c44126.htm.

U.S. Department of State. 2025 arms control treaty compliance report. Report, 2025. URL https://www.state.gov/wp-content/uploads/2025/04/2025-Arms-Control-Treaty-Compliance-Report_Final-Accessible.pdf.

Washington Naval Conference. Treaty between the united states of america, the british empire, france, italy, and japan, signed at washington, february 6, 1922. Treaty text, 1922. URL https://treaties.fcdo.gov.uk/data/Library2/pdf/1924-TS0005.pdf.

Zou, A., Wang, Z., Kolter, J. Z., Fredrikson, M., and Chaudhuri, K. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming, 2025. URL https://arxiv.org/abs/2501.18837.