RESEARCH PAPER
JULY 2025

Verification for International AI Governance

Authors: Ben Harack, Robert F. Trager, Anka Reuel, David Manheim, Miles Brundage, Onni Aarne, Aaron Scher, Yanliang Pan, Jenny Xiao, Kristy Loke, Sumaya Nur Adan, Guillem Bas, Nicholas A. Caputo, Julia C. Morse, Janvi Ahuja, Isabella Duan, Janet Egan, Ben Bucknall, Brianna Rosen, Renan Araujo, Vincent Boulanin, Ranjit Lall, Fazl Barez, Sanaa Alvira, Corin Katzke, Ahmad Atamli, and Amro Awad

Verification for International AI Governance

Ben Harack[1], Robert F. Trager[1,2], Anka Reuel[3,4], David Manheim[5,6], Miles Brundage[7], Onni Aarne[8], Aaron Scher[9], Yanliang Pan[10], Jenny Xiao[11,12], Kristy Loke, Sumaya Nur Adan[1,13], Guillem Bas[14,15], Nicholas A. Caputo[1], Julia C. Morse[1,16], Janvi Ahuja[1,17], Isabella Duan[18], Janet Egan[19], Ben Bucknall[1], Brianna Rosen[20], Renan Araujo[1,8], Vincent Boulanin[21], Ranjit Lall[22], Fazl Barez[1], Sanaa Alvira, Corin Katzke[23], Ahmad Atamli[24], and Amro Awad[1]

[1]Oxford Martin AI Governance Initiative,
[2]Blavatnik School of Government, University of Oxford, [3]Stanford University,
[4]Robert and Renée Belfer Center for Science and International Affairs,
Harvard Kennedy School, Harvard University,
[5]Association for Long Term Existence and Resilience,
[6]Technion - Israel Institute of Technology,
[7]AI Verification and Evaluation Research Institute, [8]Institute for AI Policy and Strategy,
[9]Machine Intelligence Research Institute,
[10]James Martin Center for Nonproliferation Studies, [11]Leonis Capital,
[12]Columbia University, [13]Centre for AI Security and Access,
[14]Observatorio de Riesgos Catastróficos Globales, [15]Future Impact Group,
[16]University of California, Santa Barbara, [17]Big Data Institute, University of Oxford,
[18]Safe AI Forum, [19]Center for a New American Security,
[20]Oxford Programme for Cyber and Technology Policy,
[21]Governance of AI Programme, Stockholm International Peace Research Institute,
[22]University of Oxford, [23]Convergence Analysis, [24]University of Southampton

Each author contributed ideas or writing to the report. However, being an author does not imply agreement with every claim made in the report. Authors are listed in approximately descending order of contribution.

Abstract

The growing impacts of artificial intelligence (AI) are spurring states to consider international agreements that could help manage this rapidly evolving technology. The political feasibility of such agreements can hinge on their *verifiability*—the extent to which the states involved can determine whether other states are complying. This report analyzes several potential international agreements and ways they could be verified. To improve the robustness of the conclusions, pessimistic assumptions are made about the technical and political parameters of the verification challenge.

This report has three primary findings. First, verification of many international AI agreements appears possible even without speculative advances in verification technology. Some agreements can be verified using existing hardware, while others will require major investments in developing and installing verification infrastructure. In particular, verifying the regulation of data center-based AI development and deployment appears to be possible within a few years if serious efforts are made toward that goal. One such scheme would require 1) constructing and installing narrow-purpose verification hardware in data centers and 2) creating a mutually verified data center which can run privacy-preserving computations. Second, verification for some kinds of AI-related activities is likely to face a combination of technical and political barriers, thus limiting prospects for agreement. In particular, the detailed regulation of mobile AI-enabled devices in sensitive domains—such as weapons—faces severe political challenges. Third, near-term actions in several areas, including research and development as well as state policy, can improve the prospects for future verification agreements by reducing costs and security concerns. In sum, this report outlines workable approaches for verifying international AI agreements and illustrates how investments in verification today can shape the political possibilities of tomorrow.

Corresponding author: Ben Harack (ben.harack@gmail.com).

Contents

Executive Summary

Rapid changes in the artificial intelligence (AI) ecosystem have galvanized government efforts to understand this technology and shape its future. States may seek to create international agreements over AI to capture new economic opportunities, preserve peace, and mitigate risks created or exacerbated by AI. This report examines the potential for states to undertake *verification* of international agreements relating to AI—where verification is any process by which member states can assess each other's compliance with an agreement. This analysis aims to take into account both technological limits and political sensitivities. We find that many AI agreements could be implemented verifiably today. Within a few years, targeted technical and policy efforts may allow robust and politically viable verification for a much wider range of possible agreements. One notable exception is agreements seeking to provide detailed regulation of the behavior of mobile AI-enabled devices, such as weapons, where a combination of domain-specific technical limitations and political sensitivities make it extremely challenging to design acceptable verification.

Verification can be a crucial component of international agreements. Without verification, states may find that otherwise desirable deals are unavailable, just as mutual cooperation is often unavailable in social dilemmas such as the prisoner's dilemma. Effective verification mechanisms can transform strategic deadlocks into viable compromises. Verification provides states with more political room to maneuver and with reliable information about the actions of their fellows. Therefore, it generally is in the interest of states to improve their ability to verify agreements. For example, toward the end of the Cold War, new approaches for verifying missiles expanded the set of realistic political options for arms control, eventually enabling the conclusion of agreements such as the Intermediate-Range Nuclear Forces (INF) Treaty of 1987.[1]

Both unilateral and cooperative verification techniques are valuable. Unilateral techniques—such as satellite imagery and intelligence operations—allow the behavior of a state to be scrutinized without its cooperation. By contrast, cooperative verification requires the active participation of both states—with the Prover attempting to credibly demonstrate that they are following an agreement and the Verifier examining all available information to make inferences about the Prover's compliance. Some agreements require states to take on both roles as they simultaneously try to demonstrate their own compliance while checking their peers' compliance.

Pessimistic assumptions about verification difficulty

This report aims to make claims about the possibilities for AI verification that are robust to technical and political change. To do so, it makes four pessimistic assumptions about the dif-

[1] Toivanen, 'The Significance of Strategic Foresight in Verification Technologies'.

ficulty of verification. First, we assume that decentralized training of large AI models (using multiple geographically distributed data centers) will be tractable with minimal efficiency losses compared to centralized training—thus requiring that AI-development governance and verification be applied to small data centers as well as large ones. Second, we assume that algorithmic progress may be rapid, thus drastically increasing the capabilities that may be available with a given quantity of compute. Third, we assume that AI will be employed by security-sensitive institutions such as intelligence agencies and militaries for sensitive purposes, including those relating to personal, corporate, and state secrets. Fourth, we assume that future treaties might need to subject even very sensitive uses of AI to detailed and verifiable governance, thus requiring that verification plans be compatible with secrecy and security.[2] The verification proposals described in this report are intended to be workable even if all of these pessimistic assumptions become reality. Moreover, since these mechanisms are designed to be robust in the face of evolving technical and political realities, they are also likely to be workable in conditions that are less constrained.

Agreements and their verifiability

This report examines five families of agreements relating to AI:

1. **Transfer knowledge**: A state provides AI-related knowledge to another state.
2. **Transfer resources**: A state provides AI-related resources to another state.
3. **Pool resources**: States combine AI-related resources toward a common goal.
4. **Prepare for emergencies**: States prepare to detect AI emergencies and respond to them.
5. **Regulate**: States regulate AI development and deployment according to shared rules.

These agreement families are verifiable to different degrees, as summarized in Figure 1.

The first two families of agreements—"*transfer knowledge*" and "*transfer resources*"—are similar to prior international agreements regarding the transfer of knowledge and resources. Just as with their precursors in domains such as weapons-production technologies or nuclear energy, these agreements are primarily limited by the ability of the receiving state to credibly demonstrate that transferred knowledge or resources will not be used against the interests of the sending state. Some marginal improvement in the verifiability of these agreements is possible in the next few years via privacy-preserving digital verification tools as discussed below. However, the risk of downstream misuse is expected to remain the crucial factor limiting these agreements.

The third family of agreements—"*pool resources*"—is also similar to historical analogues, but it faces less severe political problems and is therefore quite verifiable even today. States regularly create institutions to solve shared problems. By design, such international institutions tend to provide enough information to states to demonstrate that pooled resources are not

[2] These final two assumptions force us to confront the transparency-security tradeoff, where the transparency needed to demonstrate compliance with an agreement is believed to make a state less secure. This tradeoff is the central political challenge for some of the international agreements examined below.

Agreement family	Example agreement	Verifiability if implemented today	Verifiability presuming five years of serious effort
Transfer knowledge	Share knowledge of AI risks	Yes, with political limitations*	Yes, with political limitations*
Transfer resources	Share AI-specialized chips	Yes, with political limitations*	Yes, with political limitations*
Pool resources	Pool resources toward international goal	Yes	Yes
Prepare for emergencies	Computational emergency detection and repsonse	No	Maybe
Regulate	Regulate data center computations	No	Yes
	Regulate AI-enabled weapons	Very limited	Limited

*The sending state must deem the risks of knowledge or resource misuse to be tolerable.

Figure 1: The families of international agreements examined in this report and their approximate verifiability.

being misused. While most prior agreements of this kind dealt with relatively low-stakes domains, some engaged directly with core state interests such as security and prosperity (e.g., the European Coal and Steel Community). Overall, agreements aiming to pool AI resources for political purposes are highly verifiable.

The final two families of agreements—"*prepare for emergencies*" and "*regulate*"—are much more difficult to verify today, but preparation may allow some agreements in this domain to be verified robustly in the coming few years. In particular, it appears likely that a few years of intense work could allow rules about data center-based computations to be verified at scale, including the development and deployment of AI. Challenges remain for AI-enabled devices that are both *mobile* and are employed for *sensitive* purposes, such as weapons. The regulation of weapons in particular is likely to remain politically difficult even when using the verification approaches best suited to the challenge.

In sum, AI-centered agreements that bear resemblance to historical agreements over similarly sensitive technologies (such as civilian nuclear energy) are roughly as verifiable as those prior agreements were. However, some of the agreement families discussed in this report—namely the "prepare for emergencies" and "regulate" families—pose fundamentally new challenges that are the key reason why these agreements are largely unverifiable today. The central challenge in verifying these agreements is the difficulty of making credible claims about *rules* that are applied to *computations*. This challenge is the central topic of the report and will be explored further below.

The challenge of verifying computational rules and tools for accomplishing it

Demonstrating that rules are being applied to computations (such as those central to AI development or deployment) is very challenging, but hardware-centric verification schemes make it possible. Detailed knowledge of computational activities undertaken by a state is not reliably available to other states, and therefore unilateral verification is of limited use in this domain.[3] Cooperative verification has great potential, especially via verification processes that focus on AI-specialized computational hardware—hereafter called "chips" or "compute". Being a physical asset, compute is easier to place verifiable controls on than are the other crucial inputs into AI (data and algorithms). AI can only be developed and deployed on computational hardware—with compute, because of its tailoring, significantly outperforming other kinds of hardware. The anticipated centrality of compute for the future of AI means that hardware governance is likely to be relevant into the future regardless of how the relative costs of inputs change or how AI paradigms evolve.[4,5] Compute-centric governance also enables fine-grained controls which minimize disruption to compute applications. Contrary to the common notion that digital verification is more challenging than physical verification, globe-spanning verification systems already enable the operation of the Internet thanks to the remarkable abilities of modern cryptography. Similarly, civilization-scale verification of computational activities on compute hardware appears to be not only feasible but also achievable in a way that is designed to robustly protect the privacy and security of all parties.

All verification tools are imperfect, thus often requiring that tools be employed in combinations that address their individual weaknesses and achieve a desired effect. Four kinds of verification techniques are worth emphasizing:

1. States can use **unilateral** verification mechanisms to help them understand the broad shape of activities undertaken by other states—such as their compute or power infrastructure investments. States may also have access to much more sensitive information via intelligence operations, thus allowing them to double check that declarations are correct and complete. Overall, unilateral verification serves a crucial role in limiting

[3] Even though cyber attacks and intelligence operations are very capable of revealing important information, these techniques are unlikely to be sufficient for *reliable* and *ongoing* access to the *details* of all computations occurring in a rival state's data centers.

[4] An important caveat here is consumer GPUs (graphics processing units), the most powerful of which are nearly as capable as AI-specialized chips. Depending on the political goals of the agreement, a hardware-centric governance approach may have to either govern the most powerful GPUs or limit their AI capabilities at the hardware level.

[5] See also Konstantin Pilz, Lennart Heim, and Nicholas Brown, 'Increased Compute Efficiency and the Diffusion of AI Capabilities' (arXiv, 13 February 2024), https://doi.org/10.48550/arXiv.2311.15377.

the ability of states to get away with incomplete or inaccurate declarations (e.g., having undeclared sites conducting operations in violation of an agreement).

2. Hardware-enabled **on-chip** mechanisms, such as confidential computing, are embedded into the hardware undertaking the operations being verified.
3. Hardware-enabled **off-chip** mechanisms are verification techniques centered on hardware placed elsewhere in the stack (not on the AI chip).[6] For example, off-chip verification hardware could be placed in networking equipment, thus allowing verifiable claims to be made about network traffic. Off-chip hardware mechanisms can be used to provide nearly any kind of information, thus making them capable of either forming a full-stack verification mechanism on their own or supporting verification schemes centered on other techniques.
4. **Personnel**-centered verification can be employed to provide information about AI activities and their institutional context. The security challenges and potential unreliability of personnel-based verification make it less suitable to serve as an independent load-bearing aspect of an agreement. However, personnel-based schemes can be created comparatively quickly and provide a parallel verification scheme which could be politically useful either on its own or in combination with other verification efforts.

The deployment of *new* on-chip mechanisms faces significant challenges—including major issues with technical viability, political acceptability, and deployment timelines.[7] AI-specialized chips are among the most complex objects produced by humanity, employing a supply chain centered on extremely sensitive intellectual property and technologies—all spanning several countries. Creating new mechanisms is technically challenging and might require the active participation of key incumbents such as NVIDIA and Huawei. Furthermore, proving that the mechanisms are not being used for other covert purposes may be extremely difficult, since the closely guarded chip design and production processes are hard to scrutinize from the outside—and once a leading-node semiconductor has been built it is infeasible to verify its entire structure.[8] Even if a new on-chip mechanism can be designed and made politically acceptable, it would take several years for new chips to be produced and to form a meaningful part of all AI compute. This path to improved verifiability is possible, but it faces significant challenges that have no obvious solution.

By contrast, *off-chip* mechanisms might be designed, built, and deployed quickly—perhaps requiring as little as several months of focused efforts. Off-chip hardware-enabled mechanisms can be designed for narrow purposes and mutually verified either through non-invasive downstream tests or cooperative production (such as in trailing-node semiconductor fabrication facilities). Unlike on-chip mechanisms, off-chip mechanisms do not neces-

[6] For the sake of distinguishing these mechanisms from the commonly employed “on-chip mechanisms” concept, these mechanisms have a negative definition that essentially means any verification-related hardware placed anywhere *except* on the chip.

[7] Note that *existing* on-chip mechanisms—such as the components underpinning confidential computing—are extremely useful for verification, as detailed below. It is the prospect of adding *new* mechanisms that brings us up against these challenges.

[8] As discussed in the report, techniques may exist to do this, but they would destroy the semiconductor in the process, thus placing limits on their usefulness.

sarily need to be built using leading-node semiconductor fabrication techniques.[9] Overall, off-chip mechanisms appear to be capable of robustly contributing to the verifiability of computations and many other aspects of agreements.

All of these categories of verification mechanisms have potential roles in the verification of even the thorniest agreements. The next section outlines an ideal theoretical computational verification system and two ways that it can be approximated with existing or near-future technology.

Two approaches to verifiable confidential computing

This report describes two technical approaches which can approximate *verifiable confidential computing*—computing which perfectly protects an agent's privacy while also providing them with the ability to credibly demonstrate that all of their computations adhered to a set of rules. In theory, an approach that can achieve these information exchanges would allow verifiability with no information leakage and thus would fully address the transparency-security tradeoff. In reality, no approach will be perfect, but serious efforts to approximate this ideal appear to be possible. The two approaches described below share a common structure, with the first requiring existing on-chip mechanisms for "confidential computing" and the second requiring new verification hardware placed into networks.

A schematic illustration of both approaches is shown in Figure 2.

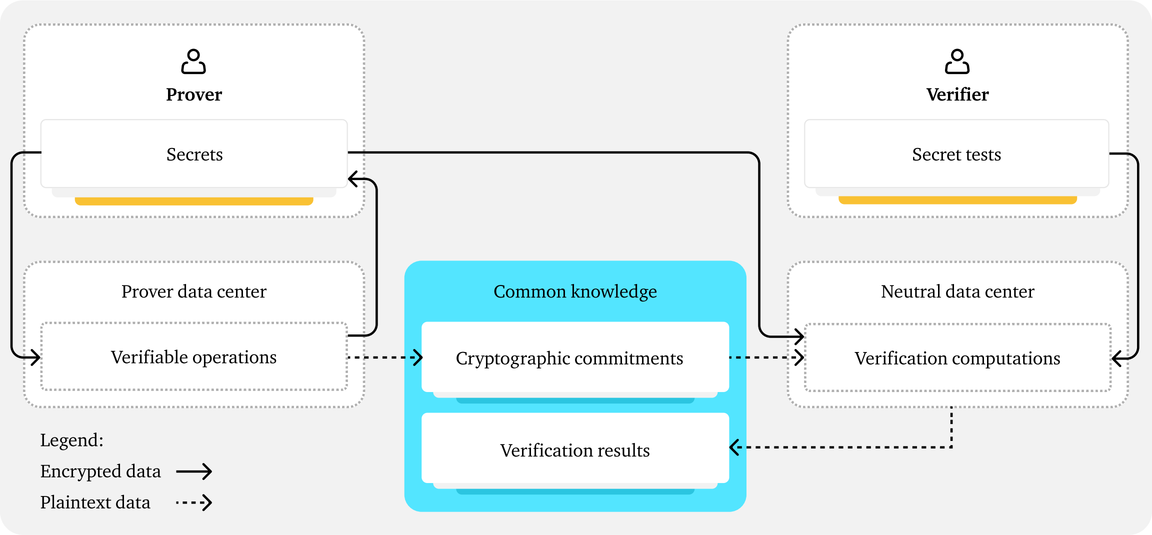

Figure 2: Schematic summary of a way to approximate *verifiable confidential computing*. This illustrates the basic structure of both of the approaches to verifiable confidential computing described in this report, since those approaches differ only in how the *verifiable operations* produce the *cryptographic commitments.*

[9] While leading performance is required for a competitive AI chip, it is not required for most functions that hardware could potentially undertake (e.g., sensors, data transmission, etc.). If leading-node fabrication is desired for a particular off-chip mechanism, it will be necessary to create a way to either verify or cooperatively produce a leading-node semiconductor—both of which appear to be open problems as of this writing.

Both approaches share a dependence on 1) cryptographic commitments, 2) a neutral mutually verified data center, 3) hardware inspection and monitoring, and 4) cutting-edge security.[10] Cryptographic commitments are short strings that "commit" to a piece of plaintext data without revealing the actual data—thus allowing an actor to later reveal the true data and simultaneously prove that it was the same data that was committed to earlier.[11] A neutral mutually verified data center is a data center that parties to an agreement cooperatively build, maintain, and secure with the express purpose of running high-sensitivity computations with neither party having access to the data. Both schemes below involve a) computational processes producing cryptographic commitments about the data involved (including all code and parameters), b) which can then be used to demonstrate within a neutral mutually verified data center that the true data has been revealed, c) thus allowing privacy-preserving computational processes to be run on the true data to determine whether they are compliant with rules. To provide mutual assurance that all hardware throughout all relevant data centers is configured correctly and not subject to physical circumvention attacks, all hardware involved in these schemes is presumed to be mutually inspected and then subject to continuous monitoring.

The first approach employs "confidential computing" features that are already available in leading compute hardware and will be more widely available in the future.[12] Confidential computing allows a set of actors (here termed the Prover and Verifier) to undertake computations using private data in which only the computational results are available to the actors (not the data itself). Overall, technologies like confidential computing allow the Prover and Verifier to mutually review code and run tests against each other's data. This kind of privacy-preserving information exchange has significant potential for technical verification. Aspirationally, it may even provide a way to answer the age-old question of "Who watches the watchers?" since mutual verification among a set of competent actors may be sufficient to demonstrate not only compliance with a set of rules, but also that all tools and data used within the system are true to purpose and not secret attempts to circumvent or undermine the governance system. Given that confidential computing features are already available on some of the leading AI chips, there is potential for a rapid rollout of this approach to govern an important portion of AI compute.

The second approach leverages networking hardware and hardware enclosures to make cryptographic commitments about all traffic. In theory, such cryptographic commitments can be just as credible as those made via confidential computing, since hardware configurations

[10] On security, the reader is encouraged to consult Sella Nevo et al., 'Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models' (RAND Corporation, 30 May 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

[11] Less technically, a commitment lets you prove that pieces of data are identical without revealing the data itself.

[12] The term "confidential computing" is used in this report to refer to not only the existing confidential computing standard, but also the entire family of technologies that can similarly enable credible multi-agent remote attestation. A full exploration of this family of technologies and techniques is beyond the scope of this report. For more on this subject, see Patrick Jauernig, Ahmad-Reza Sadeghi, and Emmanuel Stapf, 'Trusted Execution Environments: Properties, Applications, and Challenges', IEEE Security & Privacy 18, no. 2 (March 2020): 56–60, https://doi.org/10.1109/MSEC.2019.2947124.

could be inspected closely—and then monitored—to ensure that the verified networking hardware is the only such hardware available. While this scheme employs very mature technology families such as networking technology and cryptographic commitments, there are many unanswered questions about the difficulty, cost, and time involved in deploying this scheme at scale. This approach therefore requires further research and development.

These two approaches can each provide a way for a Prover to demonstrate that their computations are compliant with rules. It also appears to be possible to implement both approaches in parallel, thus providing two independent mechanisms—backed by different roots of trust—for making verifiable claims. Much more exploration is needed of these approaches and potential alternatives, but there is room for cautious optimism that at least one technical approach to verifiable confidential computing will be workable.

Political options and tradeoffs

In addition to their presumed overall prioritization of security and prosperity, states will face a number of political tradeoffs in their negotiations of AI agreements. Several such tradeoffs relate directly to verification and are thus explored briefly in this report. In their consideration of a complex AI-related agreement, such as an agreement to regulate data center-based AI development and inference, states must make several choices.

1. Which parts of the AI value chain—such as training, fine-tuning, and inference—should be governed and verified? The choice of stage to govern has ramifications for the AI ecosystem, including what AI tools can be created legally and which institutions might bear the financial and legal costs of compliance.
2. Should aspects of enforcement be embedded in the agreement directly to change the difficulty of the verification problem or to make the political agreement more robust?
3. What level of certainty of compliance is needed—and what costs in terms of money, security, delayed implementation, or slower AI computations are the various states willing to pay for a marginal improvement in certainty?
4. How gradual must agreement implementation be—and how delayed must verification processes be in relation to the activities they are verifying—in order to sufficiently alleviate the security concerns of all states?
5. Would automated code-centric verification be sufficient for the political purposes of the agreement or will human judgment be needed for all verification operations or as a pathway for rare escalation when automated systems provide ambiguous signals?
6. Where will crucial parts of the verification system—such as data centers for running verification computations or licensing systems—be located? The location of crucial hardware can shape the political viability of agreements and how states perceive the possibility of one state (or all) exiting the agreement.
7. How much AI hardware must be regulated? Key hardware families include new AI hardware, legacy AI hardware, and commodity chips such as gaming GPUs. Exempting less-

performant hardware from governance allows a significant proportion of AI work to be ungoverned—at least initially—but makes the agreement much easier to implement.

The many dimensions of political choice can be daunting, but they are also a reason for cautious optimism. When bargaining can be undertaken in many dimensions simultaneously—and when many of these dimensions allow for fine-grained choices—there is likely to be a combination of political choices that allow an agreement to fall within the zone of possible agreement for all states.[13] Similarly, there is reason to believe that a flexible set of parameters will allow states to bargain their way toward a stable equilibrium that is near the Pareto frontier, thus minimizing overall costs while maximizing overall gains.

Areas for further work

Meaningful further work on AI verification can be done on a number of different fronts, but only a few will be highlighted in this summary. Four areas will be outlined: research and development, actions that any institution can take, unilateral state action, and cooperative state action.

Research and development would be particularly valuable on seven general directions:

1. exploring the potential and limits of confidential computing—with a particular focus on resilience to extremely sophisticated cyber attacks;
2. designing and building hardware mechanisms for providing cryptographic commitments about network data;
3. collating general techniques for inspecting data center hardware as well as ways to provide ongoing monitoring to ensure that local hardware attacks are infeasible;
4. outlining a roadmap toward a neutral mutually verified data center that can realistically undertake politically sensitive computations while keeping that data robustly safe from all actors;
5. exploring the possibility of verifying and monitoring containerized data centers which can be moved to undisclosed locations to provide security assurances for security-optimizing institutions such as militaries;
6. exploring both the limits of verifying hardware that has already been created and the prospects for verifiable cooperative production of hardware at either leading- or trailing-node fabs;
7. expanding and diversifying the work on verifying whether rules are followed by digital objects—presuming plaintext access within existing privacy-preserving frameworks such as confidential computing—to include not only models but also other information such as training data, algorithm code, and inference exchanges.

[13] Paul Poast, 'Issue Linkage and International Cooperation: An Empirical Investigation', Conflict Management and Peace Science, no. 3 (2013): 286–303.

Any institution can support the development of the AI verification conversation. In particular, it would be very valuable for one or more open projects to develop transparent verification infrastructure and processes which could be leveraged by other actors. Industry players may wish to actively support such efforts in order to help foster regulatory interoperability between jurisdictions in which they do business. Another area for valuable open development would be evaluations and meta-evaluations—which are techniques for demonstrating that private evaluation code and data actually accomplish their declared goal and not some secret other goal. Even if meta-evaluations can never be perfect and even though adversaries can certainly "teach to the test" against open standards, such work has the potential to make falsified evaluations much harder to create and can thus help states as they seek to verify each other's testing frameworks. Finally, any institution, in addition to supporting direct work on these problems, can galvanize broad interest via repeated contest-like incentives such as bug bounties to find issues with proposals.

Unilateral state efforts should focus on 1) building their domestic capacity to understand the AI ecosystem and the prospects for AI verification; 2) developing their own evaluation suites (see above) while keeping at least some aspects of these suites secret to avoid allowing AI developers or other states too much ability to find ways to circumvent these tests; 3) supporting the creation and rollout of interoperable verification standards to either aid their own domestic AI industry or galvanize foreign players to create secure and verifiable AI services that they can use safely; and 4) avoiding colocation of AI facilities with other sensitive equipment, such as cutting-edge military hardware, which can help allow future inspections and monitoring of AI hardware without severe security concerns.

Cooperative state efforts can begin with the following three policies, each of which can grow from a unilateral or minilateral effort into a broader cooperative effort as the politics of verification evolve. First, states can expand their tracking of key inputs to AI such as AI-specialized chips and the equipment for producing them. Second, states can deliberately carve out space for international academic and civil society discussions on verification to enable these non-governmental communities to build common understanding of the problems and potential solutions. Third, states can monitor the situation and share information with trusted partners as they consider when to engage more broadly. Progress in AI may be extremely rapid, so states may need to make substantial efforts not only to keep up with the changing technological frontier, but also to plan ahead for the potential political demands of tomorrow.

In closing, it should be noted that for the optimistic predictions of this report to come about, political will is required. Absent political will to create and deploy verification mechanisms across key infrastructure, AI computations will remain largely unverifiable. If key states get serious about these problems, a combination of unilateral, collaborative, and open efforts should be sufficient to enable the creation of a robust verification system within a few years.

1 Introduction

As the benefits and risks of artificial intelligence (AI) have become increasingly salient, discussions have begun about how governments can act to bolster markets and avert risks. Stakeholders in these governance conversations must grapple with a remarkably fast-moving, general-purpose technology that has the potential to affect every area of human life. One crucial thread of these conversations is the prospect of international agreements that relate to AI and what they can accomplish. For many such agreements, their potential hinges on the extent to which states can *verify* whether other states are abiding by their commitments.[14,15] This report examines the verification of international agreements relating to AI, with a particular emphasis on approaches that might be technically and politically feasible in the near future.

The viability of international agreements can hinge on their verifiability because in many cases states would not want to bind their own actions unless their counterparties were similarly bound. If defection from an agreement cannot be detected, then the agreement will be unlikely to have a significant effect on state action.[16] However, if states can reliably detect defection, agreements might be possible. Since verification mechanisms are focused on revealing information about compliance, these mechanisms can open up political options—thus expanding the reach of the "art of the possible".[17]

This report is structured around the idea that verification mechanisms are best understood in the context of the agreements that they are supposed to be verifying. Myriad political factors will affect states' choices about which verification techniques to implement for a given agreement.[18] So while verification mechanisms—and verification technologies in particular—can indeed open up political options, politics will also determine which verification mechanisms are workable in a given circumstance. Therefore, the politics and the technological frontier of verification must be considered together when trying to map the space of possibilities.

Both unilateral and cooperative verification schemes are possible. Unilateral verification typically presumes that no formal or even tacit agreement exists that would enable or ease verifi-

[14] In this report, a "state" is a country.

[15] Dialogues between global academics have recently highlighted the particular need for AI verification. Bengio, Yoshua, Andrew Yao, Geoffrey Hinton, Zhang Ya-Qin, Stuart Russell, Gillian Hadfield, Mary Robinson, Xue Lan, et al. 'IDAIS-Venice' (International Dialogues on AI Safety, 2024), idais.ai/dialogue/idais-venice/.

[16] A particularly challenging domain is agreements over cyber weaponry. Unfortunately, cyberweapons can be developed and stockpiled in secret, thus making it infeasible to make robust agreements about their creation. Even worse, states often cannot discover who is attacking them in the cyber realm, since attackers can obfuscate the source of attacks.

[17] The broad definition of verification used in this report is inclusive of information activities that are also referred to as "reporting" or "monitoring". This report will use the term "verification" throughout to indicate its emphasis on the more robust end of the spectrum of such mechanisms.

[18] In this report, verification "techniques" are a more general category than "technologies". Here, "techniques" is similar to the concept of a "mechanism", referring to any way in which things could work. Meanwhile, "technologies" refer to ways of doing things, with an emphasis on enablers such as hardware and software. For example, social structures such as institutions are not referred to as technologies within this report.

cation processes. Cooperative verification refers to scenarios where states have at least some incentive to demonstrate their compliance to one another. While unilateral verification is primarily constrained by the available technologies and the domain of interest,[19] cooperative verification is typically a balancing act between diverging political needs. For cooperative verification, it's productive to consider the politics of verification in both directions. Not only is the Verifier trying to determine whether the Prover is complying with the agreement, but the Prover might also be trying to credibly demonstrate their compliance in a way that they deem to be politically acceptable.[20]

The scope of AI considered in this report includes both data center-based models and mobile AI-enabled devices.[21] Therefore, this report relates to a wide spectrum of automated capabilities, including not only models that require large-scale computing hardware, but also smaller models that can be deployed widely on devices for economic, industrial, or military purposes. Furthermore, some of the agreement and verification types discussed here are centered on key inputs such as computing hardware, data, and infrastructure—both as mechanisms for governing AI and as targets for agreements in their own right.[22]

This report includes analysis of both the civilian domain and highly sensitive domains such as state military and intelligence efforts. While the fast-moving civilian domain has been the target of most governance proposals thus far, analogous governance discussions for military AI remain underexplored.[23] This allocation of attention is expected given that AI's frontier has been driven forward by civilian institutions and purposes. However, as AI begins to be woven into all aspects of life, it will also become increasingly important for highly sensitive parts of the state apparatus. These changes warrant attention. Moreover, these sensitive domains present political difficulties that are substantially different from those in the civilian realm. As will be explored at length in this report, states can face a transparency-security tradeoff in their attempts to make verifiable deals with each other, where relatively robust verification possibilities can come at the cost of state security.[24] On the other hand, it is also possible that concerns about state security will be the very issues that will bring states to the table in the first place, just as was the case with nuclear arms control during the Cold War.[25]

[19] In an extreme scenario, if all relevant activities are clearly visible from space and all states have satellites, there would be no need for cooperative verification. Relatedly, early 20th-century naval agreements such as the Washington Naval Conference had no need for cooperative verification mechanisms because it was deemed infeasible to hide naval buildups from the other states.

[20] In this report, a Prover is a state that claims to be attempting to demonstrate their compliance with an agreement. A Verifier is an institution (e.g., another state or an international institution) that is assessing the Prover's compliance.

[21] "Mobile" here refers to devices that can be moved around. The key examples explored in this report are AI-enabled weapons.

[22] For example, the report will discuss deals where states choose to apportion AI development and deployment infrastructure according to an agreement.

[23] Matthijs M. Maas and José Jaime Villalobos, 'International AI Institutions: A Literature Review of Models, Examples, and Proposals', SSRN Scholarly Paper (Rochester, NY, 22 September 2023), https://doi.org/10.2139/ssrn.4579773; Robert Trager et al., 'International Governance of Civilian AI: A Jurisdictional Certification Approach' (Oxford Martin AI Governance Initiative, 2023).

[24] These institutions might also be particularly sensitive to regulation of their behavior due to a perception that any limitations are likely to handicap the state. This political factor is salient in the discussions in Section 1.3.

[25] This is discussed further in Section 1.3.

This report aims to make three contributions:

1. It describes **types of international agreements** that have not yet been discussed widely in the literature.
2. For each agreement type, it explores workable **verification mechanisms.**
3. Based on political needs and the technical frontier, it describes **opportunities for near-term research, investment, and policy to improve prospects** for AI verification.

The report has five sections. First, the remainder of the introduction will clarify the report's scope and then go on to summarize why international agreements over AI are desirable, why some of them need to be verifiable, and why designing verifiable agreements over AI is challenging. Second, a curated set of verification components is surveyed and analyzed. These are the ingredients for any recipe for AI verification, and they each have diverging limitations and strengths. Third, a set of political options and tradeoffs are examined for AI verification. For any given international agreement relating to AI, political choices may need to be made on some or all of these dimensions. Fourth, families of potential international agreements are described along with particular implementation options, as well as verification options that could work for those implementations. The agreements analyzed include agreements that provide international regulation of AI development and deployment. Fifth, the conclusion summarizes the key takeaways of the analysis and outlines priority areas for future work.

1.1 Scope

1.1.1 AI types

This report aims for a broad perspective, but it does emphasize certain kinds and uses of AI more than others. Specifically, we focus on a) large-scale AI[26] and b) highly sensitive uses of AI. The emphasis on large-scale AI is for two reasons. First, large models have been responsible for the majority of the rapid advances in AI capabilities in recent years, and they have been the focus of the corresponding discussions of risk and governance.[27] Second, large models and large-scale AI projects[28] are relatively easy to govern and verify compared to

[26] Multiple terms exist for large-scale AI efforts that might be correspondingly powerful and risky. These include "frontier AI", "transformative AI", and "advanced AI" as well as the more speculative "artificial general intelligence" and "superintelligence". This paper will employ none of these terms explicitly, since each term is primarily useful in other domains (such as domestic regulation) or is oriented around technical or philosophical claims rather than political impacts. This report emphasizes the international politics of AI and thus employs only generic terminology with one exception: the introduction of *systemically risky AI* in Section 4.3 (see also Appendix H) to denote AI systems with the potential to create extraordinary dangers for multiple states.

[27] ChatGPT and GPT-4 from OpenAI were widely recognized as significantly shifting the frontier of AI capability and usability. A flurry of similarly large models have since been released. Overall, these broadly capable models appear to be the primary impetus for much of the society-wide conversation about AI's opportunities and risks.

[28] Note that large-scale AI projects can involve the intensive use of resources without necessarily involving the creation of large models: for example, a project might require large amounts of AI inference using an existing model.

their smaller cousins.[29] However, it should be noted that this emphasis on large models does not imply an emphasis only on large clusters of AI-specialized compute such as large data centers. While large clusters of compute will certainly be relevant for certain kinds of governance and verification, the fact that leading AI models can be trained across multiple data centers means that, even for very large training runs, Verifiers will need to be able to verify whether trainings being completed within different data centers at different times are for the same model or for unrelated models.[30] Concretely, this report assumes that distributed training is workable with minimal efficiency losses.[31] In the same vein, this report assumes that algorithmic progress could be rapid, thus enabling greatly expanded capabilities per unit compute.[32] In sum, despite a slight emphasis on the governance of larger models, this report aims to propose verification mechanisms that could scale up in a way that allows the verification of rules even for relatively small models that are produced in great quantities.[33]

The second emphasis is on relatively sensitive uses of AI. As will be explored in more detail later, there are strong reasons to believe that in highly sensitive domains—such as state military or intelligence activities—verification of governance rules will be dramatically harder. The nature of these domains means that verification mechanisms will need to be able to do their job effectively while minimizing the revelation of information that is unrelated to the agreement. Very few verification mechanisms have characteristics that make them workable in such domains, and by comparison, low-sensitivity domains can reasonably be verified using a broader array of tools. For that reason, and because of the importance of these domains, the report conservatively emphasizes analysis of highly sensitive domains.

1.1.2 Agreement types

This report outlines a number of families of international agreements. Agreement types were chosen for inclusion based on two filters: 1) that the agreements need to be *international*

[29] The largest models as of this writing require tens of thousands of cutting-edge AI chips to create—costing a total of hundreds of millions of dollars. The creation of large models thus has a large associated footprint in computational power, money, energy, and physical space. By comparison, once created, open source large models can be modified and deployed using widely available consumer hardware, and therefore small changes can be made and minimal inference conducted for only thousands of dollars. Therefore, the creation of small models, or the deployment (as opposed to creation) of larger open source models, may have footprints that are drastically smaller. Thus these are much harder to govern.

[30] 'Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context' (arXiv, 8 August 2024), http://arxiv.org/abs/2403.05530, page 7.

[31] This choice is made for three major reasons. First, this assumption makes the verification problem harder, so if this assumption turns out to be wrong we will still have laid out plans that could work. Second, distributed training has already been demonstrated with Gemini (Georgiev et al., 2024) as well as with other recent research results (e.g, Jaghouar et al., 2024 and Peng et al., 2024). Third, calculations in Scher and Thiergart (2024) indicate that it is difficult to rule out the potential of distributed training, thus suggesting that relatively efficient distributed training is likely to be tractable.

[32] This is also the conservative assumption, since if we assume no future algorithmic progress, we could be left flailing if rapid advances occur. see Appendix E for more about the challenges for verification and some potential solutions. Furthermore, there is ample evidence of rapid historical and ongoing progress in algorithmic efficiency. Anson Ho et al., 'Algorithmic Progress in Language Models' (arXiv, 9 March 2024), https://doi.org/10.48550/arXiv.2403.05812.

[33] Important caveats to this point are described in the report, including important tradeoffs regarding the amount of compute that might need to be governed (Section 3.2) and the underexplored challenge of verifying either AI agents or AI systems that are designed to use external tools (Section 2.5.3).

and 2) that verification of the agreement must be both *possible in theory* and *provide non-trivial assurances*. Agreements are further organized into families on the basis of their verification requirements, not other political dimensions. Each of these points will be expanded in turn.

International agreements are agreements that involve *at least two states*.[34] This report does not presume that any particular states are involved in the agreements.

Only certain agreements are verifiable in theory but also have non-trivial verification needs. Verifiability-in-theory means that compliance or non-compliance with the agreement would lead to evidence that could in theory become known to the Verifier. For example, unverifiable agreements could relate to claims about *unobservable* activities or about the *intent* of actions, rather than claims about the observable implications of actions. While declarations of intent can be valuable for building and maintaining norms, they are not verifiable claims. Thus, this report will not consider agreements that are solely about mental states such as declared beliefs or intents. Instead, the report will focus on agreements with observable structure or outcomes which can be perceived.[35] For example, an agreement that stipulates rules about the location and disposition of specific physical objects would be verifiable in theory.

Furthermore, agreements that are trivially verifiable will not be explored in this report. Trivially verified agreements include agreements with obvious and readily observable consequences, such as money being sent to an account, a model being published, or a shipment of chips being delivered to a state. Any action that can be fully understood quickly and easily due to its immediate and overt real-world consequences does not need a verification regime.[36] An important caveat is that some overt actions can contain subtle deception. For example, a delivered set of chips might have been secretly modified to serve the sender's strategic goals and undermine the receiver's goals.[37] Efforts to ascertain whether a declared agreement is being secretly undermined are certainly within the scope of this report. In sum, this report considers agreements with consequences that can be legible to another state. It therefore includes categories such as knowledge or resources transfers, the pooling of resources, preparation for emergencies, and regulation (see Section 4).[38]

The question of whether agreements are verifiable in theory should not be confused with the question of whether these agreements are verifiable in practice. As described above, *verifiable-in-theory* means that a perfectly privacy-preserving verification mechanism would have the potential to verify that agreement.[39] Whether an agreement is *verifiable-in-practice* is a more multifaceted question, and grappling with this question is the primary goal of this report.

[34] This includes agreements that are bilateral—between two states; minilateral—involving small groups of states; and multilateral—aiming for broader or universal participation.

[35] There are many potential challenges in verifying consequences, including that consequences might be delayed, subtle, or deliberately hidden by the Prover.

[36] These kinds of agreements may be very useful complements to the agreements discussed in this report.

[37] Another example is Provers providing information that is deliberately falsified in a way that is difficult to catch. See Section 4.1.

[38] Tacit agreements could in theory be verified in some cases, but this report does not deeply explore this domain other than a discussion of the limitations of unilateral verification for AI regulation (see Section 1.5.3).

[39] As explored throughout the report, computational methods for such mechanisms already exist in some domains.

The report will focus on the potential for *robust* verification of the various agreement types in scenarios where states are paradoxically seeking to demonstrate compliance to each other while simultaneously seeking ways to circumvent verification.[40] Unilateral verification capabilities such as national technical means are also examined as part of the equation. Overall, the emphasis is on verification approaches which can robustly, even if imperfectly,[41] demonstrate compliance among states and other actors.[42] This report emphasizes families of verification techniques that appear to be well-suited for the political goals of states, and have the potential to reassure both sides about their mutual compliance. Other mechanisms that are less reliable, or that are likely to be politically impossible, will be mentioned but not emphasized.

Agreements are organized into families according to similarities in their verification requirements rather than other aspects of the agreement, such as their political structure, effectiveness, distribution of benefits, political feasibility, or enforcement. Other work has explored these questions,[43] and we provide a different way of organizing the discussion, which is not intended to be definitive or exhaustive.[44] This also means that the detailed political content of agreements—such as specific regulatory rules—will not be deeply explored in this report. As noted earlier, verification capabilities open up a space for political options. Therefore, the techniques discussed in this report are also intended to help future discussions about political agreements succeed.[45] Moreover, even within the realm of verification, political questions abound. These latter questions are expanded in Section 3.

1.1.3 Verification mechanisms

This report contains a survey of some verification approaches for AI. This survey is broad but not exhaustive, and aims to provide an overview of both a) the potential and limitations of the various kinds of mechanisms and b) why specific mechanisms appear to be well-suited for verifying particular kinds of agreements.

[40] This framing is sometimes termed the "covert adversary". A covert adversary is a theoretical actor who will comply with verification protocols only to the extent that they could be caught attempting to circumvent those protocols. In short, they'll cheat if they can get away with it. See Yonatan Aumann and Yehuda Lindell, 'Security Against Covert Adversaries: Efficient Protocols for Realistic Adversaries', in Theory of Cryptography, ed. Salil P. Vadhan (Berlin, Heidelberg: Springer, 2007), 137–56, https://doi.org/10.1007/978-3-540-70936-7_8.

[41] US Secretary of State George Shultz testified in 1988 regarding the INF treaty that "There is no such thing as absolute, 100 percent verification. But it is our judgment that this treaty, through its successive layers of procedures, contains the measures needed for effective verification.... The bottom line is that the verification provisions of this treaty get the job done." Rose Gottemoeller, 'Looking Back: The Intermediate-Range Nuclear Forces Treaty' (Arms Control Today, 2007), https://www.armscontrol.org/act/2007-06/looking-back-intermediate-range-nuclear-forces-treaty.

[42] The presumption is that less robust mechanisms will be easier to design and implement. While these may be very valuable to develop in order to serve particular political needs, this is not our focus.

[43] In addition to specific proposals, work comparing approaches includes; Matthijs M. Maas and José Jaime Villalobos, 'International AI Institutions: A Literature Review of Models, Examples, and Proposals', SSRN Scholarly Paper (Rochester, NY, 22 September 2023), https://doi.org/10.2139/ssrn.4579773; Robert Trager et al., 'International Governance of Civilian AI: A Jurisdictional Certification Approach' (Oxford Martin AI Governance Initiative, 2023).

[44] This report's focus on the verification of international agreements necessarily means that many other meaningful political dimensions will be underexplored or left out entirely. This prioritization reflects the goals of this report, not the overall importance of the different dimensions.

[45] Inclusion of an agreement type in this report should not be regarded as an endorsement.

Table 1.1: Agreements examined in this report.

Agreement family	Agreement types	Variants
Transfer knowledge	Share research	
	Share knowledge of AI risks and opportunities	
	Share AI development knowledge	
	Share safety-enhancing technologies	
Transfer resources	Transfer development resources	Share AI-specialized chips Share access to AI-specialized compute Training programs for AI professionals
	Provide access to AI systems	Transfer completed models Provide API access
	Share benefits	Cash transfers Deploy AI-enabled devices as aid Transfer AI-enabled devices
Pool resources	Pool resources toward an international goal	
	Pool resources toward defensive AIs	
	Pool resources toward transformative AI	
	Pursue systemically risky AI only in a singular project	
Prepare for emergencies	Computational emergency detection and response	
Regulate	Regulate AI development	Regulate data center-based AI development Regulate fine-tuning and online learning
	Regulate AI deployment	Regulating data center inference Regulating sensitive mobile AI-enabled devices

We include both widely-used mature concepts and technologies, as well as approaches which are speculative to varying degrees. However, we limit discussion of the most speculative methods, instead focusing on verification mechanisms that are available already or can potentially be made available soon (e.g., within a few years).

Mechanisms will also be examined in terms of their political characteristics, including a) their ability to verify agreement(s), b) their ability to address transparency-security tradeoffs, and c) their robustness against covert circumvention attempts. This report does not explicitly examine confidence-building measures,[46] although many of the techniques described herein would also be applicable to such efforts.

1.2 Methodology

The ideas in this report were generated through a review of the relevant literature and a series of informal cross-disciplinary idea exchanges and conversations. Overall, this document attempts to be a synthesis of the key areas of related disciplines (e.g., cryptography, machine learning, and verification in International Relations) as well as an analysis of how the knowledge of these disciplines can be brought together to shape our sense of what is possible in AI verification today, as well as what might be possible in the future. Omissions and errors are inevitable in a work of this scope. Nonetheless, the authors hope that this report informs and broadens future conversations in this crucial field.

1.3 Political needs and political will

Detailed discussions of the political needs for international AI governance are outside the scope of this report. Nonetheless, to highlight the importance of the topics at hand, we provide a brief list of reasons why verifiable international AI governance might be desirable for states. States might work together via verifiable international agreements to create new economic opportunities, preserve peace, and mitigate other risks.

Creating new economic opportunities is a clear and salient priority for many state governments today.[47] By creating effective and harmonized regulations, states can reap the potentially enormous benefits of AI.[48] Absent such harmonized regulations, cross-border trade in AI-related goods and services might be sharply limited, and entire sub-industries may fail to

[46] Confidence building measures, such as used by the Biological Weapons Convention, are useful in the absence of an agreed-upon regime, and may function as adjuncts to verification, but are not themselves intended to be verified or verifiable. See also Michael C Horowitz and Paul Scharre, 'AI and International Stability: Risks and Confidence-Building Measures' (Center for a New American Security), accessed 25 October 2024, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[47] 'Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet' (Paris AI Action Summit, 11 February 2025), https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet.

[48] Claire Dennis et al., 'What Should Be Internationalised in AI Governance?' (Oxford Martin AI Governance Initiative, 2024).

reach their full potential.[49] The creation of robust standardized rules for interoperable economic systems has been enormously influential in other domains, such as civil aviation or maritime shipping. A failure to regulate and standardize internationally might have stymied the growth of the networks for aviation and shipping that now span the globe. Similarly, the AI industry could be hampered by fragmentation, regulatory arbitrage, and the potential for industry-wide reputational damage due to a disaster. This final possibility is particularly instructive, given how the Three Mile Island and Chernobyl nuclear accidents affected the prospects for civilian nuclear energy.[50]

Preventing war is also of central concern. At least two pathways to war are already salient today. First, arms races can be extraordinarily expensive and lead to rapidly increasing perceptions of mutual threat. Today, a race toward leadership in civilian AI is already underway and AI is increasingly perceived to be central to military power.[51] An unconstrained arms race in this domain could consume vast resources that states would otherwise prefer to put toward other uses—thus raising the risk of war due to the desire to reduce their need for arming in the future.[52] Furthermore, major investments in AI-enabled military hardware may significantly increase the threat that states perceive from each other's weaponry.[53] Second, racing can also create systemic pressures toward a major war: the pursuit of greater relative power by one state can incentivize other states to wage preventive war to stave off what they perceive to be a major shift in military power. Concerns of this general shape are certainly not new,[54] but some policy elites consider this a real possibility for the near future.[55] In sum, these two pathways mean that failing to adequately govern AI internationally could lead to costly and dangerous arms races or even a major war.

[49] Sub-industries that might fail to reach their full potential without AI verification in particular could include AI services that process sensitive personal, business, or state data or which provide high-sensitivity services (e.g., government services or military & intelligence capabilities), to states themselves.

[50] Bulat Aytbaev et al., 'Don't Let Nuclear Accidents Scare You Away from Nuclear Power', Bulletin of the Atomic Scientists, 31 August 2020, https://thebulletin.org/2020/08/dont-let-nuclear-accidents-scare-you-away-from-nuclear-power/.

[51] Nestor Maslej et al., 'The AI Index 2025 Annual Report' (AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2025); James Black et al., 'Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence' (RAND, 2024).

[52] As explored in various parts of the International Relations literature, states might start wars because maintaining the status quo is too costly for them. Robert Powell, 'Guns, Butter, and Anarchy', American Political Science Review 87, no. 1 (1993): 115–32; Andrew Coe, 'Costly Peace: A New Rationalist Explanation for War' (2011).

[53] This is particularly true if AI has an effect on the overall offense-defense balance of strategic weaponry. Unfortunately, there is no guarantee that the dominance of defensive nuclear deterrence that has undergirded peace for decades will remain unchanged as AI enters the equation. See also Chris Meserole, 'Artificial Intelligence and the Security Dilemma', Lawfare (Lawfare, 4 November 2018), https://www.lawfaremedia.org/article/artificial-intelligence-and-security-dilemma; James Johnson, 'AI-Security Dilemma: Insecurity, Mistrust, and Misperception under the Nuclear Shadow', in AI and the Bomb: Nuclear Strategy and Risk in the Digital Age (Oxford University Press, 2023), https://doi.org/10.1093/oso/9780192858184.003.0005.

[54] In 1960, Thomas Schelling wrote, "A nation known to be on the threshold of an absolutely potent surprise-attack weapon may have reason to forswear it unilaterally—if there is any possible way to do so—in order to forestall a desperate last-minute attempt by an enemy to strike first while he still has a chance." Thomas C Schelling, The Strategy of Conflict (Harvard University Press, 1960), p. 133.

[55] Along those lines, Schmidt et al. argue that "[i]n the event that the identity of a winner [of the AI race] does crystallize, mere competition could devolve into conflicts driven by desperation and fear," to the extent that "some states may seem the advent of AI threatening enough to demand a nuclear response." Henry Kissinger, Eric Schmidt, and Craig Mundie, Genesis: Artificial Intelligence, Hope, and the Human Spirit (Little Brown and Company, 2024).

Other risks may also be a clear priority for many states, including the proliferation of dangerous capabilities, exacerbated inequality, and large-scale harms due to accidents or misuse. Terrorists or rogue states in possession of dangerous capabilities might be particularly likely to use them against other states.[56] On the flip side, centralization of AI's benefits could greatly exacerbate economic inequality both within and among states—which in turn can lead to political destabilization and even conflict. In response to a similar set of political needs with other technologies, states have constructed governance regimes that manage the danger of proliferation while allowing civilian use of the same technologies.[57] In particular, International Atomic Energy Agency (IAEA) safeguards have facilitated civil nuclear transfers by providing assurances to states that no transferred technology or material will be used for nuclear weapon development. A set of related multilateral export control regimes for sensitive technologies and materials was built in the decades after the IAEA safeguards regime took shape.[58]

Finally, the potential for AI to have scalable effects on the world means that AI accidents or misuse might cause harm on a vast scale.[59] Preventing both accidents and misuse is extremely difficult due to fundamental aspects of the technology and how it is being employed.[60] Moreover, competition in the economic and military domains makes these problems more likely.[61] Driven in part by increasingly urgent messaging from civil society and industry leaders, governments are increasingly recognizing these issues.[62] To manage these

[56] Markus Anderljung and Julian Hazell, 'Protecting Society from AI Misuse: When Are Restrictions on Capabilities Warranted?' (arXiv, 29 March 2023), https://doi.org/10.48550/arXiv.2303.09377.

[57] Jane Vaynman and Tristan A. Volpe, 'Dual Use Deception: How Technology Shapes Cooperation in International Relations', International Organization 77, no. 3 (March 2023): 599–632, https://doi.org/10.1017/S0020818323000140.

[58] In particular, such regimes were built for technologies and materials enabling the construction of nuclear, biological, and chemical weapons as well as long-range missiles. See 'Guidelines', Nuclear Suppliers Group; 'Common Control Lists', Australia Group; 'Equipment, Software And Technology Annex' (Missile Technology Control Regime, 14 March 2024).

[59] While most technologies are limited to local effects, AI is poised to create global effects. Two pathways to such impact include 1) interactions with our existing globe-spanning digital infrastructure and 2) the increasing intelligence and autonomy we provide to some AI systems—which might enable them to drastically expand their impacts.

[60] Brian Christian, The Alignment Problem: Machine Learning and Human Values (WW Norton & Company, 2020); Yoshua Bengio, et al., 'International AI Safety Report', 29 Jan 2025.

[61] Stuart Armstrong, Nick Bostrom, and Carl Shulman, 'Racing to the Precipice: A Model of Artificial Intelligence Development', AI & Society 31, no. 2 (2016): 201–6; Amanda Askell, Miles Brundage, and Gillian Hadfield, 'The Role of Cooperation in Responsible AI Development' (arXiv, 10 July 2019), http://arxiv.org/abs/1907.04534; Eoghan Stafford, Robert F Trager, and Allan Dafoe, 'Safety Not Guaranteed: International Races for Risky Technologies' (November 2022), https://cdn.governance.ai/International_Races_for_Risky_Technologies_DRAFT_NOV_2022.pdf; Robert F. Trager et al., 'Safety-Performance Tradeoff Model: Exploring Safety Insights in AI Competition', Modeling Cooperation, December 2022, https://spt.modelingcooperation.com/.

[62] 'Statement on AI Risk', Center for AI Safety, 30 May 2023, https://www.safe.ai/statement-on-ai-risk; 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023', GOV.UK, accessed 2 November 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

risks, governments may take a variety of actions, including some that require international cooperation in order to be effective.[63]

In sum, the political needs for international cooperation over AI may include the creation of new economic frontiers, the preservation of peace, and the mitigation of substantial risks to society. The remainder of this report aims to clarify how states can verifiably move together toward some of these goals if they choose to do so.

1.4 Importance of AI agreement verification

Verification is central to most high-stakes international agreements, and it serves three key purposes. First, a verification system aims to *detect* non-compliance, which is crucial to inform enforcement. Second, verification *deters* parties from even contemplating a deliberate violation by setting the expectation that such violations will be detected. And third, verification helps *build confidence* by permitting compliant parties to demonstrate their compliance in an open, official, systematic, and continuing way.[64]

These elements can also be illustrated by theories on how cooperative equilibria can be achieved within iterated social dilemmas. In a typical social dilemma such as the prisoner's dilemma, the all-cooperate outcome is preferred by all agents compared to the all-defect outcome—but the cooperative equilibrium is only available in specific circumstances, such as an iterated game with patient players.[65] One of the main requirements for such an equilibrium is that each agent needs to be able to detect whether the other party is defecting. This is precisely the kind of information provided by a robust verification protocol. The comparative desirability of the cooperative equilibrium is what drives agents to try to make themselves legible to one another. A further factor is the expectation of future interactions. If agents expect numerous and frequent interactions in the future, they will also expect less overall utility if they defect and the other agent detects that defection, since the other agent can be expected to punish them in subsequent rounds.[66] The expectation of punishment

[63] For example, extremely dangerous AI capabilities can be banned or verifiably placed explicitly under the command of a state. Such capabilities might include the ability to produce chemical, biological, radiological, nuclear, or cyber weapons as well as extreme social manipulation capabilities. One analogy is that no private group in the world is legally allowed to hold nuclear weapons.

[64] Coming to Terms with Security: A Handbook on Verification and Compliance (Geneva: United Nations Institute for Disarmament Research, 2003).

[65] Robert Axelrod, The Evolution of Cooperation (Basic Books, New York, 1984).

[66] In technical terms, the discount factor δ in an indefinite game can be stated as the probability of the two agents interacting again in the future. In that sense, raising the effective frequency of interactions—or the rapidity with which defection can be recognized and responded to—can push δ toward its theoretical maximum of 1. This framing highlights the rapidity of interaction and response as opposed to the common alternative framing of "patient" agents. Even fairly impatient agents can achieve cooperative equilibria if their interactions occur with high frequency.

in response to defection will increase the likelihood of cooperative behavior.[67] Ultimately, these factors mean that agents need both verification and enforcement of agreements to solve social dilemmas. This report focuses on verification but acknowledges that enforcement will also be needed.

Furthermore, this report emphasizes both agreements and verification mechanisms that focus on earlier rather than later aspects of the AI value chain. For example, we emphasize mechanisms for controlling the models that can be created in the first place, as opposed to mechanisms for controlling the use of those models. While in the analysis that follows we attempt to be systematic, we emphasize the governance of upstream factors in the production of AI systems because those factors tend to be more amenable to stable governance via iterated agreements such as those described above. This emphasis is further reinforced by theoretical and practical arguments that indicate that political agreements which focus on the "roots" of power rather than power itself—such as coal and steel resources rather than tanks—allow states to achieve more robust and useful agreements.[68] Given that there is substantial uncertainty about how powerful AI will be—either in an economic or military sense—it is also prudent to preferentially aim agreements primarily at the ways in which AI systems can be produced, since those agreements are likely to be robust to a wider range of potential futures than comparable agreements that focus only on the specific downstream effects and uses of AI.[69] The rest of this report emphasizes agreements centered on roots of power that are upstream of power itself, such as the creation rather than the deployment of models, AI-specialized compute resources, and the supply chains that create both AI chips themselves and all of their requisite inputs. While the degree to which an agreement focuses on such upstream resources is a political question and thus outside the scope of this report, this report will nonetheless emphasize iteration on upstream resources because they allow more robust verifiable agreements than similar efforts aimed at controlling downstream effects.

In some cases, feasible verification may be a necessary condition for negotiating an international agreement in the first place, especially when undetected defection is deemed unacceptable for a state's security. For example, states did not agree to the Chemical Weapons

[67] Intuitive examples of two extremes of this dynamic are interstate war and cyber attacks. Interstate war tends to involve very little ambiguity over who the parties to the conflict are—and punishment is implicit in the process via the costs of war. A state considering an overt attack on another state will know that both the defending state will correctly perceive that it is being attacked and will likely respond with its own use of force. Since war is costly, the potential aggressor knows that it will face at least some punishment if it attacks ("defection" in the language of social dilemmas). By contrast, cyber attacks do not generally allow for reliable attribution, so a state under attack generally does not know precisely who is attacking. This lack of attribution makes punishment much less credible, and thus aggressors are not as deterred by the costs of conflict in cyber attacks as they are in conventional war. These differing dynamics help clarify why interstate war is very rare while cyber attacks are ubiquitous.

[68] Thomas Chadefaux, 'Bargaining over Power: When Do Shifts in Power Lead to War?', International Theory 3, no. 2 (2011): 228–53.

[69] For example, there is no guarantee that the military use of AI will lead toward continued or renewed deterrence dominance like nuclear weapons and their delivery systems did. This uncertainty will eventually fade as the true shape of AI's effect on military affairs becomes discernible in the coming years, but for now strategic reasoning about such effects must acknowledge substantial uncertainty and thus consider a wide range of possible effects, including a substantial shift away from deterrence dominance.

Convention until the problem of verification had been solved, which took decades.[70] By contrast, some agreements cannot be fully verified due to their nature, yet states may be willing to accept them. For instance, the Biological Weapons Convention does not have a verification protocol, partly because great powers have found it unlikely that covert violations would seriously threaten their security.[71] The kinds of regulation that might be acceptable for a state is also influenced by the perceived strategic value of the weapon, including the availability of substitutes. For instance, biological weapons were perceived to have lower military utility than chemical weapons[72] and nuclear weapons.[73]

In the case of AI, demonstrating that verification is feasible may be crucial to unlocking international agreements. Given AI's potentially high strategic value, states may require significant certainty that their counterparties are not defecting from the agreement and therefore may require robust verification protocols to be part of any major agreement. If the agreement is perceived to be high-stakes, a covert defection might be regarded as an unacceptable security danger, and thus the political need for reliable verification would increase. If sufficiently reliable verification mechanisms are unavailable or perceived to be politically unacceptable, such a strategic scenario might not lead to an agreement despite such an equilibrium being more desirable for all states than the default outcome.[74] However, if defection can be quickly and accurately detected, the credible threat of reliable enforcement can create a stable cooperative equilibrium.[75]

1.5 Challenges of AI agreement verification

Three areas of background knowledge are particularly valuable for AI verification: the transparency-security tradeoff, the asymmetric burden of proof for verifying negative claims, and a broad sense of how AI is created, tested, and deployed today. Each will be explored in turn.

[70] Thomas Bernauer, The Projected Chemical Weapons Convention: A Guide to the Negotiations in the Conference on Disarmament (New York: United Nations Institute for Disarmament Research, 1990), p 19.

[71] Marie Isabelle Chevrier, 'Verifying the Unverifiable: Lessons from the Biological Weapons Convention', Politics and the Life Sciences 9, no. 1 (1990): 93–105, https://doi.org/10.1017/S073093840001025X.

[72] For example, a UN report notes the following: "The United States and the UK were of the view that the military value of biological weapons was inferior to that of chemical weapons." Thomas Bernauer, The Projected Chemical Weapons Convention: A Guide to the Negotiations in the Conference on Disarmament (New York: United Nations Institute for Disarmament Research, 1990).

[73] Regarding the relationship between biological and nuclear weapons, Richard Nixon is alleged to have said, "We'll never use the damn germs, so what good is biological warfare as a deterrent?" as well as "If somebody uses germs on us, we'll nuke 'em." David Hoffman, The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy (Anchor, 2009).

[74] Analogously, the cooperation equilibrium of the prisoner's dilemma is better for all players than mutual defection but might be unreachable unless players have sufficient ability to detect and punish defections in an iterated game

[75] Related to this, the concept of a "break out time" in nuclear verification—the approximate amount of time it would take a state to create a nuclear weapon—in turn affects calculations of the minimum politically acceptable frequency for nuclear verification processes.

1.5.1 Transparency-security tradeoff

Some verifiable agreements are intensely affected by a fundamental tradeoff between transparency and security.[76] While sufficient information transparency is necessary to verify compliance, states will attempt to limit transparency to preserve the confidentiality of information that is crucial for national security.[77] For example, on-site inspections of nuclear weapons at a military base may also reveal to inspectors important information about the host state, such as details about their capabilities and vulnerabilities—which adversaries can exploit.[78]

The severity of the tradeoff varies for different approaches to verification, and hinges particularly on the sensitivity of the information involved, the specificity of the information revealed, and the usefulness of unilateral monitoring. High-sensitivity information can make the tradeoff severe, as it was when Saddam Hussein continued to obstruct inspections by the UN Special Commission due to concerns that they might disclose sensitive information on the security regime apparatus that foreign powers could later leverage to attack his regime.[79] Information specificity is also crucial, since a verification mechanism that could demonstrate compliance without revealing any extraneous information may face no transparency-security tradeoff at all. Unfortunately, mechanisms proposed historically often required or allowed the Verifier to access information that was not directly relevant to the agreement. For example, one of the main reasons the Biological Weapons Convention lacks a formal verification protocol is that, given the thin line between offensive and defensive uses, effective verification would entail the Verifier gathering information on the civilian biotechnology industry, potentially including trade secrets.[80] A state's unilateral monitoring capabilities play a role in the transparency-security tradeoff because they shape the overall information security context in which decisions about verification mechanisms are being made. In a hypothetical scenario where unilateral monitoring—such as the actions of intelligence agencies—can be expected to have revealed all of the extraneous information that might have been acciden-

[76] Andrew J. Coe and Jane Vaynman, 'Why Arms Control Is so Rare', American Political Science Review 114, no. 2 (2020): 342–55.

[77] A similar tradeoff would of course apply to sensitive information belonging to corporations and individuals. This report does not fully explore this family of related tradeoffs. Instead, this report centers the transparency-security tradeoff since it seems to be a particularly intense tradeoff for many potential agreement types and verification mechanisms. Furthermore, if a set of mechanisms can reassure extremely sensitive institutions such as militaries and intelligence agencies that their security is being protected, then similar mechanisms can likely be used to appropriately protect commercial and personal information in most cases.

[78] Such information could be of many potential kinds. Section 1.5.1.1 discusses the sensitivity of location information for military assets. Another possibility is that a state is bluffing about its power on the international stage, and a verification mechanism might inadvertently reveal their bluff.

[79] Gregory D. Koblentz, 'Saddam versus the Inspectors: The Impact of Regime Security on the Verification of Iraq's WMD Disarmament', Journal of Strategic Studies 41, no. 3 (16 April 2018): 372–409, https://doi.org/10.1080/01402390.2016.1224764.

[80] Guy B. Roberts, Arms Control without Arms Control: The Failure of the Biological Weapons Convention Protocol and a New Paradigm for Fighting the Threat of Biological Weapons (USAF Institute for National Security Studies, 2003).

tally revealed via a proposed verification mechanism, then that mechanism does not face a transparency-security tradeoff at all—there is nothing more to reveal.[81]

Subsequent sections of this report explore the transparency-security tradeoff in further detail. While severe tradeoffs are certainly possible in the verification of international AI governance agreements, it also appears likely that workable tradeoffs can be achieved for the verification of most of the agreements explored in this report.[82]

1.5.1.1 Security sensitivity of detailed location information

One example of the transparency-security tradeoff discussed in this report is the fact that location data is useful for verification, but revealing this data can create perceived security vulnerabilities. Location data from the Prover related to personnel or digital systems could help the Verifier check whether compliance is ongoing. For example, it might be very useful for a chip-centered agreement if the Verifier could inspect data centers and find all chips in their declared locations. The problem is that if sensitive organizations like militaries believe that these chips are a crucial component of their military power, then they will not want the locations of all their chips to be knowable by the Verifier.[83] A similar issue arises if the Prover must make verifiable claims about the detailed activities of key personnel.[84]

There are reasons for cautious optimism that this challenge can be overcome. One technique that is explored later in this report allows robust verification and monitoring of resources via specialized hardware while simultaneously making that hardware impossible to locate accurately. The combination of locally stringent verification mechanisms with deliberate obfuscation of locations can allow for detailed monitoring of hardware even when that hardware is physically hidden (see Section 2.5.4.2).

1.5.2 Asymmetric burden of proof for verifying negative claims

Demonstrating or verifying the existence of an object or process is often straightforward compared to the difficulty of demonstrating the non-existence of an object or process. For example, it's trivial for a state to demonstrate that it has a key AI capability—it can simply demonstrate that capability.[85] By contrast, demonstrating that they do *not* have a given capa-

[81] A related but very different scenario involves unilateral monitoring capabilities that create enough transparency that the agreement is workable. Such scenarios occurred historically even for central military technologies, such as the naval treaties of the early 20th century and some of the nuclear arms control treaties during the Cold War. Wawrzyniec Muszyński-Sulima, 'Cold War in Space: Reconnaissance Satellites and US-Soviet Security Competition', European Journal of American Studies 18, no. 2 (30 June 2023), https://doi.org/10.4000/ejas.20427.

[82] As noted later, the verification approaches described for regulatory agreements will require serious implementation work to begin at least a year—and perhaps a few years—before full-scale verification can be done.

[83] Even below the threshold of war, a potential attacker planning a cyber attack on the Prover might benefit from knowing the locations of relevant hardware.

[84] The transparency-security tradeoff for location information might be quite salient for militaries but much less important for civilian organizations such as businesses. Therefore, high-accuracy location mechanisms might be politically tolerable for some categories of civilian technology.

[85] Similarly, the established method of proving that you have nuclear weapons is to detonate one.

bility might have a significant burden of proof; it might be challenging to produce credible evidence for the claim.[86] To acquire high confidence in the truth of a negative claim like this, a Verifier may need to be provided with significant quantities of relevant evidence of various kinds.[87]

This asymmetry shows up constantly in discussions of verification. Regardless of practical difficulties, it is epistemically straightforward for a Prover state to make itself legible to the Verifier with regards to *known* or *declared* objects, facilities, processes, and people. This cannot be said, however, for the hypothetically extant but *unknown* or *undeclared* objects, facilities, processes, or people. While a Prover might be able to provide a highly credible story about the work done by a given team, it's very hard for them to demonstrate that *no team exists* doing a particular kind of work. Similarly, while a Prover might be able to demonstrate legal codes and even enforcement actions within their domestic laws, they would have great difficulty demonstrating that they were not taking other hidden actions that countermanded these laws.

Three parts of the practice of verification grapple with this asymmetry. First, verification processes tend to be anchored on declared objects, sites, processes, and people, where robust claims are relatively tractable since there are a limited number of such declarations and they make concrete claims about reality.[88] Second, parallel efforts are made to catch omissions or inconsistencies in declarations—both to discover and to deter circumvention of agreements.[89] Third, verification processes can leverage the fact that a key input like computing power has a known and finite quantity. Given this, the larger the quantity of declared or acceptable activities that are fully accounted for, the smaller the possible scale of undeclared or prohibited activities using that input.[90]

1.5.3 Why AI verification is particularly challenging

The field of AI is vast, and thus any discussion of its verification challenges will be incomplete. This section will first compare AI with historical examples of strategically valuable technologies. Next, it will explore the challenges of achieving one of the agreement types explored

[86] This challenge is similar to the one referred to by the colloquial claim that is impossible to prove a negative. Steven D. Hales, 'Thinking Tools: You Can Prove a Negative', Think 4, no. 10 (2005): 109–12.

[87] Verification techniques for such claims can draw upon similar techniques developed in formal logic for proofs of impossibility, including the approach of *proof by exhaustion*—which aims to demonstrate a claim by exhaustively examining all possible ways to check it.

[88] Such declarations also make the AI ecosystem much more legible and possible to verify compared to an ecosystem without such declarations. For example, even with substantial access to data about hardware usage, it is very difficult to reliably classify workload types even if there are no attempts to hide the activities. See Lennart Heim et al., 'Governing Through The Cloud: The Intermediary Role Of Compute Providers In AI Regulation' (Oxford Martin AI Governance Initiative, March 2024).

[89] In general, organizing agreements around declarations can be useful for these reasons. Thanks to Mauricio Baker for making this point.

[90] A key input having a known and finite quantity is helpful towards governing AI through that input (Sastry et al., 2024) but is not sufficient on its own. For example, there could be secret production of chips, making an accounting based entirely on known chips misleading. Here still, though, there is a sense in which finitude is friendly towards verification, in that the more monetary and human capital that a country invests in known chip production, the less they will have available for secret production, so it may still be possible to make some (approximate) estimates of the scale of uncertainty introduced by these possibilities.

below—the *regulation* of AI models via rules about their development and deployment. In the process of exploring this challenge, the general outlines of the AI industry will be provided, including the key inputs to AI; how models are built and deployed; what kinds of regulations might be implemented; and finally the overall challenge of regulation. We conclude that unilateral verification alone cannot reliably verify compliance with AI regulation.

The international governance of strategically valuable technologies has a mixed history. While there have been numerous attempts to govern important technologies, including military aviation,[91] chemical weapons,[92] nuclear weapons,[93] and biological weapons,[94] where such agreements have succeeded, they have been very limited in ambition. For example, the various phases of nuclear arms control during the Cold War only managed to restrain somewhat the otherwise broad and wholesale development and deployment of these weapons. Similarly, agreements over other technologies, such as biological and chemical weapons, happened in the shadow of nuclear weapons, and at least some key actors perceived these conversations to be relatively unimportant compared to the leading strategic technology—nuclear weapons.[95] A pessimistic reading of this history indicates that no significant restraint was ever placed on strategically important military technology which had no substitutes. A more balanced reading indicates that the advent of game-changing technologies sometimes galvanized serious examination of governance structures that were unthinkable in other eras or for other technologies.[96] Overall, historical experience indicates that international governance of strategically transformative technology is very difficult. One potential basis for hope in the governance of AI is that the world of 2025 and beyond is very socially and technologically different from the eras in which these other technologies emerged, so it is possible that governance proposals that were out of reach for our ancestors will be within our grasp.

International regulation of AI must engage constructively with the shape of the technology and the industry if it is to succeed, so what follows is a brief description of each. AI models are built from three fundamental components: 1) data—the text or other media used to train the system; 2) algorithms—code that defines how data is transformed into a useful model; and 3) compute—the computational hardware on which all of these calculations take place. The fundamental resources for *deployment* are similar, though their proportion tends to be different. All three major inputs are downstream of human capital today, since it is currently humans who are doing (or guiding) the labor needed to provide these inputs.[97]

[91] Waqar H. Zaidi, '"Aviation Will Either Destroy or Save Our Civilization": Proposals for the International Control of Aviation, 1920—45', Journal of Contemporary History 46, no. 1 (1 January 2011): 150–78, https://doi.org/10.1177/0022009410375257.

[92] Thomas Bernauer, The Projected Chemical Weapons Convention: A Guide to the Negotiations in the Conference on Disarmament (New York: United Nations Institute for Disarmament Research, 1990).

[93] Waqar H. Zaidi, Technological Internationalism and World Order (Cambridge University Press, 2021).

[94] Marie Isabelle Chevrier, 'Verifying the Unverifiable: Lessons from the Biological Weapons Convention', Politics and the Life Sciences 9, no. 1 (1990): 93–105, https://doi.org/10.1017/S073093840001025X.

[95] One clear example of substitution is the U.S. renunciation of biological weapons. President Richard Nixon is recalled to have said, 'We'll never use the damn germs, so what good is biological warfare as a deterrent? If somebody uses germs on us, we'll nuke 'em.' See William Safire, 'On Language; Weapons Of Mass Destruction', The New York Times Magazine, 19 April 1998, https://www.nytimes.com/1998/04/19/magazine/on-language-weapons-of-mass-destruction.html.

[96] Zaidi, Technological Internationalism and World Order.

[97] However, over time we should expect that the provision of these inputs will become increasingly automated.

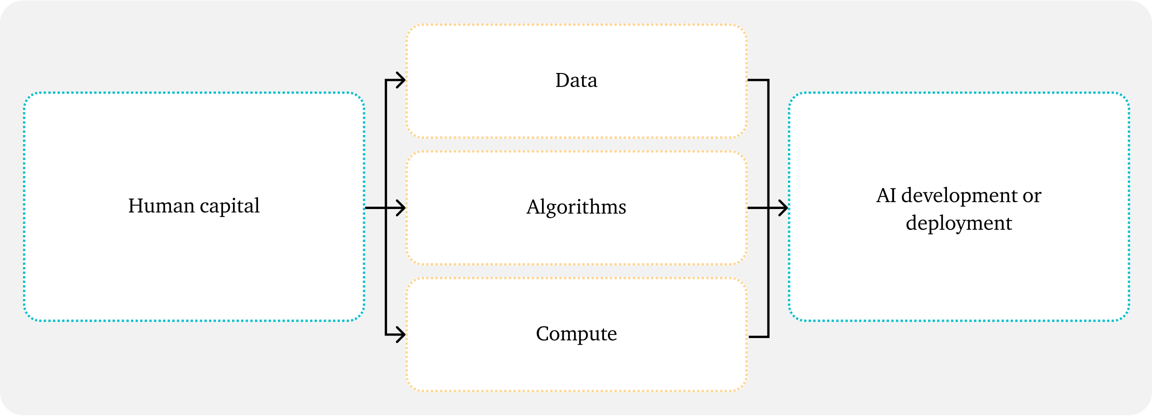

Figure 1.1: Summarized inputs into AI development or deployment. Based on a figure in Sastry et al. (2024), and modified to be inclusive of both AI development and deployment.

Of all inputs, computing hardware appears the most amenable to governance, as it enables policies regarding regulatory visibility, allocation of resources, and enforcement.[98] By comparison, it is comparably difficult to establish robust governance over personnel,[99] data,[100] or algorithms.[101]

The AI model lifecycle is often summarized as having three phases: training, fine-tuning, and inference.[102] Training (or "pretraining" in some contexts) is the initial creation of a model. Fine-tuning is the more fine-grained shaping of the model's behavior. Inference is the use of a model to do work. As machine learning paradigms evolve, the character and relative importance of these phases also evolve. For example, recent shifts toward "reasoning models" (or more generally, "inference scaling") have potentially important ramifications for the relative computational difficulty of the different phases, with inference-time compute now potentially playing a much more important role than it did previously.[103]

Regulations for AI might engage with one or more phases of the model lifecycle and one or more of the key inputs. This domain of inquiry is vast, so only a very brief summary of some relevant parts of this topic area can be provided here. Three ways to approach AI regulation are to stipulate rules about 1) how models are created, 2) how models behave, and 3) how models are controlled. First, the creation and modification of models employ a wide variety of different processes, some of which might be amenable to governance. For example, con-

[98] Girish Sastry et al., 'Computing Power and the Governance of Artificial Intelligence' (arXiv, 13 February 2024), http://arxiv.org/abs/2402.08797.

[99] Monitoring relevant people is likely infeasible and faces significant strategic problems (see Section 2.1).

[100] Controlling data is possible in theory (see Section 2.5.1), but credibly demonstrating that it has not been copied is extremely difficult (see Section 2.5.2.4). Data governance may become more workable if computing security improves drastically and if techniques such as federated learning—which allows models to be trained on private data without revealing that data—become more widely known.

[101] Algorithm governance would require personnel controls (see Section 2.1) and also brings its own particular difficulties since algorithms are potentially extremely sensitive information (see Appendix E).

[102] Changes in leading techniques are rapid, so this breakdown should be considered approximate rather than definitive.

[103] This summary does not capture the nuanced changes that this paradigm shift might create for AI development and governance. For a fuller explanation, see Toby Ord, 'Inference Scaling Reshapes AI Governance', 12 February 2025, https://www.tobyord.com/writing/inference-scaling-reshapes-ai-governance.

straints can be placed on the inputs used to create the model, such as dangerous kinds of data; the size of the model, since the largest models appear to generate many of the most salient risks; or the techniques used to train the model, since some might be regarded as very safe[104] or unsafe[105] (see also Section 4.5.1.2.2). Second, regulations might also apply to the behavior of models, including their performance on capability or safety tests that are either fully automated or involving humans (see Section 4.5.1.2.3). In this report, the term "evaluation" will generally refer to tests which can be run automatically with no human input, although the potential for human involvement will also be discussed (see Section 3.5).[106] Third, regulations might relate to how the model is deployed, with major categories including regulations for models kept within data centers (Section 4.5.2.3.1) or those that are embedded into devices which can be moved around (see Section 4.5.2.3.6).

To outline the scope of the regulatory problem, consider a scenario where regulating AI requires controls on all relevant hardware supply chains, data centers, and AI-providing organizations involved in development and deployment. A view this broad might indeed be needed, since reliable hardware governance often requires being able to verify that the hardware does not itself contain governance circumvention mechanisms.[107] While aspects of this description will be unrealistic, it provides a sense of how deep and multifaceted the challenges of AI verification can be. This summary should be seen as outlining the problem space, not proposing a solution.

Supply chains could be governed and verified at key choke points such as semiconductor fabrication to ensure that hardware matches expected designs and to ensure that required cyber and physical security measures throughout the hardware provision ecosystem are in place.[108] Ideally, hardware would be subject to verification at all high-leverage points in the supply chain, and the supply chain would use chain-of-custody requirements. Data centers could be governed and verified similarly, with all equipment installed in them subject to design review and inspection to ensure compliance with the verification scheme.[109] Most challenging of all might be the regulation of all uses of data center compute, including all workloads that companies and individuals submit to run on data center hardware.[110] Detailed

[104] See for example some of the proposals for provably safe AI. David 'davidad' Dalrymple et al., 'Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems' (arXiv, 8 July 2024), https://doi.org/10.48550/arXiv.2405.06624; David 'davidad' Dalrymple, 'Safeguarded AI: Constructing Guaranteed Safety' (Advanced Research and Invention Agency, UK Government, 2024), https://www.aria.org.uk/media/3nhijno4/aria-safeguarded-ai-programme-thesis-v1.pdf.

[105] For example, long-term planning agents may be inherently unsafe. Michael K. Cohen et al., 'Regulating Advanced Artificial Agents', Science, 5 April 2024, https://doi.org/10.1126/science.adl0625.

[106] "Audits" are also commonly mentioned as terms of art, often employing a mix of automated and human-centered tests.

[107] For more on this, see Appendix D and Section 2.2.3.3.

[108] The relevant hardware supply chains are highly international. Chris Miller, Chip War: The Fight for the World's Most Critical Technology (Simon and Schuster, 2022).

[109] See Section 2.5.2.1 for more on this.

[110] If regulation is universal, it would apply to all workloads. If regulation focuses on only AI projects involving large amounts of compute, smaller workloads can be examined more minimally. However, distributed training of AI models is workable, such as splitting a large project into a thousand smaller workloads, then any small workloads submitted to a data center operator might be components of a distributed large training run. To address this, probabilistic testing of smaller workloads, rather than universal testing, would be sufficient to catch a major circumvention attempt.

regulation of activities taking place on the AI chips would require methods to verify the software, data, algorithms, and hyperparameters that are used in the creation of AI models as well as ways to scrutinize the models themselves.[111] Further challenges include the fact that hardware stacks diverge from one another due to competition among hardware and software providers—and the size and complexity of the ecosystem is expanding dramatically. We are entering the range of millions to tens of millions of non-uniform AI-specialized data center chips deployed simultaneously around the world, with rapid evolution at all levels. Governance efforts like this are sometimes only as strong as their weakest points. Therefore, states' threat models should evolve as they make changes to the system. For example, if you shore up hardware protections, those seeking to circumvent your system might focus more on cyber attacks.[112]

This stylized description of a regulatory system illustrates why unilateral verification cannot be expected to be sufficient in domains such as regulation. While unilateral verification can certainly play a role in every agreement by providing states with information about whether the declarations by their counterparties are true and complete,[113] for at least some agreement types it cannot provide the level of certainty required for the agreement to be considered verifiable. The crux of this issue is that computation tends to be universal and fungible,[114] so unless a state has a reliable method for checking that the computations undertaken by another state abide by specified rules, they cannot be certain that the computations were compliant. Distant unilateral verification such as satellite imagery allows states to understand the rough scale of a data center, and other information sources might provide approximate information about the hardware within it.[115] However, knowing a data center's scale and hardware only tells you so much, since general-purpose hardware could be used for many different things and with many different kinds of intent.[116] Some visible *effects* of AI development can be perceived (e.g., products, weapons fielded) but most of the details of ongoing AI development are invisible by default.

Intelligence capabilities might sometimes reveal details about AI development programs.[117] However, this kind of transparency shouldn't be expected to be reliable, mutual, or politically useful. Reliability is questionable because security practices change over time. Reliable

[111] More on this in Section 4.5.1.2.2, and Section 4.5.1.2.3. See also Stephen Casper et al., 'Black-Box Access Is Insufficient for Rigorous AI Audits' (arXiv, 25 January 2024), `https://doi.org/10.48550/arXiv.2401.14446`.

[112] Furthermore, adding new hardware can introduce new vulnerabilities that are specific to that hardware or to the interfaces between different kinds of hardware.

[113] On this point, see Section 1.5.2.

[114] Universal means that any relevant hardware may undertake a particular computation. Without extra information, you can't tell which piece of hardware did the computation.

[115] This does provide a way to verify declarations about hardware within the data center, but not about the specific computations undertaken by that hardware. See also Lennart Heim, 'Limitations of Satellite Imagery Analysis for AI-Specific Data Centers', Lennart Heim (blog), 13 September 2024, `https://blog.heim.xyz/limitations-of-satellite-imagery/`.

[116] In some scenarios, even seeing the scale and type of hardware could be informative. For example, if distributed training turns out to be infeasible (see Section 1.1.1) and a state pays a large overhead to create an enormous data center. In that scenario, one can infer that there is a pressing reason why they were willing to foot the extra cost of building such infrastructure, and the most likely reason is that they are building a model of immense size. For clarity, this report assumes that distributed training is possible, so in the rest of this report, no such unambiguous hardware-based signal is expected.

[117] Intelligence capabilities include—but are not limited to—satellite data, human sources, and cyberattacks.

mutual transparency—where both actors can see what the other is doing in sufficient detail to consider those activities in compliance with an agreement—cannot be expected even among the great powers with excellent cyber capabilities. Furthermore, knowledge of violations might not be politically useful, since the disclosure dilemma means that a state may not want to reveal information from intelligence operations, lest it accidentally reveal any of their intelligence capabilities or specific sources.[118] While unilateral verification was sufficient for some parts of the Cold War nuclear era,[119] it does not appear to be sufficient on its own for many of the AI agreements discussed in this report. At the same time, unilateral verification via tools such as national technical means and intelligence agencies can be expected to provide a parallel source of information that can be cross-referenced with verifiable declarations being made as part of an agreement, and therefore can help both deter and catch attempts to circumvent the agreement (see Section 1.4). In the regulatory domains, the relative weakness of unilateral verification means that it will primarily play a supporting role to cooperative verification mechanisms, which will be required if robust verification is politically necessary.[120]

1.6 AI verification may be workable

Despite the significant challenges of verifying regulatory agreements involving AI, there are also reasons for optimism. This section will outline four reasons for optimism before outlining the remainder of the report. First, as noted earlier, AI-specialized computing hardware is much more governable than it first appears. Second, the supply of AI-specialized compute has multiple major choke points that are potentially high-leverage focal points for both international agreements and verification processes. In particular, the Dutch firm ASML is the only provider of the lithography equipment required to make cutting-edge chips, and Taiwan-based TSMC dominates the fabrication of cutting-edge chips. Overall, the supply network for cutting-edge hardware spans several countries, thus requiring supply chain governance to be international from the outset. States throughout the supply chain can take important governance action even alone, and even a small group of these states working together could powerfully shape the future of AI governance. Third, AI is drastically easier to regulate and verify than cyber weapons, despite some apparent similarities between the two domains. The most powerful and geopolitically disruptive forms of AI currently use enor-

[118] Information from state intelligence agencies may not be something that you can reveal in detail to either the other state or to an international organization, since the very act of revealing that information can also reveal how you attained it. Allison Carnegie and Austin Carson, ‘The Disclosure Dilemma: Nuclear Intelligence and International Organizations’, American Journal of Political Science 63, no. 2 (April 2019): 269–85, https://doi.org/10.1111/ajps.12426.

[119] All agreements about nuclear weapon controls up till the 1987 INF Treaty were verified unilaterally. Rose Gottemoeller, ‘Looking Back: The Intermediate-Range Nuclear Forces Treaty’ (Arms Control Today, 2007), https://www.armscontrol.org/act/2007-06/looking-back-intermediate-range-nuclear-forces-treaty.

[120] Regulation is discussed in detail in Section 4.5.

mous amounts of compute and thus require highly visible infrastructure.[121] While a cyber weapon might be created on a few hundred dollars' worth of commodity hardware in any country in the world, AI systems of even middling significance currently require at least tens to hundreds of millions of dollars to train. While verification measures for cyber weapons may be nearly impossible to achieve, AI is much more amenable to verification. Fourth, privacy-preserving verification mechanisms have been developed and appear to be applicable to AI. If appropriately implemented, these mechanisms have the potential to provide verifiability without severe transparency costs.[122]

The remainder of this report proceeds as follows. First, a set of potential verification components are reviewed to provide the reader with an overall understanding of the potential and limits of various approaches to verification (Section 2). Second, we discuss how agreements over AI—and especially regulatory agreements over AI—will face a set of political options and tradeoffs that pertain to verification (Section 3). Third, a set of possible international agreements relating to AI are outlined, along with potential implementations, verification needs, and potentially workable verification approaches (Section 4). Fourth, the report concludes with a synthesis of findings and suggests some priorities for further work (Section 5).

[121] Clarify that large-scale uses of compute is often discussed as a useful way to distinguish certain categories of potentially dangerous activity from less dangerous activity. This is a widely discussed measure because it focuses parts of the conversation on particularly dangerous activities and ensures that governance is relatively well targeted. It is not however a complete regulatory plan on its own. Sara Hooker, 'On the Limitations of Compute Thresholds as a Governance Strategy' (arXiv, 29 July 2024), https://doi.org/10.48550/arXiv.2407.05694.

[122] Some prior related works concluded that the verification of AI would be politically infeasible, but did not examine the potential of privacy-preserving technologies for addressing that problem. For example, see Jane Vaynman and Tristan A. Volpe, 'Dual Use Deception: How Technology Shapes Cooperation in International Relations', International Organization 77, no. 3 (March 2023): 599–632, https://doi.org/10.1017/S0020818323000140.

2 Survey of verification components

This section surveys some key verification components which might be brought to bear on international agreements over AI. Each of these components focuses on different objects or processes, such as personnel, digital systems (including software and hardware), electrical infrastructure, socio-technical systems, and hardware enclosures. Later sections of this report illustrate some ways in which these verification components could be employed to verify international AI agreements.

2.1 Personnel

In personnel-centered verification, human personnel play one or more key roles as the Prover seeks to make itself legible to the Verifier. The following subsections unpack three very different ways in which humans can play central roles in verification: as targets of verification controls, as intermediaries for information, and as inspectors.

2.1.1 Verifiable personnel controls

Verifiable personnel controls can be used by a Prover to demonstrate to a Verifier that personnel within the Prover's institutions are managed according to an agreement, and that no additional personnel are present. Potential controls include verifiable processes, physical access controls, digital controls, and legal controls. However, there are significant challenges with robustly verifying these due to the asymmetric burden of proof for verifying negative claims (see Section 1.5.2). A fuller exploration of this topic can be found in Appendix A.

2.1.2 Verifiable claims centered on access to personnel

Some verification approaches aim to help Provers make verifiable claims through the structured provision of access to personnel. For example, an agreement might give the Verifier the right to interview anyone in a designated group.[123] An agreement may also provide safe opportunities for those personnel to "blow the whistle" on non-compliant activities they know about within their organization. While these mechanisms can be useful in low-stakes environments, they have severe limitations in high-stakes environments, as the Prover would be able to restrict access to personnel who have witnessed non-compliant behavior or otherwise only employ loyal high-trust individuals. A longer exploration of these points can be found in Appendix B.

[123] For example, the International Atomic Energy Agency (IAEA) can employ interviews of declared nuclear personnel as part of their nuclear safeguards inspections. 'IAEA Safety Standards: Functions and Processes of the Regulatory Body for Safety' (International Atomic Energy Agency, 2018).

2.1.3 Human inspectors as a verification mechanism

Inspectors have played a key role in some of history's most challenging verification agreements, such as the Intermediate-Range Nuclear Forces (INF) treaty,[124] (Westport, Conn.: Praeger, 1998). and it is reasonable to expect that they will also play a role with AI. A related role that humans can play in a verification system is that which this report terms an *assessor*—a human that makes judgments about compliance based on information that they are provided access to. This can be similar to, or overlap with, the roles of independent auditors for other aspects of AI systems. The responsibilities, tools, and rights of inspectors tend to be listed in agreement details.[125] Later sections of this report explore aspects of the inspector's potential role. Of particular note, inspectors may play crucial roles in confirming that buildings and hardware are compliant with an agreement both at the outset and intermittently thereafter (see Section 2.5.2.1). The related role of the assessor is also explored with regards to the potential for examining information—such as code, data, and models—within tightly controlled facilities (see Section 3.5).

A key challenge is that when human inspectors visit sensitive sites, they can detect information beyond that required for their verification task. Even if garnering such information is not their goal, the danger that they would notice something security-relevant can make the transparency-security tradeoff worse. For example, this was a key concern during Cold War arms control negotiations.[126] A related policy point is that it is advisable for states to not co-locate sensitive assets, such as military hardware, with AI infrastructure if at all possible, to avoid an unnecessarily bad transparency-security tradeoff with regards to AI hardware inspections (see Appendix C.4).

However, inspectors for AI *hardware* should have a very promising transparency-security tradeoff. Using only human senses it is exceedingly unlikely that highly sensitive information would be revealed to an inspector who physically steps into a data center. What they would be able to perceive would be only the kinds of things they would be sent to examine, such as building layout, hardware inventory, and hardware connections. We should be optimistic about this possibility, because inspectors were successfully employed in much more sensitive domains, such as the INF treaty, where extensive discussions were needed to arrive at an inspection protocol that sufficiently addressed the concerns of all sides.[127] The INF treaty

[124] Joseph P. Harahan, On-Site Inspections Under The INF Treaty, A History of the On-Site Inspection Agency and Treaty Implementation, 1988-1991 (On-Site Inspection Agency, United States Department of Defense, 1993); George Rueckert, On-Site Inspection in Theory and Practice: A Primer on Modern Arms Control Regimes

[125] Consider that the verification protocol for the 1991 Strategic Arms Reduction Treaty (START) has a 500-page verification protocol. Rose Gottemoeller, 'Looking Back: The Intermediate-Range Nuclear Forces Treaty' (Arms Control Today, 2007), https://www.armscontrol.org/act/2007-06/looking-back-intermediate-range-nuclear-forces-treaty.

[126] Coe and Vaynman, 'Why Arms Control Is so Rare'.

[127] 'Memorandum of Agreement Regarding the Implementation of the Verification Provisions of the Treaty Between the United States of America and the Union of Soviet Socialist Republics on the Elimination of Their Intermediate-Range and Shorter-Range Missiles', 21 December 1989, https://nuke.fas.org/control/inf/text/inf-mouanx.htm.

case is instructive, because it illustrates that there are many ways to finely shape inspection rules so that an agreement is both tolerable to all sides and achieves its goal.[128]

The key remaining problem with inspectors is their scalability. There are already millions of AI-specialized chips in the world.[129] Physically inspecting this many chips and their many associated systems would be an enormous undertaking. Even doing this once—in detail—for an AI-specialized data center might be a substantial task. Fully inspecting all of these chips on a rapid cadence would be infeasible. Therefore, physical inspections make sense for three purposes: 1) verifying that all hardware in a data center is compliant (see Section 2.5.2.1), 2) reinspections of small portions of that hardware due to Prover-initiated maintenance or Verifier-initiated challenge inspections,[130] and 3) random inspections, presuming that some specific kinds of verification of chip activities can only be done locally.[131] All three of these activities are feasible with a reasonable number of inspectors.[132]

2.2 Digital systems

2.2.1 Cryptography

One reason that AI verification may be easier than historical arms control is that applied cryptography has greatly advanced. This allows a host of meaningful governance functions to be conducted in a way that does not reveal extraneous information.[133] In sum, cryptography can often allow the transparency-security tradeoff to be navigated successfully.

Moreover, cryptographic claims can be at least as reliable as physical claims in practice. For example, essentially the entire Internet and all technical stacks depend on cryptography that is profoundly reliable. Skepticism about the verifiability of non-material things such as digital files or computational operations is reasonable. However, we are in a world where an extraordinary proportion of all economic and social exchanges already take place through

[128] For example, nuclear weapons inspection protocols show that it is possible to vary the level of intrusiveness greatly via various techniques such as limiting access, weighing and measuring external dimensions rather than direct visual of the verified object, and shrouding. Corresponding technical measures for AI hardware would need to be devised before and during negotiation of an AI agreement that involves inspection.

[129] Agam Shah, 'Nvidia Shipped 3.76 Million Data-Center GPUs in 2023, According to Study', HPCwire, 10 June 2024, https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/.

[130] A challenge inspection is an inspection triggered by the Verifier requesting to see a particular site. This is a concept from the arms control literature. It should be noted that challenge inspections can potentially be abused if not scoped carefully, since a state might request access to a location for reasons unrelated to the agreement. Jack Allentuck, 'Challenge Inspections in Arms Control Treaties: Any Lessons for Strengthening NPT Verification?' (Brookhaven National Lab., Upton, NY (United States), 1992), https://www.osti.gov/biblio/10174104.

[131] A scheme of this kind was proposed in Shavit 2023. This scheme is not discussed extensively in this report because it depends on speculative weight snapshotting hardware features that are not yet available. However, that scheme served as the primary inspiration for much of this report, including the section on Section 2.5.3.

[132] See in particular the calculations in Shavit 2023 regarding the number of inspectors needed to cover a given set of hardware in a given amount of time, with a given chance of successfully catching a circumvention attack.

[133] A useful alternative framing of the usefulness of cryptography in making different kinds of knowledge claims is available in the 'structured transparency' literature. Andrew Trask et al., 'Beyond Privacy Trade-Offs with Structured Transparency' (arXiv, 2020), https://doi.org/10.48550/arXiv.2012.08347.

digital media where cryptography provides verifiable security.[134] Furthermore, the cryptographic concepts employed in this report are expected to be adaptable to be fully workable in a world with quantum computers.[135]

2.2.1.1 Encryption

Encryption schemes allow data communicated between intended parties to be unintelligible for any entity not permitted access. Cryptography relies on digital keys, known to parties intended to receive (and consume) the data, to decrypt the received encrypted data (known as *ciphertext*) to generate the original and intelligible form of the data (known as *plaintext*). Encryption can be symmetric—where the digital key is identical for encryption and decryption—or asymmetric—where the digital key used for encryption is different from that used for decryption. Crucial for this report is asymmetric encryption. Asymmetric encryption, also referred to as public-key cryptography, employs techniques that allow an agent (or machine) to publicly reveal a "public key" (encryption key) that can be used to encrypt data in a way that only the original agent can decrypt using their "private key" (decryption key). Exchanges of public keys enable the encryption of communications between the parties.[136] Note that for performance reasons, asymmetric encryption is often used to establish a shared session key that can be used for symmetric encryption, which is often much faster than asymmetric encryption. Such an approach eliminates the need to have a pre-shared secret (i.e., symmetric key) between any parties that could potentially communicate.

2.2.1.2 Cryptographic signatures

A cryptographic signature allows a Prover to authenticate a particular piece of data by encrypting the data or its digest (i.e., hash) using a private key known only to the Prover. Verifiers can use the public key of the Prover to verify the authenticity of the received data by decrypting the signature and relating it to the data being authenticated. Signatures discussed later in this report will include signatures generated by machines, each of which would have their own private key. Note that this is another use case for public key cryptography.

2.2.1.3 Cryptographic commitments

A cryptographic commitment demonstrates the existence of specific data *without revealing that data*. The commitment is a digital string generated using a specialized algorithm run on the private data. Algorithms commonly employed for this purpose are one-way

[134] Consider the fact that encryption schemes such as transport layer security underpin the entire internet, operating so well that they are invisible to the overwhelming majority of the population. There are astronomical returns to any attack that can break these, but they have proved robust enough to provide secure infrastructure for the increasingly digitized global economy.

[135] Note that individual cryptographic algorithms will likely need to change, and be replaced with similar algorithms that are being designed in the field of post-quantum cryptography. Thankfully, this transition is occurring more broadly in any case, and while higher computational costs of post-quantum algorithms is concerning for this transition generally, the change is likely to have minor performance costs compared to the scale of the AI systems in question.

[136] This description is conceptual and is not intended to accurately describe all schemes of this kind. Furthermore a brief description of how quantum computers can change this picture can be found in Appendix D.

hashing functions, which can be used to convert any data into a cryptographic string—the "commitment"—which is extremely difficult to falsify. Furthermore, the commitment is a relatively short string of characters, thus typically making it much smaller than the data that it is committing to.

Commitments of this kind are very useful because they allow a Prover to provide credible and future-verifiable information about their private data. At a later time, the Prover can reveal the private data associated with their commitment, thus allowing the Verifier to confirm that the hash of the revealed data is the same as the commitment provided earlier. This is often combined with partially revealed information, so that a Prover provides information about many pieces of data, and the Verifier can then pick some very small fraction which is revealed and verified. Thus, at the cost of a very small revelation of private data, a far larger set of data can be reliably verified. Commitments are credible to the extent that the hash function employed is difficult to spoof. To achieve extreme levels of credibility with a cryptographic commitment, several hashing algorithms could be employed in parallel to provide several different discrete commitments—each of which could be separately tested against the revealed data.[137]

An application of cryptographic commitments discussed in this report is that of a *model fingerprint*: a way to recognize a model once it has been created. An ideal fingerprint would attest to the specific model without revealing any of its data. Therefore, a cryptographic commitment—or several of them in parallel—could be employed for this purpose.[138]

2.2.2 AI-specialized computational hardware

General-purpose computational devices such as Central Processing Units (CPUs) are nearly ubiquitous. However, the ongoing explosion of AI capabilities is being enabled by a slightly more specialized kind of hardware. Originally employing Graphics Processing Units (GPUs), the current AI paradigm has evolved toward increasingly specialized hardware that is a better fit for the workloads that modern AI demands.[139] It is primarily this specialized hardware that is referred to as "compute" or "chips" in this report and elsewhere, since the performance differences between the specialized hardware and more general-purpose hardware are dramatic. It is possible to use commodity GPUs for some AI work, but they are not as capable as AI-specialized compute (see Section 3.2). The subsections below expand on two verification techniques that are specific to AI-specialized compute.

[137] This would be limited to only modern well-designed hashing algorithms, since some older algorithms have security problems which make it theoretically possible for the Verifier to learn things from the Prover's commitments.

[138] A related but very different problem is devising a model fingerprint that cannot be adversarially manipulated to show that two models are different when they are in fact nearly identical. This would be desirable in some governance contexts in which it is important to be able to recognize whether models are substantially different from one another. See for example Sally Zhu et al., 'Independence Tests for Language Models' (arXiv, 17 February 2025), https://doi.org/10.48550/arXiv.2502.12292.

[139] AI-specialized hardware can provide better computational performance for a given amount of energy.

2.2.2.1 Chip registry

A *chip registry* is designed to contain unique identifiers for AI-specialized chips, along with other data such as who currently owns them.[140,141] Such a registry can facilitate knowledge of which countries and corporations control which chips.

A chip registry requires that a robust unique identifier be available for each chip. For some chips, a sufficiently protected hardware-level private key might be sufficient.[142] Note that there are serious ongoing questions about the security of on-chip keys of various kinds, such as those in new NVIDIA GPUs, traditional security modules, and potentially more tamper-resistant techniques such as physical unclonable functions.[143] Chip identifiers are a fundamental requirement of many hardware governance schemes, since without them chips tend to be fungible.[144] If the registry is intended to employ real-time digital updates via interactions with the chip over the Internet, a private key for the chip is most likely required in order to ensure that the digital interactions cannot be easily spoofed. If a chip registry is considered politically workable with only rare updates based on physical inspections, then surprisingly low-tech solutions may also be workable, including the use of adhesive glitter and photographs.[145]

Location information could also be included in a registry. For example, the registry might aid hardware providers and governments in implementing country-specific regulations. A crucial issue with potential mechanisms that provide the location of chips with high accuracy (see Section 2.2.4.6 below) is that they can introduce severe transparency-security tradeoffs for sensitive organizations such as militaries (see Section 1.5.1.1). Less accurate location mechanisms can potentially help with this.[146] Furthermore, more complex schemes of hardware governance might use a "logical" location rather than physical location as part of governance rules, where a known institution (such as a state) takes public responsibility for a chip and

[140] Yonadav Shavit describes a similar mechanism, which he terms a 'chip owner directory'. Yonadav Shavit, 'What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring' (arXiv, 20 March 2023), https://doi.org/10.48550/arXiv.2303.11341.

[141] Deric Cheng, 'Evaluating An AI Chip Registration Policy' (Convergence Analysis, April 2024).

[142] NVIDIA H100s for example have a hardware-backed private key. Emily Apsey et al., 'Confidential Computing on NVIDIA H100 GPUs for Secure and Trustworthy AI', NVIDIA Technical Blog, 3 August 2023, https://developer.nvidia.com/blog/confidential-computing-on-h100-gpus-for-secure-and-trustworthy-ai/.

[143] Pim Tuyls and Boris Škorić, 'Strong Authentication with Physical Unclonable Functions', in Security, Privacy, and Trust in Modern Data Management, ed. Milan Petković and Willem Jonker (Berlin, Heidelberg: Springer, 2007), 133–48, https://doi.org/10.1007/978-3-540-69861-6_10; Wenjie Che, Fareena Saqib, and Jim Plusquellic, 'PUF-Based Authentication', in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, ICCAD '15 (Austin, TX, USA: IEEE Press, 2015), 337–44; Maria Sommerhalder, 'Hardware Security Module', in Trends in Data Protection and Encryption Technologies, ed. Valentin Mulder et al. (Cham: Springer Nature Switzerland, 2023), 83–87, https://doi.org/10.1007/978-3-031-33386-6_16; Aarne, Fist, and Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'; Md Nazmul Islam and Sandip Kundu, 'Enabling IC Traceability via Blockchain Pegged to Embedded PUF', ACM Transactions on Design Automation of Electronic Systems 24, no. 3 (31 May 2019): 1–23, https://doi.org/10.1145/3315669.

[144] One existing implementation is the Device Identifier Composition Engine (DICE) standard. This scheme depends on a Unique Device Secret (UDS) often provisioned after manufacturing as a unique value for each physical chip.

[145] This is discussed in Aarne, Fist, and Withers (2024).

[146] Asher Brass and Onni Aarne, 'Location Verification for AI Chips' (Institute for AI Policy and Strategy, April 2024), https://www.iaps.ai/research/location-verification-for-ai-chips.

can verify claims about that chip via various mechanisms, including challenge inspections.[147] This concept is explored in Section 2.5.4.2 with respect to military chips. The general idea could also be applied more broadly through technically simple (although politically challenging) mechanisms such as having all states take responsibility for all chips within their borders and continuously demonstrate that fact to their peers.[148]

2.2.2.2 Chip supply chain verification

Chip supply chain verification requires that the supply networks for AI-specialized chips can be verifiably monitored by the parties to the agreement. This could include tracking specialized equipment and materials for producing high-end chips, such as extreme ultraviolet lithography machines, in order to identify all relevant chip fabrication facilities. Verification centered on the chip supply network is important for many of the agreements described in this report. However, it is also one of the most heavily examined adjacent areas, so in the interests of space, this report will not deeply explore the structure of the chip supply network and the prospects for its governance. The interested reader is encouraged to look at other publications to understand this space.[149]

2.2.3 Hardware verifiability

Generally speaking, hardware can be quite verifiable because its mechanisms can be transparently understood via inspection and monitoring. For this reason, the verifiability of hardware undergirds much of the remainder of this report. However, four major challenges of hardware-based verification are worth spelling out in detail: 1) new hardware often needs to be developed to verification purposes, 2) miniaturization makes some forms of verification more difficult, 3) advanced semiconductors are very difficult to verify, and 4) hardware that is built to be verifiable via downstream processes might incur significant performance penalties.

2.2.3.1 New hardware is often needed for verification

While existing hardware is often relatively transparent if inspected and monitored in detail, it is often not in a configuration that allows the Prover to protect their own security if that hardware were subjected to close inspection by the Verifier. The INF Treaty was ultimately successful due to the creation of a hardware verification mechanism that allowed the United States to know that Soviet missiles exiting a particular facility were in compliance with the

[147] See Section 2.1.3 for more on challenge inspections.

[148] This is a form of light-touch verifiable regulation that is not discussed further in this report, and would be a compelling possibility for future work.

[149] Introductions to the topic can be found in these resources: Girish Sastry et al., 'Computing Power and the Governance of Artificial Intelligence' (arXiv, 13 February 2024), http://arxiv.org/abs/2402.08797; Center for Security and Emerging Technology, Saif Khan, and Alexander Mann, 'AI Chips: What They Are and Why They Matter' (Center for Security and Emerging Technology, April 2020), https://doi.org/10.51593/20190014; Akhil Thadani and Gregory C. Allen, 'Mapping the Semiconductor Supply Chain: The Critical Role of the Indo-Pacific Region' (Center for Strategic and International Studies, 30 May 2023), https://www.csis.org/analysis/mapping-semiconductor-supply-chain-critical-role-indo-pacific-region; Chris Miller, Chip War: The Fight for the World's Most Critical Technology (Simon and Schuster, 2022).

treaty.[150] The development of these verification mechanisms started years before the treaty was signed. In part, this was due to the need to find security-preserving verification mechanisms and implement them in hardware.

For digital hardware, the story may be similar. The following sections discuss several kinds of hardware-enabled mechanisms which fall along a broad spectrum, from mature, widely implemented technologies to speculative technologies lacking even a prototype. While this report emphasizes relatively mature technological capabilities, it is certainly not guaranteed that an international agreement will be able to be verified using off-the-shelf hardware for its crucial components. Components that play a central role in the verification process may need to be designed and built for that specific purpose. Moreover, to guard against the insertion of verification circumventions or other vulnerabilities, the new hardware may have to be collaboratively designed and built—a much more involved process. This latter point will be discussed again below with regard to "leading node" semiconductors (see Section 2.2.3.3).

2.2.3.2 Miniaturization can make verification more difficult

The ongoing miniaturization of digital hardware can make verification more difficult because it makes it harder to verify negative claims about digital computations or transmissions—smaller objects are easier to hide. As noted in Section 1.5.2, verifying a claim of the form "no X is occurring" often requires an exhaustive search for ways that X could be occurring. One example that will be discussed later is that of a Prover attempting to demonstrate that a data center only has specific kinds of monitored connections with the outside world (see Section 2.5.2.4). To make such a claim in today's world, the Prover would not only need to demonstrate that the major network connections (such as internet backbone cables) are adequately monitored, they would also need to demonstrate that no other hardware within the data center was capable of transmitting signals of any kind out of the facility. To make this claim robust, the Prover would have to adequately convince the Verifier that all hardware was compliant and that there was no feasible way for additional hardware to be inserted without the Verifier noticing. This problem is exacerbated by the miniaturization of both telecommunications hardware itself as well as devices through which it could be infiltrated into facilities, such as small drones.

2.2.3.3 Advanced semiconductors are difficult to verify unless cooperatively built

Verification of advanced semiconductors faces two major problems. First, leading semiconductor design and fabrication processes involve highly sensitive corporate and state secrets that will be guarded carefully. Second, downstream verification of a completed semiconductor does not appear to be technically feasible for semiconductors beyond a given level of miniaturization and complexity.

[150] Toivanen, 'The Significance of Strategic Foresight in Verification Technologies'.

Conducting mutual verification of the creation of advanced semiconductors may not be politically feasible. The leading foundries (such as TSMC) and the leading chip designers (such as NVIDIA) retain their lead over their competitors in part due to trade secrets which could be revealed if mutual verification of chip production were undertaken, and the threat of this revelation would be considered a major economic and even military security risk by the home states of those corporations. Unless we can develop a method for mutual verification of leading-node semiconductor production that does not reveal these secrets, such proposals can expect intense pushback from the firms and states involved. However, if a security-preserving mutual verification protocol could be developed for leading node semiconductor design and fabrication, such a protocol could enable highly efficient verification mechanisms to be installed in maximally performant hardware, thus addressing the concern outlined below that verifiable hardware may be less performant.

Given access to a leading node semiconductor, even a state with extensive resources is likely to have no way of non-destructively proving that the chip is compliant. No non-destructive technique is known for imaging such a chip's tens of billions of transistors spread across dozens of layers—and destructive techniques cannot be applied to chips that a state would like to use.[151] If you destroy chips in the process of learning about them, you are unable to use the chips you learn about. Furthermore, there is no known way to prove that two chips are identical, so even if a state tests dozens of chips, there is no guarantee that hardware circumventions or backdoors have not been installed in other chips.[152] These problems only apply to semiconductors above a certain level of complexity, since it is possible to reliably image single-layer semiconductors built at much larger node sizes.[153] A useful area for future work would be to examine the practical limits of downstream semiconductor verification.

2.2.3.4 Downstream-verifiable semiconductors might be less performant

Semiconductors which can be directly inspected for compliance may be less performant than their unverifiable kin. As discussed above, semiconductors that can be directly verified may need to be built with drastically less complexity and fewer layers than present-day leading node semiconductors. Each new semiconductor fabrication node has made semiconductors more complex but yielded substantial performance benefits over prior nodes, so it stands to reason that this complexity limit on downstream verifiability means that verifi-

[151] One destructive technique is to use scanning microscopes to image an entire layer, then remove that entire layer, and repeat. The resulting map could in theory allow the chip's entire design to be inspected, although the feasibility of these steps is unclear.

[152] Depending on the threat model, this may or may not be an important point. Random inspection might place significant constraints on the potential for hardware circumvention in scenarios where a broad array of hardware needs to be physically modified for a circumvention to be successful (see Shavit 2023). However, if the threat model indicates that even a single piece of modified hardware could allow circumvention for the larger system, then it is infeasible to guard against such a danger with downstream inspections. An example of a threat model like this would be something that allows a third party to use a single piece of hardware to get inside of a trust boundary (e.g., a pod of AI chips) and thus access large amounts of sensitive plaintext data.

[153] Ken Shirriff, 'Standard Cells: Looking at Individual Gates in the Pentium Processor', July 2024, http://www.righto.com/2024/07/pentium-standard-cells.html.

able semiconductors will be less performant than their leading node kin. This performance penalty is likely to be the most acute in those domains where the greatest performance improvements have been made in recent years, such as AI-specialized compute chips and their network interconnect hardware. Proposals for using semiconductors as key parts of a verification mechanism must therefore either address the challenge of cooperative production or find a way to avoid requiring leading node semiconductors within the trust boundary of the verification system.[154]

2.2.4 Hardware-enabled mechanisms

2.2.4.1 Anti-tamper mechanisms

Hardware can be produced, packaged, and monitored in ways that makes tampering with it either difficult (i.e., "tamper-resistant") or infeasible to hide (i.e., "tamper-evident").[155] Tamper resistance increases the complexity, cost, and time of efforts to modify hardware. Tamper evidence makes it possible for the Verifier to catch the Prover's efforts to modify hardware.

While simple in theory, tamper resistance and tamper evidence are very difficult to achieve if you assume that a state actor will be able and willing to spend significant time and resources attempting to break your tamper resistance mechanisms. A broad exploration of tamper resistance mechanisms is outside the scope of this report.[156] Two categories of anti-tamper mechanisms will be discussed further in this report:

1. Hardware packaging such as enclosures can employ anti-tamper mechanisms such as physical unclonable functions (see Section 2.5).
2. Hardware installations can be inspected and then monitored using video cameras and electromagnetic or acoustic sensors. However, these data streams can bring their own potential vulnerabilities, so they need to be considered carefully (see Section 2.5.2.2).

2.2.4.2 Remote attestation

Hardware mechanisms and cryptographic techniques can be employed to prove that the code running on a system—including both firmware and software—has not been tampered

[154] For example, the flexHEG report describes a way to move the AI chip outside of the trust boundary, thus enabling the verification of hardware which was built for cutting-edge performance. Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees'.

[155] "Tamper proof" is a related term that refers to theoretical hardware that cannot be modified in a non-destructive way. See also Aarne, Fist, and Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'; Kulp et al., 'Hardware-Enabled Governance Mechanisms'.

[156] For insightful introductions to these challenges and how they relate to AI-specialized chips in particular, see Aarne et al. (2024) and Kulp et al. (2024).

with.[157] The "remote" in remote attestation means that this process can in fact be conducted remotely, without the Verifiers even knowing where the system is located. The crux of this mechanism is the hardware root of trust—a private key—embedded in the chip or system of interest. Presuming that the hardware root of trust has not been violated, remote attestation is poised to provide robust results.[158] Remote attestation undergirds many of the hardware-enabled techniques discussed later in this report, and thus they share its dependence on hardware integrity.

2.2.4.3 Liveness pings

To demonstrate that a particular piece of hardware is operating with its original root of trust unchanged, the Prover can arrange to have that hardware respond to cryptographic pings that are sent by the Verifier. In so doing, the Prover can substantially limit their own ability to manipulate the hardware they are using for verification. For example, if all key parts of their hardware are responding to cryptographic pings every few seconds, they will be limited in their ability to modify any of that hardware without being noticed. This mechanism requires that the roots of trust for the hardware devices being pinged have not been copied by the Prover. This requirement could potentially be supported by mutual verification of the supply chain (see Section 2.2.3.4) and installation in the presence of inspectors.

2.2.4.4 Confidential computing

Confidential computing is a hardware-enabled computational approach that provides credible assurances to the remote user that no one else can see their code, data, or results, including the hardware operator.[159] It is premised on remote attestation, as described above, and thus depends on a secure hardware root of trust (see Section 2.2.4.2). This technology is already available for new generations of leading AI hardware, including NVIDIA's Hopper and Blackwell microarchitectures.[160] Note that while existing chips can accomplish confidential computing, upcoming generations such as NVIDIA's upcoming Blackwell chips are advertised as being able to do confidential computing in large clusters of GPUs with performance similar to unencrypted computing.[161]

[157] Onni Aarne, Tim Fist, and Caleb Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing' (Center for a New American Security, 2024); Gabriel Kulp et al., 'Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090' (RAND Corporation, 18 January 2024), https://www.rand.org/pubs/working_papers/WRA3056-1.html.

[158] See also Appendix D.

[159] The term "confidential computing" is used in this report to refer to not only the existing confidential computing standard, but also the entire family of technologies that can similarly enable credible multi-agent remote attestation. A full exploration of this family of technologies and techniques is beyond the scope of this report.

[160] Emily Apsey et al., 'Confidential Computing on NVIDIA H100 GPUs for Secure and Trustworthy AI', NVIDIA Technical Blog, 3 August 2023, https://developer.nvidia.com/blog/confidential-computing-on-h100-gpus-for-secure-and-trustworthy-ai/; 'NVIDIA Blackwell Architecture', NVIDIA, accessed 13 March 2025, https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/.

[161] The details of computational cost for confidential computing on different GPU generations is important, but will not be explored in great depth in this report. This is an area where further work would be very valuable.

Confidential computing enables a number of very useful computational abilities, only a few of which will be mentioned here. First, it allows the Prover and Verifier to both be shown code that is submitted to be run within the secure environment—thus allowing mutual code review. Second, it allows each actor to hide sensitive parts of its code and data (such as algorithms, see Appendix E) which it does not want to show to the other party while also enabling tests or evaluations to be run against that hidden code and data. Third, falsifications of any component of the system are difficult, since cryptographic commitments are used to demonstrate that the components are unchanged from their attested state—thus allowing each side to know that their code is running against the right objects even if they cannot see the plaintext versions of those objects. Overall, this kind of cryptographically-attested code and data approach therefore allows Provers and Verifiers to work together to enable security-preserving evaluation of any digital object.[162] More exploration of this concept can be found in Section 3.5.

This standard already exists, as do stacks that implement it. However, it is certainly not guaranteed that this approach will be robust enough for the political needs of agreements between states. It is certainly possible that the confidential computing stack has flaws that a cyber-competent state would be able to find and exploit to either exfiltrate data that should be hidden or damage the integrity of the verification processes. More work is needed to examine the capability and limits of confidential computing as well as to propose complementary hardware monitoring techniques that would make hardware-centered attacks much more difficult (see Section 2.5.2).[163,164]

2.2.4.5 Licensing

Hardware licensing schemes typically require that hardware only operates to its full potential if it is provided with an appropriate (encrypted and signed) license.[165] An appropriate license allows the hardware to operate at full capacity for some amount of time. Such schemes have precedent in the computing world, including in some processors provided by Intel.[166]

[162] For a practical example of the application of this technology to a cooperative verification problem in AI, see 'Secure Enclaves for AI Evaluation' (OpenMined, 12 December 2024), https://openmined.org/blog/secure-enclaves-for-ai-evaluation/.

[163] Note in particular that the implementation on NVIDIA's Hopper chips is not intended to be robust against sophisticated physical attacks. Rob Nertney, 'Confidential Compute on NVIDIA Hopper H100' (NVIDIA, 25 July 2023), https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/HCC-Whitepaper-v1.0.pdf.

[164] The limited ability of the underlying hardware (trusted execution environments and security modules) to defend against sophisticated attacks is also a key theme in Aidan O'Gara et al., 'Hardware-Enabled Mechanisms for Verifying Responsible AI Development' (arXiv, 2 April 2025), https://doi.org/10.48550/arXiv.2505.03742.

[165] See Kulp et al., 'Hardware-Enabled Governance Mechanisms'; Aarne, Fist, and Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'.

[166] 'Intel On Demand', Intel, accessed 13 March 2025, https://www.intel.com/content/www/us/en/products/docs/ondemand/overview.html.

In some verification approaches, licensing is an *enforcement* mechanism which makes some kinds of verification easier.[167] However, there is no requirement that the license-providing system be controlled by the Verifier. It can just as easily be controlled by the Prover in a way that is transparent to the Verifier, thus allowing similar verification processes to happen with somewhat different political implications (see Section 3.6).

Note that this scheme could be applied to chips via on-chip mechanisms or to larger enclosures of hardware such as pods (see Section 2.5.4.1). Depending on the model being built or operated, either approach could be reasonable. Generally, however, the most important models which require governance will be models that require more than one chip for training or inference, thus making pods the more appropriate enclosure size for this mechanism.

Two kinds of licensing systems have been discussed in the literature: offline and online. Offline licenses do not presume that a (reliable) digital connection is established between the licensed hardware and the license-providing institution.[168] Schemes like this allow chips to be kept fully "offline", in locations such as secure air-gapped facilities. By contrast, online licenses require a reliable digital connection between the hardware and the license-providing institution, thus allowing license requests and renewals to be exchanged much more quickly and often.

License renewals could be tied to a cryptographic exchange with the licensed hardware that requires a cryptographic challenge, such as a string that must be encrypted by the licensed hardware using its private key and returned to the license-providing institution. This would make it infeasible for license renewals to be requested early—a concern if license stockpiling would enable non-compliant use of the hardware. Furthermore, a late request for a license renewal could immediately trigger a request for physical inspection or other kinds of closer monitoring, since a late request could be an indication that the hardware was being tampered with or had been used in a way that violated the agreement. Similarly to the discussion above on liveness pings (Section 2.2.4.3), if license renewals happen on a quick cadence, they might make hardware tampering exceedingly difficult. An important caveat for both rapid licensing and liveness pings is that AI chips often have maintenance issues or die entirely, so loss of connection to a chip is certainly not reliable evidence of tampering—other systems of information must be layered with these mechanisms if they are to serve their purpose.[169]

[167] More complicated licensing mechanisms could also allow for the license itself to indicate how well the hardware should perform. While very hypothetical, such mechanisms could allow for fine-grained and rapid tit-for-tat between veto holders, since they could set each other's maximum hardware percentage at any level, thus allowing small increments or decrements as required by the political bargaining process. As with all agreements discussed in this report, the danger of a full exit from the agreement is assumed to always be in the background of any negotiations. Section 1.4 says more about the maintenance of political equilibria.

[168] Kulp et al., 'Hardware-Enabled Governance Mechanisms'; James Petrie, 'Near-Term Enforcement of AI Chip Export Controls Using A Firmware-Based Design for Offline Licensing' (arXiv, 28 May 2024), https://doi.org/10.48550/arXiv.2404.18308.

[169] As noted later, parallel streams of data could cover different crucial systems, such as networking and electric power. (See Section 2.5.2.2.)

2.2.4.6 Location verification

Hardware-enabled mechanisms can allow objects to be located with a configurable degree of accuracy, thus allowing governance processes to happen conditional on location. One example explored previously is an on-chip location verification mechanism, which is embedded in a chip and which allows the chip to rapidly respond to an encrypted digital challenge sent by the Verifier. Due to the fundamental limitation of the speed of light, the (very short) time required for the full exchange of the digital challenge allows the location of the chip to be approximated since it is known that the signal must travel slower than the speed of light.[170] More generally, location verification mechanisms can be added to any piece of hardware, not just chips.[171]

2.2.4.7 Enforced encryption of outbound data

Hardware-backed mechanisms can allow the Prover and Verifier to know with confidence that certain kinds of outbound data (e.g., model weights) are encrypted according to specific keys—thus protecting those resources from direct extraction. Since encrypted data cannot be read by anyone unless they hold the decryption key, such a mechanism can provide a cryptographic barrier that changes the shape of the governance and enforcement problems.[172] Later sections of this report explore how enforced encryption can enable key verification functionalities or make them more tractable (see for example Section 2.5.2.4).

A scheme of this kind requires at least two ingredients: one or more verifiable keys that will be used for encryption and a way to recognize the relevant outbound data operations. Keys could be provided by different parties depending on the governance goals, or by multiple parties to enable robust cooperative verification as explored below (see Section 2.2.4.7.1). Hardware-enabled mechanisms would need to provide a way for the various parties to confirm that the correct encryption keys will indeed be used.[173] Relevant outbound data operations also need to be identified correctly, since some outbound operations should certainly not be encrypted in this scheme (e.g., inference responses). Complicating matters is the fact that training and inference for large AI models require many chips—and perhaps even many pods of chips—to work together. Most AI-specialized hardware does not support

[170] Aarne, Fist, and Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'; Asher Brass and Onni Aarne, 'Location Verification for AI Chips' (Institute for AI Policy and Strategy, April 2024), https://www.iaps.ai/research/location-verification-for-ai-chips.

[171] At least limited access to the Internet is required, since the cryptographic challenge and its response is presumed to be over digital networks. Using standard hardware and software techniques, such access can be restricted to precisely match the requirements of the location verification, and thus this access would not be expected to provide any new opportunities for digital connections or attacks.

[172] An analogy to verified encryption would be that costly processes—such as having armed men travel between sites carrying a briefcase full of sensitive data—can be replaced with cheap and scalable cryptographic processes that accomplish the same governance goal.

[173] In parallel, additional hardware-enabled mechanisms might be put in place to prove that data was encrypted using a specific key. This hypothetical parallel approach could be an interesting area for further work.

encryption over the highest-bandwidth interfaces, although this is changing with NVIDIA's upcoming releases.[174]

At least one existing method can implement enforced encryption of selected outbound data, and one proposed method is promising. Existing confidential computing techniques (see Section 2.2.4.4) could allow the Prover and Verifier to mutually verify that the code operating on the hardware enforces encryption using the appropriate keys on the appropriate operations. A more speculative method is that of Flexible Hardware-Enabled Guarantee (flexHEG) systems, which would embed rules into hardware and into code that would run on a secure processor adjacent to the AI-specialized hardware.[175] In either case, operations involving larger models would likely require (for efficiency purposes) that this mechanism be implemented at the level of the pod rather than merely at the level of the chip (see Section 2.5.4.1).

Selective encryption of outbound data has three main challenges. First, this mechanism is only as robust as the technical mechanisms it depends on—such as confidential computing and flexHEG discussed above.[176] Second, encryption adds some overhead to data transfer operations, and in certain circumstances this overhead might be substantial. Third, this mechanism cannot protect data from side or covert channel attacks that exfiltrate the data in unexpected ways, such as model extraction attacks via inference calls.[177]

2.2.4.7.1 Doubly-encrypted outbound data

In a special variant of the above mechanism, selected data could be verifiably locked by *both* the Verifier and the Prover until after verification processes have completed. The encryption of outbound data would be enforced with two or more keys, with at least one coming from the Verifier and one from the Prover. Since each party knows that the data was encrypted using a key that they provided, they gain some assurance that no one (including the other party) can access the plaintext. Such a scheme would allow politically useful information protocols: for example, it can ensure that data cannot be examined by any party until it arrives in a neutral verification data center (see Section 3.6.1). Following privacy-preserving verification processes in that data center, the Verifier could provide the Prover with the decryption key for that data, thus enabling the Prover to decrypt their own copy—while still ensuring that the Verifier never gains full access to the plaintext data.[178] In sum, this mechanism would allow the Prover to demonstrably submit their digital objects to governance and verification in a way that prevents the Prover from copying or using those digital objects until after the governance processes have completed.

[174] 'NVIDIA Blackwell Architecture', NVIDIA, accessed 13 March 2025, https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/.

[175] James Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024.

[176] Off-chip encryption hardware might be able to solve this problem, but it is unclear how relevant data would be identified. Perhaps *all* outbound data could be encrypted, thus raising the credibility of this mechanism, but potentially raising its efficiency costs substantially in some domains.

[177] Jiacheng Liang et al., 'Model Extraction Attacks Revisited' (arXiv, 8 December 2023), https://doi.org/10.48550/arXiv.2312.05386.

[178] To ensure that the data can never be decrypted again in the wild, the Prover can also destroy all copies of their private key once they have secured access to their data.

This scheme may be politically valuable. First, it demonstrates that even the Prover cannot access their data without the compliance of the Verifier. This places the verification process within the AI development process and thus helps clarify the Prover's seriousness about demonstrating their compliance. Second, the Verifier could delay revelation of the decryption key following verification processes, for example if there were inconsistencies in the results. This places some pressure on the Prover to continue to engage seriously with the Verifier about resolving the information problems, since they would normally like to gain access to the data that they have paid to create, such as a finished model (see Section 3.5 for more about how repeated and escalating efforts can be made to resolve verification issues). If this process were taking place with a very large model, the costs in hardware time and money that are embodied in the model would be substantial, thus incentivizing the Prover to move effectively to demonstrate that the model is compliant so that they can quickly gain access to their valuable model.[179]

2.2.4.8 Networking hardware

Networking hardware enables machines to communicate with one another. Verification of networking hardware can enable the Prover to demonstrate how information can flow throughout their infrastructure. Furthermore, if networking hardware can be reworked to enable a combination of data retention (for the Prover) and cryptographic commitments (for the Verifier), it can allow for the verification of all data that passes through it. This concept is expanded in Section 2.5.2.3.

2.2.5 AI-enabled devices

An AI-enabled device is a physical device that has an AI model embedded into it. In most of the discussion that follows, these devices will be presumed to be mobile (i.e., not limited to data center locations). Furthermore, later sections that discuss their governance will assume that these devices are operating with highly sensitive technologies and in sensitive environments, with autonomous weapons being the archetypal example (see Section 4.5.2.3.6). All of these assumptions are intended to bias the discussion in the direction of taking seriously the most challenging verification problems for devices of this kind. As with other such conservative assumptions in this report, we hope that if we consider maximally challenging verification problems seriously, this will yield insights that can be readily applied to less challenging domains.

2.3 Electrical infrastructure

Building or running AI requires computational hardware, which in turn requires electric power. Prior work has emphasized that even very coarse-grained information about electric

[179] In the worst case, such an interaction could lead to the end of the agreement between the Verifier and Prover. In this case, the Prover would presumably never gain access to the resource at the center of the disagreement. They would have to recreate it again if they wanted it.

power usage could be useful for verification of international AI agreements.[180] While macroscopic power signatures reveal very little about the computations that are being undertaken, they can certainly help reveal undisclosed data centers.[181] That is, if an international agreement over AI requires states to disclose via declarations the locations of some or all of their data centers, information about the state's electrical power infrastructure could be used to check for omissions in those declarations (see Section 1.5.2). Since electrical power infrastructure is itself strategically valuable, revealing detailed information about its structure would be subject to the transparency-security tradeoff. Addressing these concerns is theoretically possible through the inventive use of a stack similar to that summarized in Section 3.5, where sensitive information from both the Prover (declarations about electric grid structure) and Verifier (estimates of electric grid structure from unilateral monitoring) can be combined securely to test verification-related claims. A major challenge with this sort of information exchange is that the Verifier could potentially garner more information than intended from the verification-related claims, if they structure the data they provide such that the responses from the verification process reveal important security-relevant information that the Prover wants to keep hidden.[182] Similarly, it is plausible that the Prover could learn things about what the Verifier believes to be true and thus discover important security-relevant details about the Verifier's ability to unilaterally garner information. In sum, this technique appears to have significant potential, but it may also have significant remaining challenges that deserve further study.

2.4 Socio-technical systems

Verification schemes are inevitably embedded in larger socio-technical systems. This means that verification needs both to be robust to changes in the broader systems, and to serve a role in ensuring that the guarantees apply despite the surrounding socio-technical architecture. A verification regime that relies on operators needs to be robust not just to technical circumvention, but to operators intentionally changing the systems. For this reason, even narrow technical schemes need to engage with socio-technical infrastructure.

2.4.1 Institutional digital infrastructure

Verification schemes may relate to or include broader digital infrastructure (beyond data centers) such as telecommunications systems and information technology equipment (including personal computing equipment used by key personnel). Institutional digital infrastructure could include privileged internal information systems such as internal communication, in-

[180] Akash R. Wasil et al., 'Verification Methods for International AI Agreements', arXiv.org, 28 August 2024, https://arxiv.org/abs/2408.16074v1.

[181] Even with fine-grained power usage data drawn from the equipment within a data center, it is not yet possible to reliably know what kinds of workloads are being run—even in the absence of Prover efforts to obfuscate their actions. For more on the state of the art of workload classification, see Lennart Heim et al., 'Governing Through The Cloud: The Intermediary Role Of Compute Providers In AI Regulation' (Oxford Martin AI Governance Initiative, March 2024).

[182] For example, a poorly designed exchange could reveal substantial information about the declared structure of the power grid to the Verifier (such as the locations of all major power loads).

cident reporting, and potentially other aspects of a regulatory apparatus. Verification might engage with this infrastructure in two major ways: 1) by demonstrating that controls on the infrastructure are credible and 2) by demonstrating other claims via information that the infrastructure can provide. Each will be explored in turn.

2.4.1.1 Verifiable digital infrastructure controls

In some cases, the Prover would like to demonstrate that their institutional digital infrastructure abides by specific controls. This is somewhat easier than the similar personnel-based approach (see Section 2.1.1) because the function and rules of each component of digital infrastructure can in theory be made clear to the Verifier, and adherence with the rules can be double-checked. For example, a Prover can declare that one of their general-purpose computational systems will only ever act in a strict and particular pattern (e.g., sending messages at certain intervals, encrypted with a specific key, with specific kinds of content). A Verifier can install relatively simple hardware systems to monitor the general-purpose system to ensure that it adheres to these limitations. Importantly, the Verifier-installed systems can be mutually verifiable, thus reassuring the Prover about their capabilities (see Section 2.5.2).

The central challenge of verifiable digital infrastructure controls is proving that the infrastructure under examination is the only infrastructure that matters for a particular governance issue. For example, even if the Prover shows the Verifier a regulatory information system, there is no guarantee that the Prover is not maintaining a secret parallel system that contains key information relating to the governance question. This is a variant of the asymmetric burden of proof for verifying negative claims (see Section 1.5.2). To demonstrate that the digital infrastructure being shown to the Verifier is the only relevant infrastructure, the Prover might have to find innovative ways to demonstrate that no secret parallel infrastructure could exist.[183]

2.4.1.2 Verifiable claims centered on access to digital infrastructure

Similarly, the Prover might provide structured access to their digital infrastructure in order to demonstrate other claims to the Verifier, such as claims about activities that are monitored by that infrastructure. In theory, the Prover can provide some kind of access to key digital infrastructure so that the Verifier (or the Verifier's privacy-preserving code, as discussed in Section 3.5) can check whether specific claims are true. As noted above, the digital infrastructure would need to be verified throughout to ensure that the Prover does not retain other ways of manipulating its content or filtering information that it receives. Unless compliance is checked in a privacy-preserving way, this approach would face significant political challenges, since it would potentially reveal information that is sensitive and not directly related to the verification of the sender's behavior.

[183] Exploration of this question goes beyond the scope of this report. It should be noted however that data infrastructures are physical, thus allowing the Prover to demonstrate that key buildings or computing systems have no other way to communicate than what is shown to the Verifier. A related concept for data centers is discussed in Section 2.5.2.1.

2.5 Enclosures and security boundaries for AI hardware

Much of the discussion about AI verification so far has centered on making chips hard to tamper with and installing governance mechanisms on them.[184] This report takes a broader stance for the reasons outlined in Section 3.4. In this report, enclosures and security boundaries are not assumed to be *only* around the AI-specialized chips. Overall, this report emphasizes the potential for off-chip mechanisms to help accomplish governance goals, including in synergy with on-chip mechanisms such as confidential computing.[185]

Enclosures can help:

- Partition hardware into separate units (or assemblages), each of which can make verifiable claims as they perform their functions. The act of partitioning hardware provides opportunities for both governing its interactions with other hardware and verifying that governance. Verification hardware may not need to be embedded within the hardware units themselves. Instead, verification hardware could be placed in between the different units.[186]
- Make it possible to make verifiable claims even if untrusted hardware (e.g., a CPU or a prior generation GPU) is doing the heavy lifting. The flexHEG report specifically talks about moving the trust boundary to allow for the verification of activities undertaken on GPUs that lack on-chip verification mechanisms.[187] This is an important insight, because it allows verification mechanisms to work on 1) legacy hardware and 2) chips that are too complicated to be inspected downstream (see Section 2.2.3.3).
- Provide verifiable hardware packages that match the scale and character of the computational operations that are needed. Some models are small enough to fit on a single chip (e.g., for running inference), while others are best run on a large pod of many hundreds of GPUs.
- Mitigate threats of tampering and circumvention at the level of a complete system, not just a single chip—see Section 2.2.4.1.
- Focus verification processes on relatively simple hardware that can be mutually trusted.
- Allow a Prover to make overlapping claims (e.g., about the behavior of chips, pods, modules, and entire data centers)—thus making it much harder for them to circumvent verification mechanisms.

To accomplish these things, enclosures need to be *secure* and *mutually verified*. Each of these concepts is expanded in the subsections below. There are also political questions about the

[184] Aarne et al. (2024); Kulp et al. (2024).

[185] Prior work has described one example of such an enclosure, the FlexHEG. This section expands on some of the concepts from that proposal and shows how these ideas can open up both technical and political options. Petrie et al. (2024).

[186] This description presumes that hardware on such boundaries would have access to the information needed for governance and verification. As will be discussed below, this assumption might require that cryptographic schemes be chosen which allow credible commitments to be made about even encrypted data. see Section 2.5.3.

[187] Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees'.

location and physical control of these enclosures which are examined in Section 3.6. This section will first unpack the concepts of security and mutual verification before going on to describe verifiable claims that can be made, as well as different sizes of enclosures that might be very useful for AI governance and verification purposes: pods, containerized data centers, and traditional data centers.[188] Future work will be required to understand how we might scale up this approach to address political needs.

2.5.1 Security

If hardware or software is to be trusted for sensitive operations, it must be trustworthy. The domains of physical and cybersecurity research and practice are enormous and cannot be adequately summarized here. Institutions intending to use hardware to make verifiable claims must urgently prioritize security. A report by Nevo et al. (2024) lays out many important aspects of the problem.[189] Of particular note are their findings that a) the problem space is extremely complex and multifaceted, and b) few, if any, institutions are realistically able to defend against sustained attacks from the most capable institutions in the world.[190]

The need to make computational infrastructure *verifiable* as well as secure unfortunately introduces new challenges. In order to achieve high credibility, the information flows that support verification of compliance must be relatively continuous—therefore opening up further avenues for attack. While a full discussion of the overlapping and diverging needs of security and verification is beyond the scope of this report, two concepts are worth noting. First, while verification creates new pathways for attack, it also helps bring attention and resources to the problem of robustly securing infrastructure. For example, as explored in Section 2.5.2.2, any data pathway that could plausibly be used for an attack is another candidate domain for additional verification mechanisms. The extraordinary attention and diligence that such mechanisms involve might in fact *improve* security beyond what it might have been in an unverified equilibrium. Second, particular aspects of the technical and software stack for verification must be developed either collaboratively or openly. Collaboration could mean that technical teams from rival states would work together to create and test mechanisms, as American and Soviet teams did for nuclear verification during the Cold War. Open development could be fully public, involving open-source scrutiny from not only the primary states involved in the agreement but also third party states, corporations, and civil society. While the states that are most important to future agreements over AI (such as the United States and China) have enormous technical capabilities, it is still possible that they will miss something if they develop these mechanisms alone or even in concert with each other. Open development could allow for a level of adversarial testing that even the great powers would have difficulty matching. A particularly compelling form of incentive is also possible for open standards:

[188] These subsections generally refer to "data centers" for enclosures in general, since most of these concepts have been most thoroughly examined with regards to data centers. However, all of the concepts can apply to enclosures of any scale.

[189] Sella Nevo et al., 'Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models' (RAND Corporation, 30 May 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

[190] Scher and Theirgart (2024) also note that data center security is a crucial area of work.

bug bounties. If substantial monetary prizes are offered to anyone who can find important flaws in the proposed mechanisms, very intense scrutiny can be expected.[191]

Other aspects of security are overtly political. As noted later, enforcement patterns can shape verification (see Section 3.1.1). These choices can also shape some aspects of security. For example, if it is politically intolerable that stolen hardware still functions, secure facilities such as data centers might install various mechanisms to ensure that hardware that is physically seized cannot be used.[192] Related to enforcement is the question of what happens if one party or the other exits the agreement—a point which is explored in Section 3.6.

2.5.2 Mutual verification of hardware and code

For at least some kinds of verification, if digital stacks are to be used as part of verification mechanisms, those stacks must themselves be verifiable. For example, in order to verify in detail that regulations are being followed regarding AI development or deployment (Section 4.5), there must be heavily automated scrutiny of computations and data. In order to be acceptable to the Prover, the systems undertaking such scrutiny must have been demonstrated to be compatible with their security requirements. Mutual verification of hardware and code is one way to do that.

Existing technologies make possible the mutual verification of digital stacks. To implement this, however, significant effort is needed to actually build (or rebuild) stacks into a form that is mutually verifiable, and then ongoing monitoring (and perhaps rare inspections) would be needed to demonstrate that no hardware circumventions are being undertaken.

How can this be done? Previously discussed technologies such as remote attestation and confidential computing (Sections 2.2.4.2 and 2.2.4.4) allow for software to be verified if the hardware stack has also been demonstrated to be compliant. As noted in Appendix D, digital stacks must typically be verified "from the ground up" (with the exception of cryptographic techniques that allow us to largely ignore the details of communications infrastructure between two secured systems). As noted in Section 2.2.3, some forms of hardware can be verified directly while others would require cooperative creation. Furthermore, as noted in Appendix C.6, mutual verification is expected for the verification infrastructure, but complete secrecy is needed for at least some of the data that is used for evaluating compliance.

Drawing on concepts explored earlier such as confidential computing and hardware-enabled mechanisms, the subsections that follow unpack some of the important verification capabilities that are enabled by the mutual verification of hardware and code.

[191] Simple "game" rules might incentivize early action. Bounties could require *unique* exploits, so it makes sense to register your finding as fast as possible. Rises in bounty value (perhaps with each consolidated version of the open standard) would incentivize more intense scrutiny.

[192] For example, many nuclear weapons have permissive action links, which prevent them from being used by unauthorized people. Similarly, some hardware governance ideas for AI have included proposals for hardware that when faced with a request to undertake non-compliant computations would either refuse to operate or self-destruct. James Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024.

2.5.2.1 Verifiable claims about the equipment contained in a data center

If a Prover wants to demonstrate to a Verifier that particular hardware (and no other hardware) is in a data center and is arranged in particular ways, they have ample tools for accomplishing that goal. Nuclear arms control agreements have employed extensive inspection and monitoring of buildings, equipment, and personnel.[193] Similarly, a Prover seeking to demonstrate claims about their data center can employ well-established methods for proving that their buildings abide by design specifications and that hardware identity, configuration, and connections match expectations, and they can demonstrate that these claims remain true via monitoring of portals (transit points and other choke-point locations)[194] and personnel.

Verifier-employed inspectors might be needed intensively in the early phases of hardware compliance verification and then only brought back intermittently afterwards. They might come back in response to a) hardware reconfiguration requests from the Prover (e.g., to update equipment or replace failed hardware), b) challenge inspections from the Verifier,[195] or c) random inspections to check ongoing compliance.

The Prover can make the hardware of the facility tamper-evident in a variety of ways, thus making a hypothetical hardware tampering attack more difficult and raising the Verifier's confidence that no such tampering is occurring in the facility. Crucially, most mechanisms for tamper evidence would have zero effect on the Prover's security, thus making them a potentially desirable way to reassure the Verifier.

A data center could be monitored by overlapping systems of automated sensors, which could detect a tampering event but could not detect the actual computational activities on the chips.[196] If done correctly, this kind of monitoring also has zero effect on the Prover's security, so they can employ these mechanisms extensively to continuously reassure the Verifier that no circumvention attempts have occurred. These systems would be installed and monitored with the full cooperation of the Prover and Verifier.

As noted earlier, the ongoing miniaturization of technology raises the prospect of new digital circumvention techniques that were impossible in prior decades (see Section 2.2.3.2). Expectations for building design, hardware configurations, anti-tamper packaging, and monitoring must take this into account.

2.5.2.2 Securing and verifying all channels

Section 2.5.1 noted that guarding against attacks requires substantial effort, but also argued that each potential attack channel provides another avenue for verification. This point is

[193] Mauricio Baker, 'Nuclear Arms Control Verification and Lessons for AI Treaties' (arXiv, 8 April 2023), https://doi.org/10.48550/arXiv.2304.04123.

[194] Brian Jennings et al., 'Advanced Portal Monitoring for Arms Control Treaty Verification' (Oak Ridge National Laboratory, 2024), https://www.osti.gov/servlets/purl/2472697.

[195] See Section 2.1.3 for more on challenge inspections.

[196] Sensors could include visual-range and infrared video cameras, acoustic detectors, radar, and other sensors such as accelerometers.

worth emphasizing. To maximize both security and verifiability, it is useful to assume that all hardware that is physically proximate to computational or networking hardware could be used for attacks and should therefore be considered as another pathway for credible verification. For example, in addition to the two key channels laid out for digital verification based on computing hardware (Section 2.2.4.4) and networking (Section 2.2.4.8), one can also add other privacy-preserving information flows by leveraging the information provided by sensors and connections attached to key components such as GPUs, CPUs, and racks. Such channels could include a) power usage; b) electromagnetic spectrum or thermal monitoring; c) acoustic sensors and accelerometers. Each of these has the potential to be the pathway for a security breach. Equally, however, they could be used to help demonstrate to the Verifier that the declarations made in the primary channels are true and complete. For example, while high resolution power usage data can be used to exfiltrate data under certain conditions,[197] it is infeasible for aggregated data to present the same danger.[198] Furthermore, if the data provided via the power usage sampling is treated as sensitive and therefore subject to the same kind of privacy-preserving digital verification techniques as model inputs (e.g., training data and algorithms) or model behavior, then the remaining danger of this information channel can be mitigated further (see Section 2.5.3). The residual danger of an attack along any of these channels can be compared with the gains in transparency that could be reaped.[199] Furthermore, since each hardware mechanism could have its own separate hardware root of trust, physical hardware circumvention attacks would grow increasingly complex as additional parallel data streams are added—since any one of them makes it possible to notice a circumvention attempt.[200] This approach is certainly not a simple or complete answer to the transparency-security tradeoff, but this frame does suggest that encapsulating channels into new, verifiable data streams can be compatible with security needs while also drastically increasing the Prover's ability to demonstrate their compliance. Extensive practical work on each channel would be needed in order to make this vision a reality.

[197] The detailed power usage of a GPU (sampling at 1000 Hz) running an AI model can be used to exfiltrate data even if the stack is employing confidential computing. Lagerros (2025).

[198] Aggregate data can still allow you to accurately estimate compute used and—depending on the resolution and the AI paradigm in use—can help with workload classification. See Lennart Heim et al., 'Governing Through The Cloud: The Intermediary Role Of Compute Providers In AI Regulation' (Oxford Martin AI Governance Initiative, March 2024).

[199] One specific danger for military assets is that any local measurement with an associated timestamp can be used to locate that object via "triangulation" which allows the location of objects to be estimated based on the speed of sound or light along with a sharp signal such as a loud sound or an electromagnetic pulse. Potential mitigations for this danger include never revealing this data outside of privacy-preserving analysis systems and also carefully scrutinizing the assessment code to ensure that it could not be used for this kind of attack.

[200] If securing channels like this is possible, this approach can also provide ways for the Prover to demonstrate to themselves that their system is secure against an increasing number of possible attacks. Taken to its logical extreme, a successful effort of this kind could place so many constraints and checks on the digital system that the Prover might become relatively confident of the integrity and security of their system. However, in considering such a hypothetical it is important to note that presently there are no civilian AI data centers that are believed to be robust against serious attacks by a competent cyber state. Getting to robust security will require years of intense and sustained efforts. Sella Nevo et al., 'Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models' (RAND Corporation, 30 May 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

2.5.2.3 Verifiable communication within a data center

Data center hardware can be set up to allow the Prover to make credible claims about the content of all digital exchanges within the cluster. This can be done by employing networking hardware capable of both hardware-enabled full text logging and the generation of a stream of cryptographic commitments which are sent to the Verifier (see Section 2.2.4.8). The commitments sent to the Verifier are credible because the Verifier can see precisely how and when they are being generated. As explored in Section 2.5.3 below, a stream of commitments of this kind can allow the Prover to securely demonstrate their compliance with regulations at a time and place of their choosing while simultaneously protecting their security.

In theory, verifiable networking data streams like this could be installed between all communicating devices within the data center. In practice, this is impractical. For example, the sheer quantities of data exchanged between AI chips communicating in a pod (see Section 2.5.4.1) might be enormous. A more practical approach would involve making targeted use of this capability to make verifiable claims about the inputs and outputs of key enclosures such as pods or the data center itself.

A number of questions about this scheme are ripe for future exploration, including:

- Can existing highly performant networking hardware implement this scheme in a way that could be trustworthy for the states involved? If not, how feasible would it be to produce relatively performant hardware that can be mutually trusted (see Section 2.2.3.3)?
- If neither category of high performance hardware is available, how can less-performant but trustable hardware be best arranged to provide verification without substantially reducing performance (Section 2.2.3.4)?
- Can this scheme remain credible and practical if it uses gateway machines for each pod to make cryptographic commitments about digital objects in their entirety rather than each raw packet or data block? This would mean that network verification only sees exchanges between the pods, not what goes on within them. If this could work, how big can pods be while remaining governable and practical?[201]
- If no single piece of mutually trusted hardware can possibly fill this role, could two or more independently designed and untrusted systems be installed which make the same cryptographic commitments, with only the Prover's system copying the plaintext data?[202]

[201] Calculations provided in Scher and Thiergart (2024) indicate that inference workloads require remarkably little external bandwidth and that there is a real possibility of training protocols that allow for efficient distributed training. If these calculations are a good estimate of reality, then there should be substantial room to maneuver in the design of network topologies to support inference and training as desired by the data center operator.

[202] Extending this to three systems (sourced respectively from the Verifier, Prover, and one credible third party) could allow for further checks, such as automatically alerting all parties if a commitment mismatch occurs.

2.5.2.4 Digital perimeter: Verifying data center communication with the outside world

A *digital perimeter* is a physical, digital, or hybrid boundary system that prevents data from entering or exiting the enclosed area without being noticed. In a cooperative verification setting, a verifiable digital perimeter allows the Prover to credibly claim that data cannot cross the boundary without the Verifier noticing.

There are four ways to implement a digital perimeter with the technologies described in this report. All of these approaches presume that hardware within the perimeter has been extensively verified to demonstrate that it is incapable of other forms of communication, that physical access by personnel is controlled, and that appropriate safeguards are in place to ensure that additional hardware cannot be infiltrated (see Section 2.5.2 and Section 2.2.3.2). The four approaches discussed in this report are:

- **Air gap**: An air-gapped data center with no ability to send or receive any kind of remote signals.[203] In such cases, secure delivery of digital files can be arranged using physical deliveries of encrypted data and separate deliveries of decryption keys.[204]
- **Network verification**: Verification of all network traffic is possible with specialized hardware (see Section 2.5.2.3). This is possible not only at the data center boundary (inputs and outputs of the whole data center), but also for its subcomponents (e.g., racks or pods) which themselves can be in enclosures. Overall, a nested set of verifiable networks is possible, which would ensure that any communications into or out of the data center would leave multiple traces for verification processes to catch and examine. Circumventing such a system digitally would require circumventing at least two of the hardware roots of trust (the data center gateway and at least one pod gateway).[205]
- **Cryptographic boundary:** Data center gateways enforce encryption on all inputs and outputs.[206] Both the Prover and Verifier provide cryptographic keys to specialized hardware that will enforce that all exchanges use those keys—decrypting all inputs with both

[203] This requirement is much more challenging than it might appear at first glance. Preventing wired electronic communications is relatively easy, but preventing all meaningful electromagnetic signals is much more difficult. All potential modes of transmission would need to be examined and accounted for, including acoustic transmissions and personnel-based leaks. Furthermore, as noted in Section 3.5, it is unclear how an air gapped facility can be reliably verified.

[204] Decryption keys can even be delivered for the precise machine(s) that will be examining the relevant information. By encrypting a key with the public key of the target machine, one can ensure that only the target machine can decrypt the resulting ciphertext as long as the private key cannot be stolen (digitally or physically) from the target machine. Strict data center management and monitoring can help guard against the latter possibilities, since with sufficiently intense attention from both the Verifier and the Prover, physical and digital attacks will be very difficult to hide. Note that achieving security of this level will be very challenging. For more on this, see Sella Nevo et al., 'Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models' (RAND Corporation, 30 May 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

[205] Note that this scheme presumes physical controls. If physical access is available for an attacker, the guarantees provided by this scheme are drastically weakened.

[206] A more intense variant of this would double-encrypt data at the pod level as well as the gateway level, thus requiring at least four decryption keys to reveal the data—two from each primary actor. Another variant would have three or more sets of keys, with one or more controlled by neutral parties in addition to the sets controlled by the Prover and Verifier.

decryption keys and encrypting all outputs with both encryption keys (see also Section 2.2.4.7.1).[207] In sum, this allows both parties to be more confident that neither party can meaningfully copy the data without the consent of the other party, since neither party can decrypt the output data unilaterally. Presuming that the hardware is defended appropriately, this moves the digital perimeter into cryptographic space—thus allowing more freedom with copying encrypted output files.[208] This approach could be desirable on its own, or it could be employed in **conjunction** with other digital perimeter designs to allow for layered checks.[209]

- **Combination** of network verification with cryptographic boundary: Combining the two approaches above would provide both commitments about all traffic. Network verification would make copying very difficult, and the cryptographic boundary should protect the data even if an attacker manages to make a copy.

In closing, a word of caution. Even if you can inspect all data passing into or out of a data center, you won't know its full meaning. Since Verifiers cannot see the inner workings of the data center, the Prover could undertake sophisticated exfiltration attempts.[210] A digital perimeter is therefore an imperfect verification scheme. It can significantly limit the Prover's ability to use their hardware in covert ways that violate an agreement, but alone it is not sufficient to guard against all reasonably possible circumvention attempts. Therefore, to achieve extremely high certainty, other mechanisms should be used in conjunction with a digital perimeter, as explored in the next section.

2.5.3 Verifiable confidential computing

To address the transparency-security tradeoff, the ideal setup of a Prover's data center would approximate *verifiable confidential computing*. Here this is defined as computing that is confidential in that no one else can tell what the Prover did, but the Prover retains the ability to prove (at a time and place of their choosing) what they did in their computations—*exactly* and *completely*.[211]

[207] Encryption and decryption keys would be different, and each of the parties would maintain control of their own private keys.

[208] Copying an encrypted output file does not reveal the information contained in it. The only agents who can read the file are those in possession of all of the needed decryption keys.

[209] This is a mechanism that combines enforcement and verification. It is included in the report because it can substantially change the verification needs of an agreement.

[210] One hypothetical example is that the Prover figures out how to hide data in the least significant bits of outputs (such as timestamps or floating-point numbers), thus exfiltrating information without the knowledge of the Verifier because those bits might normally be assumed to be noise. Noise-free replications of machine learning computations might be possible, but may also be difficult (see Shavit, 2023).

[211] This term should not be confused with *verifiable computing*, which allows for proof that a computation was done, but does not allow for proof that no other operations were done. Another difference between the two is that verifiable confidential computing assumes that the Prover wants the ability to keep their information secret, while verifiable computing does not make that assumption.

There are two aspects of this theoretical ideal:

- **Security**: Confidential computing, or a similar set of security features that allow the Prover to be confident that no one can steal their data during any of the planned computational phases.
- **Verifiability**: The Prover has the ability to robustly demonstrate what they did with their computations. Here it is presumed that they choose to do so in a strictly privacy-preserving way, but in theory they could also simply reveal the relevant data publicly if they deemed that politically acceptable.

No existing system is known to accomplish this goal, but an approximation of this ideal standard appears to be possible using existing hardware. Here is one potential approach, which is also illustrated in Figure 2.1 below.

- **Mutually verified data centers**: Two data centers have mutually verified hardware (see Section 2.5.2.1). Presumably, one is in the Prover's territory and the other is in neutral territory (see Section 3.6.1).
- **Digital perimeters**: Each data center has a digital perimeter (see Section 2.5.2.4).
- **Verifiable operations**: Within the digital perimeter of the Prover's data center, they have arranged hardware that allows them to create one or more sets of cryptographic commitments.
    - Two primary methods for such commitments are explored in this report. These could in theory be implemented in parallel for more robust verification (see Appendix C.5):[212]
        - Confidential computing allows the Prover and Verifier to remotely attest the system's integrity as well as the code that will be run and cryptographic commitments for all input data, output data, and hidden code (see Sections 2.2.4.4 and 3.5).
        - Verifiable networking, where all data exchanges between key nodes of the network can be tracked by cryptographic commitments and later revealed in full (see Section 2.5.2.3).

[212] This report will not explore the full details of how these schemes can be combined, and this should be an area of ongoing research. In general, complex combinations of cryptographic systems can create new failure modes. For this reason, it is worth noting at the outset that some confidential computing details might need to be carefully designed to support any combination. As a concrete but notional example of this, rather than using hardware roots of trust directly as private keys for network traffic, it may be necessary for the code on the secure enclaves to generate a public/private key pair that is *specific for the workload* (such as executing a training plan) and which can be used as the only key for encryption and decryption of network traffic within the "Verifiable operations" hardware. The reason that separate keys for each workload are desirable is because any key used to encrypt network traffic would have to be revealed within the "Verification computations" box on the right side of the diagram. This revelation allows the cryptographic commitments made on the network to be connected to the revealed plaintext data. Revealing a hardware root of trust private key from one or more of the Prover's machines is not advisable even if it were a practical possibility, since those hardware roots of trust are a key part of the Prover's confidence that their computations were secure in the first place.

    - Furthermore, additional supporting streams of cryptographic commitments are possible if needed to increase the credibility of the Prover's claims, drawing on data recorded by other sensors placed in the data center (see Section 2.5.2.2).

- **Verification computations**: Within the digital perimeter of the neutral data center, confidential computing is used to allow the Prover and Verifier to cooperatively undertake verification computations without revealing any sensitive information (see Section 3.5).
- **Common knowledge**: Both the verification results and the cryptographic commitments relating to the verifiable operations are common knowledge for the Prover and Verifier.

In sum, the Prover is able to demonstrate that their computations are compliant while preserving their security. Furthermore, presuming that the Verifier believes that the evaluations they create for testing compliance are likely to catch non-compliance, this overall exchange can allow the Verifier to achieve high confidence in the Prover's compliance.[213] An important facet of this certainty is that this scheme allows the Prover to demonstrate the compliance of *all* computations that take place within a designated portion of their infrastructure. Furthermore, in this scheme the central compliance tests are run against the plaintext data and not a noisy proxy such as electric power draw or heat emissions, therefore allowing detailed compliance checks rather than broad guesses about activities.

This approach has limitations in its scope. First, it is a digital verification system, so all rules must be operationalized in digital operations over digital objects. These operations might be enormously complex, such as having a specialized AI system read text or watch video feeds, but they must be digital. Second, this approach is not sufficient (alone) to undertake verification of computations that include external tools such as the Internet. This might be a salient domain of verification given that there already exist AI tools that use the internet extensively as part of their workflow.[214] The usage of external tools would allow the AI model to both exfiltrate and infiltrate data, thus raising potentially unresolvable problems regarding verification. Third, it presumes that verification of the digital objects can be undertaken by software alone. The potential for adding humans to this scheme is explored in Section 3.5.

This stack avoids the centralization of power, since the Prover remains in control of their own hardware and data. The neutral data center is a focal point for governance processes, but there is no requirement that only one facility be used for that purpose. Multiple neutral facilities in different jurisdictions might be desirable for providing multiple independent checks of compliance for crucial activities (e.g., in case of a red flag or political challenge).

This stack also allows the Prover to demonstrate their compliance with a large number of different requirements. This infrastructure approach allows the Prover to build a single system which can allow them to demonstrate their computations' compliance with global rules,

[213] Furthermore, further streams of data can be added which can increase certainty of compliance without any large changes to security (see Section 2.5.2.2 and Appendix C.5).

[214] 'Claude Can Now Use Tools', Anthropic, 30 May 2024, https://www.anthropic.com/news/tool-use-ga.

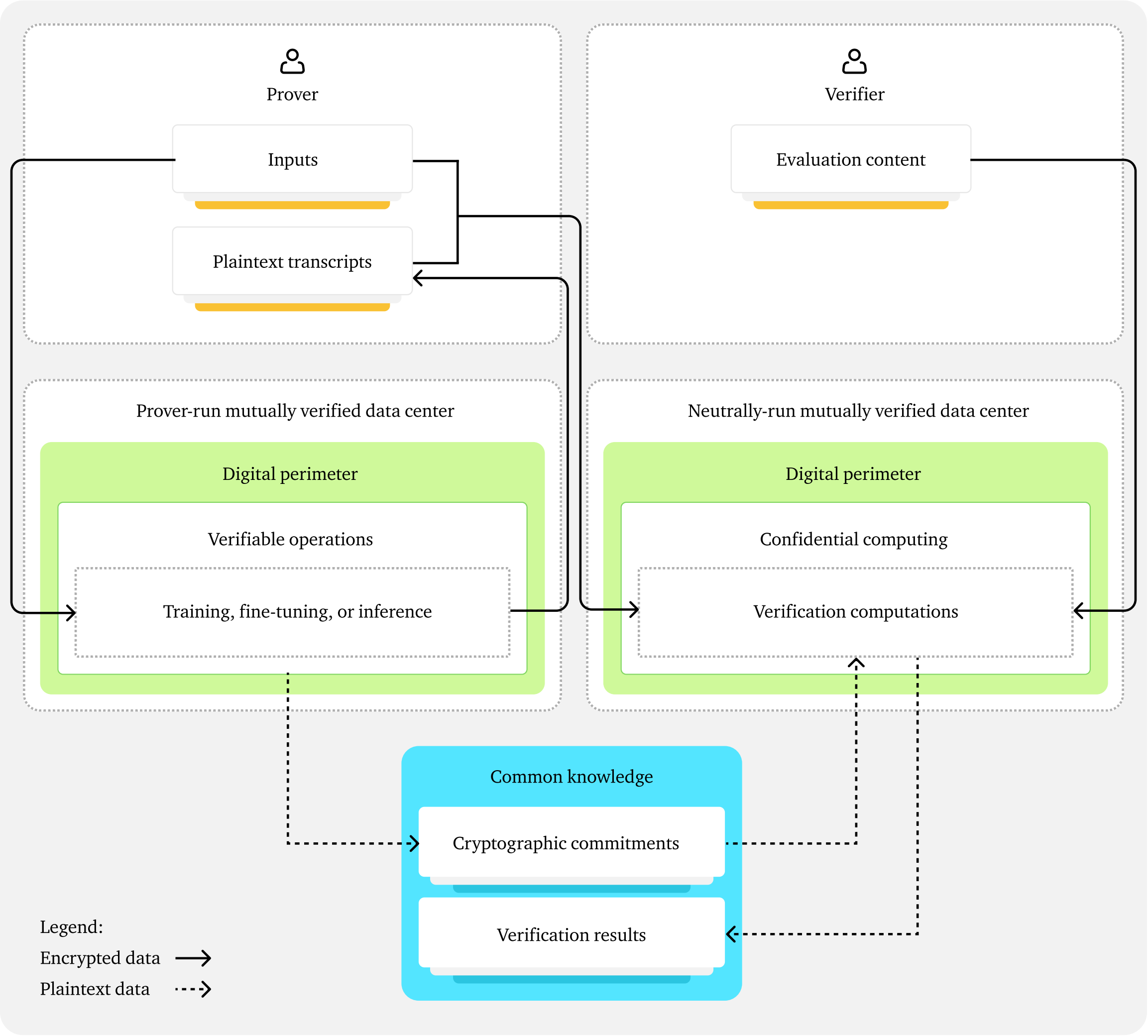

Figure 2.1: A schematic summary of verifiable confidential computing as applied to training, fine-tuning, or inference, and the associated verification of claims about rules followed during those processes. The conceptual process is as follows: The Prover keeps secret their *inputs* and the *plaintext transcripts* of their computing operations. The Verifier keeps secret their *evaluation content*. The *Prover-run mutually verified infrastructure* allows computing operations to be kept secret while also producing a stream of *cryptographic commitments* which become *common knowledge*. The cryptographic commitments are used in the *neutral mutually verified infrastructure* to prove that the data revealed securely within this infrastructure by the Prover is in fact the true data that was produced during the earlier computations in the Prover's infrastructure. Privacy-preserving *verification computations* are run on the true data and the evaluation content, producing a set of *verification results* which become common knowledge. Finally, the sensitive data is verifiably deleted from the neutral infrastructure.

minilateral rules (e.g., potential OECD standards), regional rules (e.g., European Union laws), and even domestic law.[215]

2.5.4 Different enclosure sizes for AI verification

Enclosures provide physical and logical separation between their inner hardware and the outside world, thus allowing governance and verification processes that are otherwise difficult to achieve. For AI verification, one proposal is to enclose individual AI-specialized chips in a way that makes them easier to govern and verify,[216] while many other proposals—as well as much of the discussion in this report—focus on data centers as the primary governable and verifiable unit of computational hardware. Both extreme ends of this spectrum are certainly valid, and they provide different kinds of verification abilities with different tradeoffs. It is very plausible that a combination of both will be desirable for states engaging in detailed agreements about their AI development and deployment. However, other, intermediate hardware enclosure sizes may also be extremely important. Each of the following subsections describes an enclosure size and why it might be distinctly useful.

2.5.4.1 Pods

A pod is a unit of AI-specialized computational hardware that contains dozens to hundreds of AI-specialized chips which are intended to be used together for related tasks such as model training or inference. Today, more than one AI-specialized chip is required to train or quickly run inference on even medium-sized models. AI-producing organizations attempt to design pods that are optimal for their workloads, but causation can also flow in the other direction, as available hardware units can strongly shape which arrangements are optimal. For example, NVIDIA is producing 72-GPU rack-sized machines,[217] which may lead AI-producing organizations to organize their workloads around that pod size.

Since the pods are designed to accomplish crucial workloads for training or inference,[218] they are likely a useful unit (or agglomeration) of hardware to enclose and govern. For example, a pod is particularly well-suited for being the level of hardware organization for making verifiable claims about model training steps or inference exchanges, both of which are discussed elsewhere in this report as key operations that states might seek to govern and verify. Pod governance and verification could perhaps be conducted through a specialized gateway machine. The hardware internals of pods could be inspected and their overall physical integrity

[215] Using such a scheme for domestic compliance checks could allow domestic regulators to very credibly demonstrate that they lack the ability to expose private personal or business data via their compliance checks. Such an approach to regulatory oversight might therefore face less political pushback from groups that are sensitive about personal or business secrets. To implement this approach in high-resource environments, a domestic regulator might choose to build their own “neutral” facility with the mutual verification of industry and civil society groups that seek to defend the privacy of their constituents. In low-resource environments, this scheme could allow compliance demonstrations for all applicable regulations even if no new infrastructure is added.

[216] James Petrie et al., ‘Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees’, 23 August 2024.

[217] ‘NVIDIA GB200 NVL72 GPU – Optimized for AI and Data Centers’, NVIDIA, accessed 15 March 2025, https://www.nvidia.com/en-gb/data-center/gb200-nvl72/.

[218] Different pod sizes will likely be desirable depending on the workload.

could be monitored (see Section 2.5.2.1 and Section 2.5.2.3). To allow for improved verifiability and a reduced vulnerability to proximate side-channel attacks, pods could be physically isolated by at least a few feet from other data center hardware.

One challenge with pods is how to handle maintenance issues with their components—especially their AI-specialized chips, since those often have issues. If a chip dies, it might be possible for the pod's hardware to verifiably show that it has been turned off (see also Section 2.5.2.2 for parallel ways to make this claim). However, the data center operator might want to replace problematic components quickly, or otherwise update or tweak hardware on a regular basis. These operations would incur a risk to the verification process unless the Verifier had an inspector oversee the operation, just as they would have done for the data center or pod previously to verify it in the first place (see Section 2.5.2.1). Further work should be done to understand the efficiency overhead of having Verifier inspectors present for all hardware work on chips, including those within a pod.

2.5.4.2 Containerized data centers

Small self-contained data centers might be verifiable in important ways while remaining highly secure for extremely security-sensitive organizations such as militaries. Detailed location information for military hardware can be a key vulnerability (see Section 1.5.1.1). This raises the question of whether digital verification processes might be abstracted in a way that reliably protects location information while still allowing politically important verification to occur. This section describes one potential approach to solving this problem: containerized data centers.

A containerized data center would be contained in a standard shipping container of whatever standard size is convenient for the Prover.[219] The hardware within would be mutually verified in a controlled setting (see Section 2.5.4.3) to demonstrate precisely what hardware was inside and in what configuration (see Section 2.5.2.1). The container shell (or a layer within) would be the enclosure security boundary, and this would encapsulate the verified AI hardware. Hardware mechanisms could be installed to provide parallel channels of information to demonstrate that the inner hardware has not been tampered with (see Section 2.5.2.2)[220]—with all digital exchanges being mediated through Prover-run network infrastructure to allow them to obfuscate the location.[221] Once the inner hardware has been verified and the verification data streams enabled, the container could be moved by the Prover to any location.

[219] The reasoning behind this suggestion is similar to the reasons why some weapon systems are hidden in containers—they are generic, mobile, and common. Furthermore, a shipping container may be large enough for a non-trivial unit of computing power while also being small enough to not rely on unique or rare infrastructure—since such rare infrastructures could give away its location.

[220] Depending on the level of security needed, many techniques could be used in parallel. Cryptographic exchanges could use the hardware roots of trust for the whole enclosure as well as subsystems such as chips, pods, network switches, and monitoring hardware—thus making circumvention more difficult. While stealing one root of trust might be reasonably tractable for a state actor given enough time, breaking several parallel systems without triggering any alarms should be much less feasible.

[221] The Prover could of course cut off this communication, but that would be noticed by the Verifier. Some amount of enforcement could be built into the agreement, such as an automatic requirement that an enclosure that loses communication for X seconds is automatically required to be brought in for reinspection.

In sum, the container would house one or more pods of AI-specialized chips with a clear security boundary around them—thus enabling the Prover to demonstrate to the Verifier via credible data streams that the activities of the chips are compliant (see Section 2.5.3) while simultaneously demonstrating that the inner hardware has not been tampered with. The Verifier does not know the actual location of the hardware, but retains sufficiently continuous digital contact to be relatively confident that the enclosure is not being tampered with. Presuming that the roots of trust are not broken, this allows the Verifier to monitor and verify the hardware continuously, even if they don't know where it is.[222]

Three major issues of this approach are worth highlighting, along with speculative avenues of mitigation for each. First, data streams from the enclosure might contain implicit location information, such as electromagnetic (EM) signals or acoustic noise with recognizable signatures. Hypothetically, a Verifier attempting to locate the containers could trigger “loud” events of either kind—such as an EM pulse or a minor earthquake—and observe the exact timing of the expected spike in the readout. This would allow them to locate each container using a small number of such events triggered from different locations.[223] This type of attack could potentially be mitigated through a combination of acoustic and EM shielding and by processing all data through privacy-preserving evaluations only. Since the Prover can demand to see the code of such evaluations, they should subject it to extreme scrutiny. Sensors that are lower accuracy and lower time resolution may also help—since sensitivity below a certain level may make them infeasible vectors for location attacks. Finally, there may be ways to obfuscate the precise timing of events within the enclosure to ensure that verification code cannot extract precise reactions. While none of these mitigations can be expected to be perfect, they may allow the security concerns of the Prover to be addressed sufficiently for a deal to happen.

Second, significant inefficiencies might be introduced by the requirement that hardware not be manipulated after it leaves the verification facility. If hardware malfunctions must either be left unaddressed or require another visit to the hardware verification facility, they could end up costing a moderate amount in terms of downtime, relocation, and reinspection. One potential mitigation is for the participating parties to plan for hardware malfunctions at a certain rate and ensure that their AI computations can proceed robustly regardless—thus reducing the rate of needed visits to the verification facility. Some over-capacity or graceful degradation might be required for such a scheme.

Third, only limited agreements might be possible in areas where there are extreme security concerns. It may or may not be realistic for agreements over military hardware to require byte-by-byte commitments that are then verified in another facility. If this is the case, it

[222] This approach embraces the concept of a *logical* location or a *relative* physical location with respect to a verified and monitored space—the enclosure. This approach can be contrasted with the idea of an *absolute* physical location (see Section 2.2.2.1).

[223] The details of such an attack are beyond the scope of this report.

is worth noting that even less specific or aggregate data could still be very useful for states attempting to reassure one another for security purposes.[224]

Containerized data centers are a speculative proposal, intended as a workable way to provide deep and continuous verification of rule enforcement while also ensuring that states can keep their military AI resources secure. The general approach appears worth exploring further, but it should be regarded as highly speculative until further work has been done.

2.5.4.3 Hardware verification facility

A hardware verification facility is a mutually verified and secured facility that allows other objects to be brought in and verified by both parties.[225] Cooperative facilities of this kind can solve information problems that may not be solvable via either sequential verification (where one actor verifies the object before handing it off to the other actor) or unilateral action. For example, certain kinds of hardware scrutiny might involve substantial and complex testing hardware. While each of the Prover and Verifier might be able to verify that a piece of hardware is compliant, if either actor subjected the hardware to intense tests in an opaque facility, the other actor might rightly fear that a hardware circumvention attack was being undertaken. No ordering of sequential verifications can solve this problem, but cooperative verification could do so.[226]

If both the Verifier and the Prover need to simultaneously prove that mobile hardware is both compliant and unaltered, a hardware verification facility is one way to address this problem. As noted for containerized data centers (see Section 2.5.4.2) and for mobile AI-enabled devices (see Section 4.5.2.3.6.c), hardware could be brought into facilities to allow parallel efforts by inspectors and engineers from both parties. Similarly, hardware that supports verification of networking or other channels might need to be mutually verified in this

[224] Consider a Verifier and Prover who are convinced that AI *development* (not deployment) is the primary way by which power would shift between them, and thus they seek to monitor each other's development with care. In this scenario, they could agree to undertake only certain kinds of operations (such as inference) in containerized data centers, with all major development work taking place in larger data centers with well-known locations. This agreement both simplifies their verification needs and provides them with some strategic stability. Judging from historical precedent, states can be expected to be leery of verifiably regulating in detail their use of final systems such as weapons, but they tend to be somewhat more willing to regulate what they build and maintain during peacetime. AI development being in larger centralized data centers means that such development is *easier to target*. So if a state begins the process of exiting the agreement, their development resources will be at risk but *not their warfighting resources*. This disincentivizes first strikes to some extent, since the targeted state would have a very good likelihood of retaining at least some of its mobile inference capacity. In this scenario, since these states are assumed to believe that only AI *development*—but not inference—can substantially shift relative power between them, holding that development under greater scrutiny and at greater vulnerability to attack would make it more possible for the states to bargain over the "roots" of their power and thus avoid creating unacceptable power shifts that could lead to war. For more about why bargaining over the roots of power is advantageous, see Thomas Chadefaux, 'Bargaining over Power: When Do Shifts in Power Lead to War?', International Theory 3, no. 2 (2011): 228–53.

[225] This is an enclosure to the extent that it includes a strict security boundary with physical and digital transit restrictions. Unlike the other enclosures discussed, it is not primarily focused on housing AI-specialized computing hardware.

[226] If the primary concern is that the *final* actor who takes the hardware into their facility can modify it, with the further presumption that the other actor could detect that modification if they test the device afterwards, then there are potential sequential protocols (such as using a radiological random number generator to randomize the testing order in a testing phase of indefinite length) that could address this problem as long as the verification processes are fast and cheap.

manner, since both sides would want to ensure that the hardware is designed only to achieve its narrow purpose and has not been modified by the other party (see Section 2.2.4.8 and Section 2.5.2.2).

3 Political options and tradeoffs in AI verification

3.1 Which stages of the value chain should be governed?

A key political question for the creation of a verifiable international agreement is which stage of the AI value chain should be governed and verified. For regulatory agreements (see Section 4.5), interventions can be aimed at different parts of the timeline of AI development and deployment, ranging from before the model is created up until it is running inference or being deployed on a mobile device.[227]

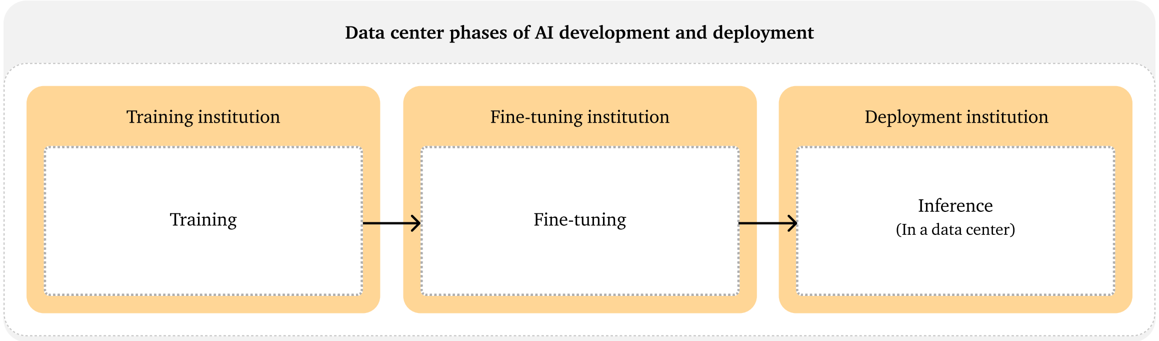

Figure 3.1: A highly simplified AI value chain through to data center inference. A single institution may also manage multiple stages.

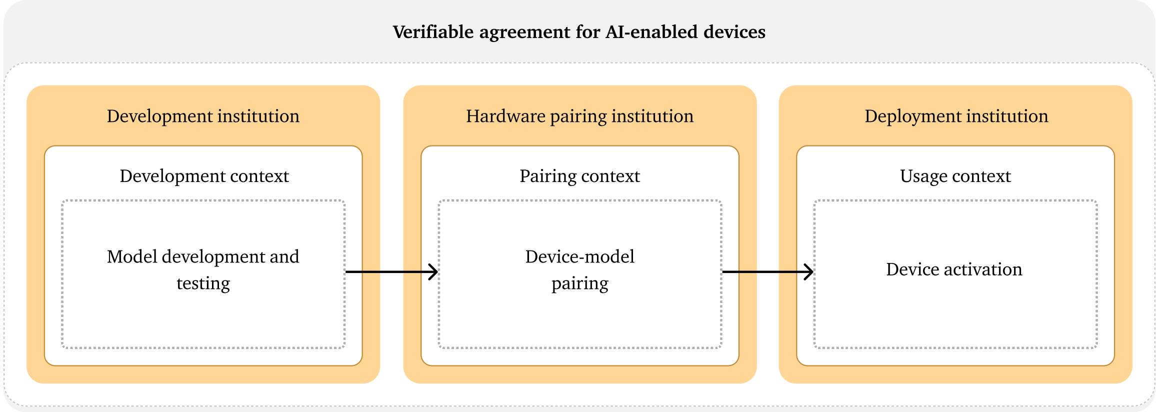

Figure 3.2: A highly simplified illustration of an international agreement over the AI value chain for AI-embedded devices. One institution might manage multiple phases of this process.

[227] See Section 1.5.3 for an introduction to these terms and stages.

Governing each phase comes with potential costs and benefits, and not all rules can be applied at each phase. In particular, rules that require differential model behavior in different contexts cannot be implemented via rules on model creation.[228] For example, differentiating between white hat and black hat cyber efforts may be impossible without detailed knowledge of the usage context—so a model responding to queries would have no way of differentiating between the two.[229] However, as noted in Section 1.4, upstream rules can have a significant effect on the downstream ecosystem and even on the states' perceptions about macroscopic concerns such as relative power. Upstream rules have the potential to more powerfully shape the industry, but they are too blunt to solve every governance problem.

The different phases can also be seen as multiple opportunities to ensure that rules are actually being followed and thus reassure the parties to the agreement (see Appendix C.5). In this sense, governance of each phase could diverge to some extent in its details, but in each downstream layer it would also be possible to check some aspects of the governance that took place in the earlier layers.[230] For example, if certain types of data are forbidden in training, downstream evaluations can also test for related capabilities that the AI could only have learned from the forbidden data.[231]

3.1.1 Enforcement shapes verification

Most forms of agreement enforcement are outside the scope of this report as they include the vast range of actions that are available to states, including diplomatic actions, sanctions, and war.[232] However, enforcement of some rules is implicit in some of the proposed verification approaches described here, as these enforcement choices can significantly shape the associated verification problems. This report does not claim that baked-in enforcement of these kinds would be enough to ensure that a given agreement would be deemed politically credible. At most, the points discussed below delineate the shape of the strategic problem without fully determining it.

Three categories of enforcement are highlighted in this report. First, a number of proposed hardware-enabled mechanisms are deliberately restrictive in ways that change the shape of the verification problem. Licensing, for example, can shift verification and enforcement to before the computations of interest take place (see Section 2.2.4.5). Similarly, enforced encryption of outbound data can help actors move verification computations to a time and

[228] Hypothetically, future models might be sophisticated and well-understood enough that detailed context-specific downstream rules could be effectively enforced through upstream governance—such as rules about model inputs or behavior—but for now this remains a speculative idea.

[229] In this example, the model does not have knowledge of its usage context and therefore cannot tell if it is responding to queries from appropriate authorities (such as "white hat" cybersecurity professionals) or nefarious actors ("black hat" hackers). The social implications of the model's responses might be drastically different in the two scenarios even if the queries themselves are identical.

[230] This is analogous to the "swiss cheese" defense model that has been proposed in other areas, including AI safety. Md Shamsujjoha et al., 'Swiss Cheese Model for AI Safety: A Taxonomy and Reference Architecture for Multi-Layered Guardrails of Foundation Model Based Agents', in 2025 IEEE 22nd International Conference on Software Architecture (ICSA) (IEEE, 2025), 37–48, https://ieeexplore.ieee.org/abstract/document/10978931/.

[231] A remaining problem on this front is that of password-locked sandbagging models, discussed in Appendix G.

[232] For the theory of enforcement and verification used in this report, see Section 1.4.

place that is better suited for both, providing credible assurances of information security to each actor while also allowing political flexibility through the timing of verification (see Sections 2.2.4.7 and 3.1.3 as well as Appendix C.2). Second, robust hardware security can involve mechanisms that disable the hardware if it is removed from its expected configuration (see Section 2.5.1). Third, enforcement is also implicit in choices about where to locate resources, since parties exiting an agreement might be able to seize any resources located on their territory (see Section 3.6).

3.1.2 Strict governance of even one part of the AI value chain can be useful

It may be possible for strict decisions at one level of the AI value chain to substantially change the governance problem for downstream levels. Among other possibilities, upstream controls can shape which actors can participate in downstream activities and also simplify the associated verification.

Here are several illustrative examples of potential interactions of this shape, roughly ordered from least speculative to most speculative:

1. Placing controls on AI-specialized hardware—and on the technologies that allow the creation of that hardware—can shape which actors can gain access to a given level of capability (see Section 1.3).
2. Sufficient control of inference-specialized hardware[233] could ensure that only verified models are allowed to run at large scale.[234]
3. If model training data can be controlled sufficiently—or robust unlearning employed at scale—then it might be possible to create models that are verifiably more safe for broad use and thus amenable to both open sourcing and widespread use.[235]
4. Strict governance of model creation can:
    a. ensure that copies of important models cannot be made. This helps mitigate the danger of proliferating capabilities, as well as assisting with the governance of fine-tuning and inference (see Section 4.5.1.4).

    b. ensure that models with robust CBRN capabilities are not created (see Section 4.5.1.2.2).

[233] Currently, AI hardware is fairly general-purpose, with some differential abilities available for different classes of chips. Speculatively in the future we might see intense hardware specialization toward different workloads.

[234] This is only moderately speculative since there are only a very small number of suppliers of such hardware, thus allowing key governments to both track and govern this hardware if they choose to. Such tracking will certainly not be perfect since there are many untracked chips in the wild, but even tracking a substantial and growing portion of the world's inference capacity could have very meaningful governance ramifications.

[235] As of this writing, it is unclear whether models in the current paradigm can be made robustly safe in this way even if a capable team attempted to do so. For more on unlearning, see Fazl Barez et al., 'Open Problems in Machine Unlearning for AI Safety' (arXiv, 9 January 2025), https://doi.org/10.48550/arXiv.2501.04952.

    c. ensure that models beyond a certain size[236] are infeasible to create (see Section 4.5.1.2.1).

5. Common tools such as Internet browsers could be upgraded to allow them to check certificates for all AI tools that they interact with, similar to how transport layer security ("https") is strongly incentivized by browsers today.[237]

3.1.3 Timing of verification

Verification activities are not always closely tied in time to the activities that they are verifying. In fact, at least four different moments in time might be relevant for complex schemes: 1) when rules are actually implemented by the Prover, 2) when the Prover provides a commitment[238] to the Verifier about what was done, 3) when the Verifier is able to check the Prover's claims, and 4) when post-verification enforcement takes place if needed.

The time gaps between these phases can be due to either technical or political choices. For example, technical limitations may mean that in-flight verification of inference rules is infeasible at scale (see Section 4.5.2.3). Relatedly, the security sensitivity of both the Prover and Verifier might cause them to want to use a highly specialized verification facility (see Section 3.6.1), rather than Prover-operated hardware, for checking compliance with rules about AI development or deployment.

Generally speaking, the Verifier would like to keep delays small to minimize the time that a Prover could be acting out of compliance. Equally, however, the Prover might generally prefer larger delays, since escrowing compliance data can also help them alleviate their concerns about security risks (see 'Cryptographic escrow' in Appendix C.2). It might also be the case that both actors might desire *tighter* timelines to make their agreement more robust, since as noted in Section 1.4, rapid iteration can allow for more robust cooperative equilibria.

3.2 What activities need to be verified?

A central political question for the negotiation of verifiable agreements is: *What activities need to be verified?* A related question, discussed below, is: how much certainty of compliance is needed? (See Section 3.3)

The breadth of activities to be governed and verified could touch on many domains, but only one will be discussed here: compute hardware. One of the most technically challenging types of agreement to verify that is discussed in this report is the regulation of data center-based

[236] Often model size is measured by the model's number of parameters or quantity of compute used in its creation.

[237] This proposal would require significant changes in the entire digital ecosystem which are probably infeasible in the short run. For example, all APIs could be required to carry certificates of their host institution and the AI models they employ. Such a systemic scheme could perhaps borrow from ideas raised by the Coalition for Content Provenance and Authenticity, visible at https://c2pa.org.

[238] See Section 2.2.1.3.

AI development and deployment (see Section 4.5.1 and Section 4.5.2.3), and the primary proposed mechanism involves verifiable controls on computational hardware.

Presuming that a verifiable agreement is being negotiated, one of the open questions will be how much hardware is “in scope” for the agreement and must therefore be regulated and verified. If the negotiators decide to focus on only the kinds of hardware with maximal AI capabilities, they will focus on cutting edge AI-specialized chips.[239,240] If even prior generations of AI-specialized hardware are of concern, then older AI-specialized chips would also be included. Going further still, if commodity hardware such as consumer GPUs is deemed a concern, then a staggering number of chips would be in scope.[241] Similarly, if states are negotiating the rollout of a governance and verification agreement for AI, they might choose to begin with the cutting edge chips and broaden from there, thus allowing their efforts to focus on the most important resources first. Furthermore, verification activities—or the hardware’s compatibility with the transparency-security tradeoff—may be dependent on chip features such as confidential computing, and only relatively new chips (as of this writing) have that capability.

3.3 How certain must the Verifier be of the Prover’s compliance?

Another salient political dimension for the negotiation of verifiable agreements is how certain the Verifier needs to be of the Prover’s compliance. Note that this is related to the question discussed above regarding the activities to be verified. These two dimensions may be politically tied together or may vary somewhat independently. Consider the scenario where a foolproof (perfect certainty) verification system is installed which covers 80% of the Prover’s computational hardware. How concerned should the Verifier be about that remaining 20%? Other than the general guidance provided above about the relative importance of different kinds of hardware, there is no short answer to this question. Depending on the political reasons for the agreement, 20% of activities being unseen might be politically reasonable or unreasonable.

Presuming that some portion of activities have been designated to be verified, the question still remains in a more specific form: for the verified activities, how much certainty of compliance is needed? No agreement in history has been perfectly verifiable, because perfect

[239] A few million cutting-edge AI-specialized chips exist today.

[240] Note that some current high-end GPUs are capable enough at AI operations that their governance may be needed—or future consumer GPUs might need to be limited in their ability to undertake AI operations. Erich Grunewald, ‘Are Consumer GPUs a Problem for US Export Controls?’ (Institute for AI Policy and Strategy, May 2024).

[241] The vast majority of consumer GPUs are significantly slower and less capable than the hardware used by frontier model developers, or are embedded in consumer devices such as gaming consoles and PCs, and are otherwise not very useful for large-scale AI computations. This report assumes that decentralized training and inference is possible, but nonetheless assumes that AI-specialized compute has a significant computational advantage over other hardware—and this advantage is increased when compute can be placed in large concentrations. While consumer GPUs are unlikely to play a key role in creating or running the largest and most capable models, these devices may nonetheless be relevant for governance depending on the political needs being addressed. See also Grunewald, 2024.

certainty was not reasonably achievable or practical. In some cases, such as the INF Treaty, rather high certainty was achieved through years of nuanced negotiations and a tight focus on a specific kind of activity.[242]

Political actors will need to decide how much certainty is desired, and what level of certainty the agreement will require given other tradeoffs such as time, costs, and security risks. Luckily, fine-grained calibrations of certainty might be possible, because there are a relatively large number of tools available that can be ratcheted up or down in their intensity depending on the level of certainty needed—particularly for hardware-centric governance. Verification mechanisms can be combined and even arranged into independent systems that allow for a variety of different ways of catching non-compliance (see Appendix C.5). States can also vary the breadth of their agreement (Section 3.2), the phases under verification (Section 3.1), and many specific facets of how information is collected and revealed (for one example among many, see Section 2.2.4.3). Agreements which focus on the verification of personnel are much more constrained, since oversight of personnel is not scalable in the same way as hardware-enabled oversight. For those agreements, the level of realistically achievable certainty might be much lower than what is achievable with hardware.[243]

3.4 On-chip vs off-chip hardware mechanisms

If hardware mechanisms are to be employed as part of the verification stack, one key question is whether these mechanisms should be installed within AI-specialized chips themselves or within other hardware components. On-chip mechanisms are potentially desirable for a number of reasons, but they are also subject to particular challenges which other forms of hardware mechanisms do not have.

On-chip mechanisms are potentially desirable in many ways, only some of which will be explored here.[244] First, AI-specialized chips are perhaps the highest-leverage factor in an AI governance system, since they are physical, countable, and come from a highly concentrated supply chain. Second, crucial parts of the AI value chain are heavy computational workloads on these very chips. Third, data center-quality AI chips are currently very distinct from the semiconductors used in all other domains—including in general-purpose computation and embedded devices across the economy. Fourth, on-chip mechanisms could allow for the creation of mechanisms that are difficult to tamper with, thus making it infeasible for low-resource actors to circumvent the rules.[245] Fifth, on-chip mechanisms allow for close coupling between verification and enforcement mechanisms, such as licensing systems un-

[242] Part of the challenge of the INF verification provisions was distinguishing very similar missiles from one another when one missile was governed under the agreement but the other was not. See Toivanen, 'The Significance of Strategic Foresight in Verification Technologies'.

[243] Of course, for domains that are amenable to personnel and hardware based verification, both families of approaches could be employed in parallel.

[244] Aarne, Fist, and Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'; Kulp et al., 'Hardware-Enabled Governance Mechanisms'.

[245] As discussed in Section 2.2.3.3, the sheer complexity and layers in advanced chips make them difficult to understand, let alone modify in a hardware-level attack. By contrast, simple single-layer semiconductors built on old semiconductor nodes might be successfully attacked by relatively unsophisticated actors.

der which remote permission would be required in order to unlock the full capabilities of the chip.[246]

The challenges of on-chip mechanisms are similarly numerous and nuanced, so only a brief description of a few of them will be provided here. First, utilizing on-chip mechanisms for governance purposes requires that new chips be produced with those mechanisms which can then gradually supplant existing chips—a process requiring at least a few years of lead time. Therefore, if an agreement required a new on-chip feature, the preparation of that feature would need to occur years before the agreement could be fully implemented (see Section 2.2.3.1). Second, conducting mutual verification of on-chip mechanisms on leading edge semiconductors would require extensive verification of the activities of leading chip design and fabrication companies, who will be extremely protective of the secrets that enable their competitive advantages. Without mutual verification during chip creation, states would have to either use verifiable semiconductors (most likely older node chips which are less performant) or accept that they cannot verify that the chips are not compromised or backdoored somehow (see Section 2.2.3.3 and Section 2.2.3.4).

In sum, on-chip mechanisms have great promise due to factors such as their close proximity to a key node of governance and their potential tamper resistance, but they also seem likely to face political challenges centering on the tension between the necessity for mutual verification and companies' desire to protect extremely valuable trade secrets. By contrast, off-chip hardware mechanisms are somewhat more conceptually distant from the computations at the heart of the AI value chain, but they have much more potential for politically feasible mutual verification.[247] Both on-chip and off-chip hardware mechanisms appear worthy of greater scrutiny.

3.5 Security-preserving digital verification: Are humans needed in the loop?

Many of the verification mechanisms discussed in this report require a way to securely evaluate sensitive information without revealing either the information or the detailed contents of the evaluations to anyone.[248] This subsection describes a continuum of ways that this problem can be solved, thus allowing the Prover to demonstrate their compliance to the Verifier without the Verifier or the Prover learning any information which is not strictly required. The continuum explored below ranges from purely automated privacy-preserving computations to a scenario where human assessors are given wide latitude to explore the provided information to ascertain compliance. This continuum contains a risk-risk tradeoff. Purely automated evaluations are somewhat less trustworthy for the Verifier than a human assessor

[246] Kulp et al., 'Hardware-Enabled Governance Mechanisms'.

[247] For example, early work on hardware enclosures has illustrated how untrusted AI chips can be made governable and verifiable by enclosing them in mutually verified hardware. See Section 2.5 and James Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024.

[248] As discussed in later sections, digital objects being checked for completeness and compliance could include training plans, models, training transcripts, inference plans, and inference transcripts.

(who would also be augmented with all evaluation tools). Equally, however, the Prover would rightfully worry that a human assessor would find ways to remember or transmit secret information beyond that needed to verify compliance with the agreement.

Confidential computing allows the verifiable execution of a set of computations which accomplish a computational task without revealing extra information to any of the parties (see Section 2.2.4.4).[249] Importantly, confidential computing allows both the Prover and the Verifier to review code before it is run—but it also allows them to protect crucial data from each other.[250] Furthermore, this approach may address the computational side of the "*Who watches the watchers?*" problem—where verification processes themselves must be subject to verification, with the Prover being able to verify that the Verifier is running appropriate code.[251] Confidential computing capabilities are available on recent GPUs, such as the H100 from NVIDIA, so nothing fundamentally new needs to be added to leading AI-specialized hardware in order to allow confidential computing.[252] In a nutshell, confidential computing would allow sensitive data from both the Prover (model and inputs) and the Verifier (evaluation tools, content, and even AI systems) to be present on the same system while retaining full information security for both parties. This allows even extremely elaborate verification computations, which might be required for sufficient transparency that the Verifier is reassured about the Prover's compliance.

Another more speculative approach to this problem is that of zero-knowledge proofs. A zero-knowledge proof allows a Verifier to know a governance-related fact with certainty despite not having direct access to any of the data that proved that fact. Zero-knowledge proofs are a potentially ideal approach for the transparency-security tradeoff, but it is unclear whether all (or even most) AI governance questions can even be answered using zero-knowledge proofs, and the computational burden of existing approaches is enormous.[253] While zero-knowledge proofs may be an ideal solution to verification in some senses, they remain impractical to apply at scale as of this writing.[254]

[249] Aarne, Fist, and Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'.

[250] 'When Data Sharing Is a Problem, PySyft 0.9 Is the Solution', OpenMined Blog, 6 August 2024, https://blog.openmined.org/announcing-pysyft-09/; Andrew Trask and Irina Bejan, 'Privacy, Security, and Innovation – Friends Not Foes' (Center for Security and Emerging Technology), accessed 24 January 2025, https://cset.georgetown.edu/event/privacy-security-and-innovation-friends-not-foes/.

[251] For more on the "Who watches the watchers?" problem, see Appendix F.

[252] It should also be noted that confidential computing is only strictly required for the privacy-preserving verification facility. The mechanisms outlined in this report can be accomplished without confidential computing capabilities in the AI development and deployment infrastructure.

[253] Sanjam Garg et al., 'Experimenting with Zero-Knowledge Proofs of Training', in Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23: ACM SIGSAC Conference on Computer and Communications Security, Copenhagen Denmark: ACM, 2023), 1880–94, https://doi.org/10.1145/3576915.3623202; Haochen Sun, Jason Li, and Hongyang Zhang, 'zkLLM: Zero Knowledge Proofs for Large Language Models' (arXiv, 24 April 2024), https://doi.org/10.48550/arXiv.2404.16109; Tobin South et al., 'Verifiable Evaluations of Machine Learning Models Using zkSNARKs' (arXiv, 22 May 2024), https://doi.org/10.48550/arXiv.2402.02675; Suppakit Waiwitlikhit et al., 'Trustless Audits without Revealing Data or Models' (arXiv, 6 April 2024), https://doi.org/10.48550/arXiv.2404.04500.

[254] If progress on zero-knowledge proofs is dramatic, they might rapidly become the most practical approach for verifying many AI-related questions, thus supplanting many of the conclusions of this report.

Coming back to the topic of confidential computing for the privacy-preserving evaluation of digital objects, there are five non-trivial political and technical challenges that must be navigated. First, the location of new verification-specialized data centers could be a politically salient issue, potentially requiring a nuanced and multi-pronged solution (see Section 3.6 below). Second, algorithm code will also need to be verified, ideally in a way that is fully automated, since problematic algorithm code could allow for substantial circumventions of regulations (see Appendix E). Third, suites of evaluations need to be developed to address the multifaceted problems inherent in the assessment of digital objects such as model inputs (data, hyperparameters, and algorithms), model training transcripts,[255] models themselves, inference plans and so on (see Appendix G). This ecosystem is young and rapidly developing, but it must develop enormously if states are to depend on it for high-stakes deals.[256] Fourth, cyber attacks on the confidential computing stack might be possible, and this possibility deserves further scrutiny.

Fifth, physical access to the hardware might allow a state to violate confidential computing, including by potentially exfiltrating sensitive data or adding code that changes the behavior of the system.[257] One way to guard against many of these attacks is by monitoring the hardware stacks, an issue explored in Sections 2.5.2.1 and 2.5.2.3. More speculatively, proposals exist for tamper-resistant enclosures that are robust enough that they could be employed without external hardware monitoring. One specific approach, described in Petrie et al. 2024, combines computational and tamper-resistant elements into “Flexible Hardware-Enabled Guarantee” (flexHEG) mechanisms.[258] While the specific governance and verification recipes described there differ substantially from those in this report, their work provided multiple dimensions of inspiration.

Finally, there is the question of whether humans should also be in the loop or potentially serve as a point of escalation when automated evaluations indicate that something is amiss. Human-directed assessment of digital objects might be useful for a number of reasons. For example, human assessors might 1) act as a flexible stop-gap while automated evaluations are still being developed, 2) serve as a final source of overall judgment that goes beyond the narrow checks that (small) evaluation systems are capable of, 3) serve as a more serious “red team” of the assessment results as informed by the context, and 4) serve as an escalation process when automated evaluations indicate that something is amiss or the Verifier raises a challenge about a particular digital object. This kind of human involvement in the verification of sensitive digital assets is not without precedent. Presently, technology providers

[255] Shavit, ‘What Does It Take to Catch a Chinchilla?’; Choi, Shavit, and Duvenaud, ‘Tools for Verifying Neural Models’ Training Data’.
[256] ‘Evals’, 19 May 2023, https://github.com/openai/evals; ‘METR: Model Evaluation and Threat Research’, accessed 30 September 2024, https://metr.org/.
[257] Kulp et al., ‘Hardware-Enabled Governance Mechanisms’.
[258] Petrie et al., ‘Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees’.

such as Microsoft and Huawei go to significant lengths to provide governments with access to sensitive information such as source code via tightly controlled physical environments.[259]

However, the advantages of having humans in the loop are matched by very significant potential downsides. Most crucially, while it is possible to verifiably wipe the memory of a digital computer to preserve privacy, no such operation can be done with human assessors. Therefore, sensitive information seen by the human assessor may be revealed later to the Verifier or other actors, thus raising the Prover's concerns about the security of their information. Moreover, at least for the largest digital objects, such as training data sets or completed large models (which are gigabytes to petabytes in size), there is a real question about whether a human assessor can achieve insights that are not already provided by their tools. However, for assessments of much smaller objects, such as inference plans (see Section 4.5.2.2.1), humans may indeed have the cognitive capacity to meaningfully engage with the content and provide non-trivial insights into its logic.

How might humans be included in the loop in a relatively privacy-preserving way? Here is one potential approach. The Prover and Verifier set up a neutrally located data center for which they both verify and monitor the hardware, personnel, and physical security (see Section 3.6.1). For scenarios requiring extreme security, but lacking a need for rapid verification, this facility could also be air-gapped.[260] Within that facility, privacy-preserving evaluations could be run against digital objects, as described above. Additionally, human assessors could be provided access to specific parts of the assessed data or to the outputs of aggregate measures of the data. The access granted would need to be agreed to by the Prover via a confidential computing voting system. After conducting their evaluations as agreed by the Prover and Verifier, the assessor would be able to reliably transmit one bit per judgment to the Verifier and Prover.[261] In so doing, these assessors in their highly controlled context would perform the trusted-third-party equivalent of a zero-knowledge proof—wherein the Verifier would learn *only* whether the provided information was in compliance.

Following their assessment work, human assessors might need to be subject to strict physical and digital controls depending on the sensitivity of the information they were provided access to. Moderate intensity examples of such controls might include a ban on working for any state or key AI firms for a certain number of years. Extreme intensity examples might include living and working in air-gapped locations (verified by the Prover) for years. The

[259] 'Transparency Centers', Microsoft, 29 October 2024, https://learn.microsoft.com/en-us/security/engineering/contenttransparencycenters; 'Huawei Cyber Security Evaluation Centre (HCSEC) Oversight Board Annual Report 2021' (Huawei Cyber Security Evaluation Centre Oversight Board, 20 July 2021), https://assets.publishing.service.gov.uk/media/60f6b6be8fa8f50c7a1b9ffd/2021_HCSEC_OB_REPORT_FINAL__1_.pdf.

[260] Air gapping does introduce some questions about how mutual monitoring of the facility would work, since information is needed in order to monitor. If monitoring systems themselves must be air gapped, this raises the question of how these monitoring systems can in turn be monitored to ensure that neither party has circumvented them. Data centers which are wired can directly address these challenges by designating which kinds of outbound information flow are permitted and carefully controlling them (presuming that one or both actors might try to subvert them).

[261] Precisely how this bit would be transferred is a question for future work. Redundant mechanisms might be advisable, including even some extremely low-tech options like flags which can be visually monitored from far outside the facility.

extreme end of this spectrum appears unworkable in most potential scenarios, but might be a contingency reserved for high-stakes scenarios which cannot be resolved in any other way. If the Verifier is challenging the Prover about their compliance in a specific respect, but no available privacy-preserving evaluation is capable of providing the assurance needed, then one or more human assessors could be called in to provide the needed transparency while also protecting the security of the Prover. The substantial discomfort they might endure by being segregated from most of society for a time might be a price they and their state are willing to pay in order to avoid a cascading end to an important governance agreement.[262]

Regardless of whether humans are in the loop or not, this scheme allows the Prover to control what they reveal to the Verifier's agents (automated or human) and to the Verifier themselves—thus preserving the Prover's security. Equally, the Verifier is able to make judgments about the risk they perceive from the digital objects being assessed even if they cannot directly see any of the object's details, thus allowing them to achieve sufficient transparency to know whether the Prover's declarations with regards to these objects are correct and complete. If the Prover attempts to hide or exclude some of the data, the Verifier has many tools for noticing. If the Verifier attempts to exfiltrate secrets via their evaluations processes, the Prover can defect such efforts.[263] While it should be expected that states experience some degree of information leakage via other channels such as state intelligence agencies or open source intelligence, the verification mechanism described here has the potential to be very robust against accidental or inadvertent revelation of sensitive information.

3.6 Location and physical control of crucial hardware

The physical locations of key parts of the verification apparatus can have political ramifications. States may trust hardware located in their own territory far more than hardware located elsewhere. Furthermore, as they consider any potential agreement, states will also consider what happens if one of the parties chooses to exit that agreement, which could allow hardware to fall into the hands of the hosting state.

Regardless of where hardware is located, it could in theory be placed under *local* control of any shape. For example, embassies are typically physically defended by the states they belong to, not the states in which they are embedded. Such local control is possible for hardware crucial to verification, but there are important challenges. First, the host state certainly retains the ability to seize that location by force if needed. All verification schemes must assume that such an exit from the agreement is possible even if it is deemed unlikely. Second, locating a facility within a host state can give that host state both special access to that facility and (consequently) special reassurances that it is not being used against their wishes.

[262] One way to slightly reduce the burden of this approach would be to have assessors look at high-sensitivity information early in their tenure as an assessor, and then low-stakes information afterwards. This would allow them to be productively engaged in important work while they are under travel limitations.

[263] Recall that code is mutually inspected before running. Portions of the code and data that are hidden via confidential computing techniques can themselves be subject to evaluations (and perhaps even an escalation to a human assessor) by the counterparty—thus making it much more difficult for nefarious code to reach its target. Air gapped facilities are even more robust against this kind of attack.

Even if local control is maintained, the host state would be able to limit which equipment and personnel can access the site, and otherwise monitor signals that emerge from that facility including Internet exchanges. Having a facility within their own territory provides that state with additional reassurance that their data is not being stolen or the integrity of their computations compromised without their knowledge. Equally, however, the counterparty would perceive the same problem in reverse, as their sensitive data would be at greater risk of theft and their computations at greater risk of being manipulated. These concerns are not necessarily equal in gravity, as for example when a Prover attempts to demonstrate that their highly sensitive (e.g., military) model is compliant with relatively generic regulations. However, both concerns can be intense, as in the case where a Prover is proving their model compliant while the Verifier is using a variety of costly and sensitive evaluations to determine the model's compliance (see also Appendix C.6).

Addressing challenges like these will require political nuance beyond the scope of this report. Nonetheless, one broadly defined approach will be described in some detail below: the neutral mutually verified data center.

3.6.1 Neutral mutually verified data center

A neutral mutually verified data center—a *neutral data center*—is a data center that is under neutral institutional control (e.g., a third-party state or an international institution) and which is verified and monitored by both the Prover and Verifier (see Section 2.5.2). This section briefly explores the ways in which such a data center could be physically controlled, the purposes for which it might be used, and the overall feasibility of this approach.

A neutral data center is presumed to be under mutual control as well as under mutual verification. The Prover and Verifier can *both* exert physical control over the facility through a combination of cooperation and layered security checks. In normal operation, nothing gets in or out without them both agreeing. A second potential dimension of neutrality is the host state, which could be a state that is not strongly aligned with either the Prover nor the state or states that form the Verifier. Combining both mutual local control with a relatively neutral host state is desirable for neutral data centers that perform sensitive verification operations.

In the context of this report, the primary purpose for a neutral data center would be verification operations. As noted in the previous section, the Prover and Verifier both seek to protect their own sensitive data during verification operations and thus each would prefer to have more control over the verification data center's context. The additional certainty that one party might gain from a location change to their own territory might be counterbalanced with an *increased* perception of risk for the other party. Both sides might also fear a physical attack on the data center that could reveal some highly sensitive data—and thus they might each reserve the right to unilaterally wipe or even destroy all memory-capable computational devices in the facility (see Section 2.5.1). Given the extreme sensitivity of the operations that would be undertaken by a verification facility, choosing its location might be a key part of the negotiations for a deal. From the vantage point of this analysis, it seems like hosting verification facilities within a neutral state might be one of the best overall options.[264]

Neutral data centers might also be employed for much more ambitious computational workloads such as AI development or inference. There are at least four salient advantages of pooling large amounts of computational resources into neutral data centers: First, since verification can be run on adjacent hardware within the same data center, digital objects never need to leave the data center in order to be verified—thus reducing the complexity of the overall system as well as reducing the cyber attack surface. Second, neutral data centers are likely to be somewhat more trustworthy for the Verifier compared with Prover-run data centers, even if the Prover tries quite hard to demonstrate their compliance with their own data centers. Third, larger neutral data centers likely mean fewer total data centers will need to be monitored and controlled, thus reducing the overall cost of the verification system. Fourth, neutral data centers that would be placed at risk if the agreement ended are also a credible commitment mechanism for the Prover. By placing some of their compute in a neutral data center, they can make it infeasible for them to exit the agreement while retaining all of their compute. Therefore, exiting the agreement will come with additional costs. As the politics of AI governance evolve, it might be politically advantageous for participating states to gradually (and in lockstep) increase the amount of their compute that is located in neutral territory so that they can iteratively reassure each other about their commitments while also making the agreement more robust over time.

Construction of these kinds of facilities could plausibly be started immediately.[265] As of this writing, the primary technical barriers to building such a facility are security issues with no publicly-known robust solutions.[266] Mutual verification of such a facility would make these problems more difficult. Neither of these challenges should be underestimated and neither should be presumed to be impossible. Given the specialized nature of the narrow verification-only neutral data centers described above, it is entirely possible that the facility could be largely (or fully) air-gapped along with all of the tools and personnel employed to continuously verify its integrity. Neutral data centers play a crucial role in several of the verification approaches laid out in this report. For this reason, their development (or the potential discovery of their infeasibility) is a major crux for the future of AI verification.

[264] It should be noted that other than the security concerns regarding the information being processed in the neutral data center, there is no reason to believe that the neutral data center provides significant diplomatic or military power to the host state. The verification schemes outlined in this report hinge on the provision of trustworthy neutral compute, but that neutral compute provision does not provide much—if any—power to the host state. This concern would be mitigated even more fully if a set of several neutral facilities were set up in different countries, thus disallowing any one of them from using the threat of shutdown as a way to extract concessions from other actors who need this service for other reasons.

[265] Repurposing an existing facility might also be possible, though verifying every inch of an existing building might consume some of the benefits of avoiding having to build a new one.

[266] Sella Nevo et al., 'Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models' (RAND Corporation, 30 May 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html.

3.6.2 Licensing system locations

If licensing systems are chosen as one of the components of a verification regime, there are also political dimensions to the technical options available.[267] There are three major options for licensing system locations: Prover-controlled, Verifier-controlled, and distributed. Each has a set of technical and political factors shaping its desirability.

If a licensing system is controlled by the Prover and verified by the Verifier,[268] the Prover could exit the agreement while retaining full control of their hardware. As noted above for neutral data centers, being able to exit the agreement costlessly might be politically desirable for some reasons, but it also means that any agreement of that kind is less robust to shocks or changes unless other enforcement provisions are made. Note that it might be possible to obfuscate the location of the licensing system or systems to allow the Prover some assurance that such systems cannot be targeted with physical attacks.[269] Overall, this approach places a lot of power in the Prover's hands, but can still allow the Verifier to be confident that the Prover is abiding by the agreement.

Verifier-controlled licensing systems have very different implied politics. If the Prover exits the agreement, they will not be able to use the licensing system to enable their hardware. Depending on the implementation of the licensing systems, the Prover's hardware might be extremely difficult for them to re-enable in this scenario.[270] Verifier-controlled licensing systems thus might raise the cost of exiting the agreement for the Prover. Equally, however, it does imply that the Verifier might have some arbitrary political control over the Prover's ability to use some of their own hardware. This could be useful as a credible commitment, especially if conducted mutually, but it does look like a vulnerability at first glance, making the politics of this scheme more challenging. Mutual vulnerability is a centerpiece of mutual deterrence in nuclear strategy, so it is plausible that something similar may be done for licensing. In some ways, licensing locations can mirror the politics of large neutral data centers located outside of a state's territory (see Section 3.6.1).

Distributed licensing systems can employ cryptographic techniques such as secure multiparty communication or secret sharing[271] to create a licensing system of any desired political

[267] This section uses the term licensing to refer to hardware licensing as described in Section 2.2.4.5.

[268] This scheme presumes that the Verifier has credible information about the licensing system such that they are robustly sure that they are seeing every license issued. This could for example use a digital perimeter (see Section 2.5.2.4). All of these licensing approaches also require that the Verifier is sure that the target hardware in question cannot operate secretly without a license

[269] See Section 2.5.4.2.

[270] Simple license-driven locks on pod or rack-level enclosures might be easy to retool into other configurations. On-chip licensing schemes might be extremely difficult to circumvent depending on their design and maturity. See Aarne, Fist, and Withers (2024).

[271] Yehuda Lindell, 'Secure Multiparty Computation', Communications of the ACM 64, no. 1 (January 2021): 86–96, https://doi.org/10.1145/3387108; Amos Beimel, 'Secret-Sharing Schemes: A Survey', in Coding and Cryptology, ed. Yeow Meng Chee et al., vol. 6639, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer Berlin Heidelberg, 2011), 11–46, https://doi.org/10.1007/978-3-642-20901-7_2.

shape.[272] License-granting powers can be spread among a number of different parties, who must then work together to provide licenses. According to some sort of voting rules, these parties can choose together whether to grant a license—and presumably no single party can provide a license alone. Depending on the protocol, all stakeholders might have a credible reassurance that no valid licenses have been generated without their active participation (e.g., consensus). A consensus algorithm removes the sensitivity associated with a single location or license provider, since all stakeholders must have signed in order for a license to be sent. The challenge with a consensus algorithm, however, is that all key holders are implicitly veto holders, and every location that they hold their keys becomes a security liability for the Prover.[273] Therefore, non-consensus algorithms must also be considered seriously.[274] Other than the shift from physical to cryptographic licensing schemes and the implied politics therein, the politics of a distributed licensing system resemble those of the Verifier-controlled licensing schemes described above, with the Prover potentially facing significant costs for exiting an agreement. The key advantages of a distributed licensing approach are its up-front flexibility, its lack of need for a single centralized licensing system which could be a security liability, and its potential for change as political challenges evolve.

3.7 What kind of institution should oversee verification?

3.7.1 Direct vs indirect verification

Various institutional frameworks are possible for AI governance and verification. One dimension of note is whether the verification is done directly or indirectly via other institutions. Direct verification is when a state or a centralized international institution directly verifies the detailed behavior of another state.[275] This can be contrasted with indirect verification, where an additional layer of regulatory apparatus is allowed to serve as an intermediary. Typically, such intermediary institutions exist at the domestic level.[276] For example, a **jurisdictional certification approach** to AI governance would enable an international organization to judge whether (and how completely) states are embedding international AI governance standards into domestic law—an approach that has analogues in the international governance of civil-

[272] If one or more mutually controlled locations are still needed, keys can also be managed via cyber-physical processes such as key-control ceremonies. 'The DNSSEC Root Signing Ceremony', CloudFlare, accessed 2 October 2024, https://www.cloudflare.com/dns/dnssec/root-signing-ceremony/.

[273] If there are five keys stored in five locations, all of which are needed for the Prover to receive licenses, then the physical destruction of any of them would disallow all future licenses under a consensus algorithm.

[274] This report will not explore this challenge further, but will note in passing that cryptographic schemes are widely available for managing distributed keys and voting rights in nuanced ways.

[275] Even though international institutions are created by states, they are still capable of direct verification. For example, if an international verification organization had the mandate, personnel, and authority to directly inspect activities for compliance and continuously report on that compliance, it would certainly be conducting "direct" verification according to the definition provided here since it would have immediate access to the personnel, equipment, and activities needed to assess compliance. By contrast, any institution which is assessing compliance through proxies such as domestic regulatory agencies is indirect.

[276] One salient example of such an institutional type is the currently ongoing creation of national AI Safety Institutes in different countries.

ian aviation, maritime traffic, and finance.[277] Alternatively, a third party state can be brought in to perform the analysis. The **peer certification approach** allows states to send groups of experts to inspect each other's compliance with the agreement.[278] Direct verification is likely to be more desirable to Verifiers because it provides more reliable information. Correspondingly, Provers are likely to prefer indirect verification, which can allow them to employ information barriers to protect their most prized secrets.[279] While the question of direct vs indirect verification will certainly be important for future discussions about AI governance institutions, it will not be explored further here. The text in the remainder of this report implies that verification will be undertaken directly by a Verifier, but this should not be taken as a stance on this political question.

3.7.2 Is an international institution desirable?

Depending on the agreement, an international institution might be desirable for managing processes and pooling resources, but it may also not be needed. Institutions are particularly valuable when making agreements with complicated cooperative protocols among three or more states. A related, verification-like challenge is that of making international institutions legible enough to be trusted by their member states—but not too leaky to accomplish sensitive tasks.[280] This problem overlaps with the goals of this report, but compared with AI verification, the design and management of institutions is a well-established field and will therefore not be discussed further here. Overall, this report does not take a stance on whether an international institution is desirable for implementing the agreements described below.

[277] Trager et al., 'International Governance of Civilian AI'.

[278] One example of this approach is the mutual evaluations system within the Financial Action Task Force. 'Mutual Evaluations', Financial Action Task Force, accessed 13 July 2023, https://www.fatf-gafi.org/en/topics/mutual-evaluations.html.

[279] For example, see Section 'Mitigating Proliferation Dangers from Governance Processes' in Trager et al., 'International Governance of Civilian AI'.

[280] This is a perennial debate regarding international organizations in sensitive domains, such as the IAEA. Robert L. Brown and Jeffrey M. Kaplow, 'Talking Peace, Making Weapons: IAEA Technical Cooperation and Nuclear Proliferation', Journal of Conflict Resolution 58, no. 3 (1 April 2014): 402–28, https://doi.org/10.1177/0022002713509052; Nicholas L. Miller, 'Why Nuclear Energy Programs Rarely Lead to Proliferation', International Security 42, no. 2 (1 November 2017): 40–77, https://doi.org/10.1162/ISEC_a_00293; Rebecca Davis Gibbons, 'Supply to Deny: The Benefits of Nuclear Assistance for Nuclear Nonproliferation', Journal of Global Security Studies 5, no. 2 (1 April 2020): 282–98, https://doi.org/10.1093/jogss/ogz059.

4 International AI governance agreements and their verification requirements

This section provides an overview of possible international agreements relating to AI and their verification requirements.

The specific agreements within each category are deliberately kept artificially narrow for the purpose of examining the verification requirement of each component. These should be regarded as ideal types, not full proposals for real-world agreements. Real agreements would likely combine multiple approaches and perhaps even agreement types or categories. In this sense, this exercise should be seen as mapping out the verification requirements for some of the different components which may be mixed together into future agreements.

For some of these agreement types, it is possible to imagine variants that include some states but not others. For example, agreements with fewer states would be regional or "minilateral" while agreements with many states may be reasonably described as multilateral or even global. For non-global agreements, verification protocols may have to take into account the existence of non-participating states, since those states may have capabilities that are relevant to the verification being examined. In contrast, a global agreement may allow us to assume that actors with a certain level of capabilities do not exist outside of the agreement.

Due to limited space, only a very small number of verification mechanisms will be listed in each section below. These were chosen because they appear to be a compelling combination of a) effective, b) technologically mature, and c) politically viable. If increased confidence in compliance is desired, the methods described in the following subsections can often be combined or implemented in parallel (see Section 3.3 and Appendix C.5).

4.1 Transfer knowledge

For commercial, economic, or political reasons, states might seek to transfer knowledge internationally.[281] In the domain of AI, knowledge transfers can be accompanied by concerns about the proliferation of capabilities, security, intellectual property,[282] and economic or military competitiveness. The kinds of "knowledge" discussed here are non-trivial to verify and non-public.[283] Either the sending or the receiving state may want to verify knowledge transfers. The receiving state might want to confirm that the information is *authentic and*

[281] A related challenge is that of transferring physical resources. See Section 4.2.

[282] Given the proprietary nature of many AI development processes, agreements must carefully outline intellectual property (IP) protections. This can include defining the scope of shared knowledge—such as focusing on non-proprietary techniques or general methods—ensuring commercially sensitive aspects are not exposed. These protections give sending states the assurance that their IP rights are safeguarded while supporting meaningful knowledge transfer.

[283] See Section 1.1 for more on why this report focuses on certain kinds of agreements.

Table 4.1: Agreements examined in this report.

Agreement family	Agreement types	Variants
Transfer knowledge	Share research	
	Share knowledge of AI risks and opportunities	
	Share AI development knowledge	
	Share safety-enhancing technologies	
Transfer resources	Transfer development resources	Share AI-specialized chips Share access to AI-specialized compute Training programs for AI professionals
	Provide access to AI systems	Transfer completed models Provide API access
	Share benefits	Cash transfers Deploy AI-enabled devices as aid Transfer AI-enabled devices
Pool resources	Pool resources toward an international goal	
	Pool resources toward defensive AIs	
	Pool resources toward transformative AI	
	Pursue systemically risky AI only in a singular project	
Prepare for emergencies	Computational emergency detection and response	
Regulate	Regulate AI development	Regulate data center-based AI development Regulate fine-tuning and online learning
	Regulate AI deployment	Regulating data center inference Regulating sensitive mobile AI-enabled devices

Agreement family	Example agreement	Verifiability if implemented today	Verifiability presuming five years of serious effort
Transfer knowledge	Share knowledge of AI risks	Yes, with political limitations*	Yes, with political limitations*
Transfer resources	Share AI-specialized chips	Yes, with political limitations*	Yes, with political limitations*
Pool resources	Pool resources toward international goal	Yes	Yes
Prepare for emergencies	Computational emergency detection and repsonse	No	Maybe
Regulate	Regulate data center computations	No	Yes
	Regulate AI-enabled weapons	Very limited	Limited

*The sending state must deem the risks of knowledge or resource misuse to be tolerable.

Figure 4.1: The families of international agreements examined in this report and their estimated verifiability.

complete.[284] By "authentic", we mean true to the extent known by the sending institution and produced or summarized via an agreed-upon process. By "complete", we mean that no relevant knowledge is being withheld in violation of the agreement. For its part, the sending state might want to confirm that the information is protected appropriately by the receiving state (i.e., not spread or resold beyond the rules of the agreement). The remainder of this subsection will describe some examples of knowledge transfer agreements and then explore the related verification challenges facing the sending and receiving states.

[284] Related concepts in civilian nuclear verification are "correctness" and "completeness". This report uses "authenticity" instead of "correctness", since some agreements may be over information that is not known to be correct, but is produced via authentic processes. For the IAEA definitions, see Laura Rockwood, 'IAEA Safeguards: Correctness and Completeness of States' Safeguards Declarations', in Nuclear Law: The Global Debate (The Hague: T.M.C. Asser Press, 2022), 205–22, https://doi.org/10.1007/978-94-6265-495-2_10.

4.1.1 Examples

4.1.1.1 Share research

Particular kinds of AI-related research findings could be verifiably shared among relevant institutions such as AI Safety Institutes or international regulators. For example, an agreement might stipulate that all research of a particular category be shared.[285] Verification of this kind may be desirable for an international regulatory institution or for multilateral collaborations that include domestic AI regulators or research efforts.[286]

4.1.1.2 Share knowledge of AI risks and opportunities

Industry-wide data on AI risks and opportunities could be verifiably sent to states or to an international organization tasked with summarizing the industry and the underlying science.[287] For example, states might commit to sharing information on newly discovered AI-related hazards. The analogy of the World Health Organization's International Health Regulations indicates that such agreements are desirable and possible, but that same analogy is also a cautionary tale, as states have failed to consistently report information during major health incidents.[288]

4.1.1.3 Share AI development knowledge

States verifiably share knowledge about how to develop AI, with the goal of enabling better or faster AI development.[289] While this is certainly related to the research-sharing described above, this domain also includes non-research knowledge such as the technical and practical expertise needed to effectively develop and deploy AI systems.[290]

4.1.1.4 Share safety-enhancing technologies

States may choose to systematically share certain kinds of safety-enhancing technologies with each other. One historical example of this was the United States choosing to share nu-

[285] As discussed further in the subsections below, demonstrating *completeness* in such an agreement might be extremely difficult. While it is comparatively easy to send and verify research, it might be impossible to know whether other (secret and undeclared) research had taken place.

[286] Lewis Ho et al., 'International Institutions for Advanced AI' (arXiv, 11 July 2023), http://arxiv.org/abs/2307.04699; Marta Ziosi et al., 'AISIs' Roles in Domestic and International Governance', 2024.

[287] Ho et al., 'International Institutions for Advanced AI'; Hadrien Pouget et al., 'The Future of International Scientific Assessments of AI's Risks' (Oxford Martin AI Governance Initiative, August 2024).

[288] Raphael Lencucha and Shashika Bandara, 'Trust, Risk, and the Challenge of Information Sharing during a Health Emergency', Globalization and Health 17, no. 1 (18 February 2021): 21, https://doi.org/10.1186/s12992-021-00673-9.

[289] One analogy to this kind of effort is cyber capacity building, which seeks to augment the cybersecurity abilities of other states. Andrea Calderaro and Anthony J. S. Craig, 'Transnational Governance of Cybersecurity: Policy Challenges and Global Inequalities in Cyber Capacity Building', Third World Quarterly 41, no. 6 (2 June 2020): 917–38, https://doi.org/10.1080/01436597.2020.1729729.

[290] Such knowledge transfer could also provide information about key infrastructure for AI development. Information flows could include specific modalities such as confidential technical reports, infrastructure design documents, and site visits.

clear permissive action link technology with the Soviet Union, in an attempt to help them centralize control of their nuclear weapons.[291]

4.1.2 Verify that knowledge shared is authentic and complete

If the sending state commits to sharing authentic and complete knowledge with the receiving state, the receiving state faces the challenge of verifying these aspects of the knowledge.

4.1.2.1 Verifying information authenticity

There are at least four ways to verify information authenticity:

1. **Verify directly**: Test using replication or other computational checks (e.g., replicating technical results or following proofs). This is robust for specific domains, but not applicable to all forms of knowledge. Some knowledge may also be too costly to verify directly even if it can be done in theory.

2. **Verify via access to key personnel**: Sending states provide the research of interest by providing access to personnel who can share knowledge about that research. This is workable in low-stakes environments but very difficult to make credible in high-stakes or secretive environments (see Section 2.1.2).

3. **Verify via process and access**: The process by which the knowledge is provided is credible. For example, verifiable claims might be made via access to specific digital infrastructures (see Section 2.4.1.2).

4. **Verify via other methods**: States may have other ways to verify the authenticity of knowledge, including comparisons with the findings of their own state's intelligence services or comparisons with the knowledge of other trusted states.

Not all information can be verified in these ways. For example, when the U.S. was considering sharing sensitive nuclear control technology with Pakistan (to help secure Pakistan's arsenal), the Pakistani government worried that the transferred technologies might allow the U.S. to prevent them from using their own nuclear weapons—and no approach available at the time was able to resolve these uncertainties.[292]

4.1.2.2 Verifying information completeness

To verify that information is complete, Verifiers must acquire evidence that nothing important was withheld by the sending state. Depending on the domain, this could be an extremely challenging goal and may even be impossible (see Section 1.5.2).

There are at least two general ways to accomplish this in practice:

[291] Jeffrey Ding, 'Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies', European Journal of International Relations, 27 April 2024, 13540661241246622, https://doi.org/10.1177/13540661241246622.

[292] Jeffrey Ding, 'Keep Your Enemies Safer'.

1. The sending state provides evidence that they are providing all relevant information. This might be accomplished by providing that all potential sources of this kind of information are covered by the information transmission mechanism.[293]
2. The sending state might provide different kinds of parallel evidence that employ different approaches and data sources, thus allowing the receiving state to compare the evidence. The presumption is that while the sending state might be able to easily manipulate one kind of information process, they might have difficulty manipulating many such processes. See also Appendix C.5.

4.1.3 Verify that transferred knowledge is safeguarded appropriately

Some agreements may require that the receiving state must protect the knowledge received with verifiable controls on personnel, digital infrastructure, or AI infrastructure (see Sections 2.1.1, 2.4.1.1 and 2.5.2). These controls must address at least three problems:

1. They must ensure that the transferred knowledge is not re-transferred outside of the bounds of the agreement without the original sender's consent.
2. They must ensure that the transferred knowledge is not diverted to forbidden uses within the receiving country. For example, the sending state might require that the receiving state not use transferred resources for military investments.
3. They must ensure that the transferred knowledge remains satisfactorily protected against theft by other actors.

4.2 Transfer resources

States may also seek to transfer resources other than knowledge across borders. Once again, they may be motivated by commercial, economic, or political incentives. Reasons for resource transfers might fall anywhere on a spectrum from purely market-driven to purely politically motivated. Non-commercial resource sharing agreements could be undertaken for a variety of reasons.[294] These arrangements can serve various purposes: supporting inclusive economic growth, fostering technological self-determination in developing countries, and advancing the political objectives of sending states (for example by strengthening strategic partnerships or encouraging the adoption of international agreements).[295]

[293] For example, all research of a given type is conducted in specific facilities by a limited number of people—and a transmission infrastructure provides credible assurances that all of the relevant information is being transmitted.

[294] Claire Dennis et al., 'Options and Motivations for International AI Benefit Sharing' (Centre for the Governance of AI, 2025).

[295] Several international agreements include resource-sharing mechanisms as a key component. For example, Article IV of the Nuclear Non-Proliferation Treaty promotes nuclear cooperation for peaceful purposes, which the International Atomic Energy Agency pursues through its Technical Cooperation Programme.

There are two major verification challenges with resource sharing: 1) the verifiable transfer of resources from the sending state to the receiving state and 2) the subsequent protection of such resources to prevent processes such as proliferation, diversion, or resale. The following subsections explore the verifiability of these agreements depending on the type of resource transferred. We find that verifying secure transfer is far more intricate for some assets (e.g., pretrained models) than others (e.g., financial proceeds).

4.2.1 Transfer development resources

States may want to verifiably share with (or sell to) other states a portion of their AI development capabilities. Such transactions would allow a broader group of states to construct their own AI systems. The subsections below explore three different development resources that could be transferred.

4.2.1.1 Share AI-specialized chips

An agreement could facilitate the transfer of some quantity of AI-specialized chips. In order to assure the receiving state that the correct proportion of chips is being shared, and that they have not been tampered with, the sharing agreement may require some controls on the chip supply chain, perhaps also including use of a chip registry.[296] The receiving state will have to verifiably control the downstream uses of these chips to guard against resale or unapproved transfers. Such controls would apply to data centers and potentially domestic regulators (see Section 2.5.2.1).

4.2.1.2 Share access to AI-specialized compute

An agreement could facilitate *remote* (cloud-based) access to AI-specialized compute resources. An agreement of this kind might be very desirable for many potential receiving states, since AI-specialized data centers are somewhat rare and are heavily concentrated in a few countries. Meanwhile, sending states might desire an agreement of this kind so that they can economically benefit from exporting compute operations without losing physical and legal control of the hardware.[297]

Such access might be provided via sender-domiciled cloud computing intermediary institutions or via more direct access by receiver-employed institutions:

- **Access via receiver-controlled intermediary institution**: In this scenario, the sending state allows agents of the receiving state to have direct access to some AI-specialized compute, such as a data center or a portion of a data center. The receiving state might be allowed to use the infrastructure under controlled conditions. To ensure mutual security and compliance with the agreement, the sending state could implement ver-

[296] See Sections 2.2.2.1 and 2.2.2.2.

[297] For a related exploration of how some aspects of AI governance can be provided through cloud providers, see Lennart Heim et al., 'Governing Through The Cloud: The Intermediary Role Of Compute Providers In AI Regulation' (Oxford Martin AI Governance Initiative, March 2024).

ifiable controls over the infrastructure in question (see Section 2.5.2). Meanwhile, the recipient state would need to establish auditable systems to prevent the unauthorized transfer or resale of access (see Section 2.4.1.1). Personnel controls may also be necessary to ensure that only authorized institutions, such as government agencies or strategic industries, utilize this privileged compute access within the bounds of the agreement (see Section 2.1.1). The sending state may also need to be able to verify that the compute is being used in accordance with specific rules (see Section 4.5).

- **Access via cloud services**: The sending state provides access to its AI-specialized compute resources via the Internet, thus allowing the recipient to engage in AI activities without having their own on-premises hardware. The sending state could demonstrate that the hardware is secure in ways that disallow them from stealing information or downgrading the quality of computational resources provided (see Section 2.5.2). As noted above, the sending state may need to verify compute use (see Section 4.5) and the recipient may have to prove that they are appropriately protecting the resource.

In either case, the compute would still be located in the sending state, but it would be under different local management in the two scenarios. Given that the compute is located in the sending state, whether they are able to credibly assure receivers of ongoing access (i.e., that they won't be cut off for political reasons) is primarily a question of political signalling and political commitment, and thus outside the scope of this report.

4.2.1.3 Training programs for AI professionals

An agreement enables the training of AI professionals—including via exchanges. Receiving states would need to ensure that the sensitive knowledge received by their professionals is appropriately protected via verifiable controls on personnel and potentially also digital infrastructure (see Sections 2.1.1 and 2.4.1.1). This is related to the previous section on verifiably transferring knowledge (see Section 4.1), but refers to activities that go beyond merely transmitting and verifying data. Consider that training programs are somewhat likely to involve on-site activities within one or both states, with appropriate security controls in place. States may also confirm the authenticity of training in somewhat different ways, since the acquisition of skill might be somewhat more difficult to confirm than the authenticity and completeness of data as discussed above.

4.2.2 Provide access to AI systems

Whether or not development resources are shared, it is also possible to *share access to AI*, since access does not require ownership. Two kinds of access are explored: transfers of completed models and remote access via API.

4.2.2.1 Transfer completed models

An AI model can be created in one state and then transferred securely to another state. The sending state may have to provide verifiable information about the training process as well as potentially the fine-tuning processes undertaken to create the model (see Section 4.5.1). The receiving state would have to protect the model with verifiable controls on personnel, data centers, and institutional digital infrastructure—all of which might be facilitated through domestic regulation if private actors play a role in any of these processes (see Sections 2.1.1, 2.4.1.1 and 2.5.2).

4.2.2.2 Provide API access

Receiving states could be allowed to use particular APIs provided by a sending state. APIs (application programming interfaces) are Internet-enabled exchanges of information between a client and a server, all of which are encrypted in transit. APIs allow users to benefit from both models and AI-specialized compute that are physically located in other states. Such an access-sharing approach may be desirable, because the energy, engineering, and institutional requirements for AI-specialized data centers make it infeasible for most states to build them within the next several years, even if these facilities will only be used for inference and not the more expensive development steps.[298]

Guaranteeing reliable API access over time presents distinct political challenges, as the sending state retains full physical and legal control over the data centers and infrastructure. As noted above in Section 4.2.1.2, sending states may have to make significant political commitments in order for continuity of access to be regarded as credible. Receiving states might consider it dangerous to build valuable public or private infrastructures atop API access provided by a sending state which has not made serious commitments to ensuring that such access is ongoing.

A more technical question is that of verifying the integrity of the API. To verify that the API's behavior (e.g., model, performance, security) is being provided as per the agreement, the sending state can provide parallel streams of verifiable evidence from their hardware which can demonstrate that the hardware is in a compliant configuration, has not been altered by the sender's agents, and that no data is being copied without the receiving state's express permission (see Section 2.5.2 and its subsections, as well as Section 4.5 for verifiable rules about AI development and inference).

For their part, the receiving state may have to provide credible information about how they are ensuring that access to the API is not being re-sold or misused. This could include verification of personnel controls, domestic regulation, digital infrastructure, and data centers.[299]

[298] The politics of AI-specialized compute is also increasingly contested, thus making its acquisition more difficult for many states.

[299] Note that smaller or non-AI-specialized data centers can exist in the receiving state as part of more traditional digital infrastructure for a state.

4.2.3 Share benefits

States may also choose to verifiably share some of the *benefits* of their AI systems—including cash transfers or the productive use of AI-enabled devices (see Section 2.2.5).[300]

4.2.3.1 Cash transfers

An agreement could include monetary transfers between states. The sending state might agree to transfer money based on the financial performance of certain entities, which would require some verification of financial data.[301] The receiving state may not have to provide any verifiable assurances in this approach.

4.2.3.2 Deploy AI-enabled devices as aid

AI-enabled systems can be used to provide economic help to other states without transferring ownership of those systems. For example, future AI-enabled systems could be used for infrastructure development. Receiving states might require verifiable information about the devices themselves, and if they receive such information they may have to verifiably protect it via controls on personnel and digital infrastructure (see Sections 2.1.1, 2.4.1.1 and 4.5.2.3.6).

4.2.3.3 Transfer AI-enabled devices

A sending state can also transfer AI-enabled devices to a receiving state, with the receiving state guaranteeing that those devices will not be passed onwards to any other state.[302] Verification of such an agreement would be very similar to the agreement described above. The only major addition is that domestic regulation in the receiving state would have to verifiably demonstrate that they will retain ongoing control of the AI-enabled devices in a way that makes their redirection or theft unlikely.[303]

4.3 Pool resources

States may also choose to pool their resources into a cooperative project that is overseen by a new third-party institution. They may do this in pursuit of a wide variety of international goals, including contributing to the Sustainable Development Goals, building "defensive" AIs intended to protect humanity from AI misuse or accidents in the future,[304] and pursuing potentially dangerous research in a cooperative setting with international oversight. Pooling resources can allow costs to be shared among many states. Furthermore, pooling resources

[300] For a broader take on the concept of benefit sharing which overlaps with other agreement categories in this report, see Lennart Heim, 'AI Benefit Sharing Options', Lennart Heim (blog), 28 September 2024, https://blog.heim.xyz/ai-benefit-sharing-options/.

[301] Verification of financial data is not examined in this report.

[302] One precedent for this kind of agreement is re-export limitations for weapons.

[303] This might imply a form of indirect rather than direct verification. See Section 3.7.1.

[304] Yoshua Bengio, 'AI and Catastrophic Risk', Journal of Democracy, October 2023, https://www.journalofdemocracy.org/articles/ai-and-catastrophic-risk/.

can be politically useful as a signal of intent as well as a commitment mechanism.[305] The rest of this subsection describes some examples of this category of agreement before unpacking three components: internal regulation of the project, resource provision, and guardrails against resource redirection.

4.3.1 Examples

4.3.1.1 Pool resources toward an international goal

States can verifiably pool a portion of their resources toward an international goal such as achieving the Sustainable Development Goals. This kind of agreement overlaps heavily with the agreement types described above.

4.3.1.2 Pool resources toward defensive AIs

States can verifiably pool resources toward the research, development, and deployment of "defensive" AIs which can help humanity mitigate the danger of hypothetical uncontrolled or "rogue" AIs.[306] This proposal bears significant similarity, both in overall structure and its verification requirements, to proposals discussed below (see Sections 4.3.1.3 and 4.3.1.4). Presuming that defensive AIs may eventually be used in sensitive contexts such as cyberdefense, verifying the capabilities and activities of the institution(s) building them will be crucial. Potentially, many domains of verification will be relevant, with a particular emphasis on the verifiable control of the institution's personnel, digital infrastructure, data centers, training capabilities, and fine-tuning capabilities (see Sections 2.1.1, 2.4.1.1, 2.5.2 and 4.5).

4.3.1.3 Pool resources toward transformative AI

Two or more states could pool their resources with the intent of creating transformative AI—AI capable of transforming human society on the scale of the agricultural or industrial revolutions.[307] This project could therefore be one of the largest AI projects and it would be aimed at having major and broad effects on the world—as opposed to the more narrowly scoped projects aimed at specific international goals or creating defensive AIs.[308] While this proposal is not explored in greater depth here, the following sections explore categories of

[305] Resources placed into a pooled effort are resources that may be difficult to repurpose back to unilateral efforts. This can also be made part of a deliberate strategy to create increasingly robust agreements (see Section 3.6). Pooling of resources can make states less threatening to each other and also make verification of claims related to the state's remaining resources easier.

[306] Yoshua Bengio, 'AI and Catastrophic Risk', Journal of Democracy, October 2023, https://www.journalofdemocracy.org/articles/ai-and-catastrophic-risk/.

[307] Holden Karnofsky, 'Some Background on Our Views Regarding Advanced Artificial Intelligence', Open Philanthropy Project (Blog), Open Philanthropy Project, 2016, https://www.openphilanthropy.org/research/some-background-on-our-views-regarding-advanced-artificial-intelligence/.

[308] The governance issues of such a project go far beyond the scope of this report, since there is a real potential for such a project to dramatically affect humanity's future. It is worth noting however that the governance structures built for managing such a project internally might be very similar to the internal governance requirements of the singular project approach described below. In both approaches, tremendous emphasis must be placed on reliable governance as well as safe development and deployment.

verification that would be needed for such an agreement, including for both AI development and AI deployment (see Section 4.3.1.4 and Section 4.5).

4.3.1.4 Pursue systemically risky AI only in a singular project

States verifiably pursue *systemically risky AI*—AI with the potential of significantly affecting humanity as a whole (see Appendix H)—only in a singular cooperative project. States may have many reasons for building such an agreement, including avoiding war and minimizing the risk of the loss of human control (see Section 1.3). Presumably, the aim of a centralized project would be to avoid a costly race toward extremely powerful AI among states[309] while also increasing the safety, transparency, and political control of the development of powerful AI systems. Such an agreement would require the pooling of AI resources and an associated ban on this category of AI development by any other institution. Such an agreement would bear a resemblance to the "Pool resources toward transformative AI" concept described above, though coupled with a broad effort to regulate AI development widely to ensure that this project has no significant competitors (see Section 4.5 and Section 4.5.1.2.1 in particular). This proposal similarly has many political details and ramifications that go far beyond the scope of this report.

4.3.2 Components

Three components of these kinds of agreements relate to AI verification: regulating how AI is created or used within the project, how resources are provided to the project, and how resources are controlled to ensure that they are used for appropriate purposes. Regulation is discussed extensively below (Section 4.5). Resources can be provided in verifiable ways, including funding, chips (see Section 4.2.1.1), computing infrastructure access (see Section 4.2.1.2), API access (see Section 4.2.2.2) and personnel (see Section 2.1). Verification of the appropriate usage of those resources would require controls on information and resources as described earlier (see Sections 4.1 and 4.2).

4.4 Prepare for emergencies

Preparing for AI-related emergencies is too broad a topic to fully explore in this report. Given that AI might be used in any domain by any actor, there are a vast array of potential scenarios that a state might deem an emergency. In light of this breadth, this report examines the verifiability of the *computational* side of emergency response. While an AI-related emergency may have many effects on the world, fundamentally, AI depends on computations. Therefore, both the detection of and response to an AI emergency may be centered on AI computations and the hardware that enables them.

[309] Domains of AI that are not systemically dangerous might continue to be sites for intense competition.

4.4.1 Computational emergency detection and response

Preparing for the computational side of potential AI emergencies requires building verifiable systems which are capable of: a) detecting dangerous AI systems and AI-related computational events, b) alerting states and other stakeholders about the danger, and c) responding to the emergency.[310]

To detect dangerous AI systems or computational events, detection systems must have extensive access to the computations taking place. Such systems are highly analogous to the computational verification systems discussed below (Section 4.5), and that section should be consulted for details. Overall, verification of such systems may be politically possible depending on the sensitivity of the information that would be transmitted. As of this writing, detailed verification appears unworkable, but there is significant potential for highly verifiable systems within a few years.

Alerting states and other stakeholders about AI-related dangers requires that credible mechanisms exist to transmit crucial information to those stakeholders at the appropriate time. Such information transfer systems would also resemble the technical underpinnings of regulation as discussed in Section 4.5. As with all information transfers discussed in this report, questions of political feasibility are likely to hinge on the scope, limits, and risks of the information that could be provided via any proposed mechanism (see also Section 4.1).[311]

The prospect of responding to an AI emergency raises even more questions. Mechanisms that preserve state ability to act would face similar technical challenges as those outlined in the section on regulation below (Section 4.5). If such mechanisms are intended to be used on the international level (e.g., by a group of states via a protocol such as a vote), they would likely face extreme political challenges. For example, the same mechanisms that could pause or permanently disable AI systems exhibiting dangerous behavior in an emergency could potentially be misused by an adversary. An adversary might be able to unilaterally weaponize such mechanisms to gain an advantage in war or industrial competition. Equally, if small groups of states can veto the use of these mechanisms, there might be very few situations in which the mechanisms would actually be employed.[312] The potential for failures in both directions is real, and thus any attempt to build verifiable emergency response powers into an international agreement would have to grapple with questions like these. The potentially devastating security consequences of mistakes may make such mechanisms extremely chal-

[310] Andrea Miotti and Akash Wasil, 'Taking Control: Policies to Address Extinction Risks from Advanced AI' (arXiv, 31 October 2023), http://arxiv.org/abs/2310.20563.

[311] It should be noted that states cannot be expected to reliably report events to the international community, even if they are very important. For example, the Soviet Union attempted to keep the meltdown at Chernobyl secret. Olga Bertelsen, 'Secrecy and the Disinformation Campaign Surrounding Chernobyl', International Journal of Intelligence and CounterIntelligence 35, no. 2 (3 April 2022): 292–317, https://doi.org/10.1080/08850607.2021.2018262.

[312] One analogy is the authority of the UN Security Council to manage questions of war and peace and how action is rare due to the permanent members of the council holding a veto over all such decisions.

lenging to implement.[313] Mechanisms less suited to misuse might need to be found in order to address these concerns.

Our discussion of emergency response will be left underspecified in this report in favor of more emphasis on regulation (described below). The reader should understand that discussions about preparing for AI emergencies are in their infancy and thus should not take any of the claims provided here as definitive on the subject. The goal of this section was to outline some of the computational aspects of emergency preparation and highlight how these aspects significantly overlap with the verification of AI regulation.

4.5 Regulate

In this agreement type, states verifiably regulate one or more aspects of AI development or usage according to international standards. By attempting to discuss AI regulation broadly, this section casts a very wide net, including both the civilian and military spheres as well as the full spectrum of model sizes and uses. Many different governance goals and verification approaches are possible within this category, so the discussion in this report will be necessarily limited.

The vast potential range of potential rules and target models within this category also implies very different levels of sensitivity. Some domains might be relatively benign (e.g., commercial models trained on no particularly sensitive data) while others will be maximally sensitive (e.g., models created and used by state militaries or intelligence agencies). Overall, this discussion focuses on verification mechanisms that are potentially workable in the maximally sensitive cases, since generally the less sensitive cases are easier problems to solve.

The full space of potential regulations is beyond the scope of this report, so only a few examples will be provided. Other works have explored this space and this remains an area of active research.[314]

While regulations could refer to any number of technical systems, institutions, or individuals, this section will focus on regulations that place technical (digital) requirements on AI development and AI deployment. This focus is justified primarily by the observation that there is a well-established literature on how to verifiably regulate institutions and individuals, but little has yet been said about how to verifiably regulate AI development and deployment.

The following subsections will focus on hardware-centric verification of regulation rather than personnel-based methods. This is because hardware is better able to scrutinize the details of computational operations, and hardware-enabled verification is more scalable. To augment these approaches, personnel-based verification of AI regulation could be under-

[313] As with all mechanisms discussed in this report, the severity of imperfections will be weighed by political decision-makers against their reasons for considering such agreements in the first place. Even agreements with fairly severe issues might go ahead if the problems they solve are much larger or more likely than the problems they introduce.

[314] Markus Anderljung et al., 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (arXiv, 11 July 2023), http://arxiv.org/abs/2307.03718; David Manheim et al., 'The Necessity of AI Audit Standards Boards' (arXiv, 11 April 2024), https://doi.org/10.48550/arXiv.2404.13060.

taken in parallel. However, this faces a much more severe transparency-security tradeoff than the hardware mechanisms discussed here and may therefore be ill-suited for verifying regulatory efforts that involve sensitive information (see Section 2.1.2).

4.5.1 AI development

Regulating AI development means applying rules to the creation and modification of AI models. As noted in the introduction, model development is often split into two major phases in current development paradigms: training and fine-tuning.[315] Training typically employs large general datasets and large amounts of compute to produce a relatively general-purpose model. Fine-tuning is then employed to modify the model behavior in a more fine-grained way.

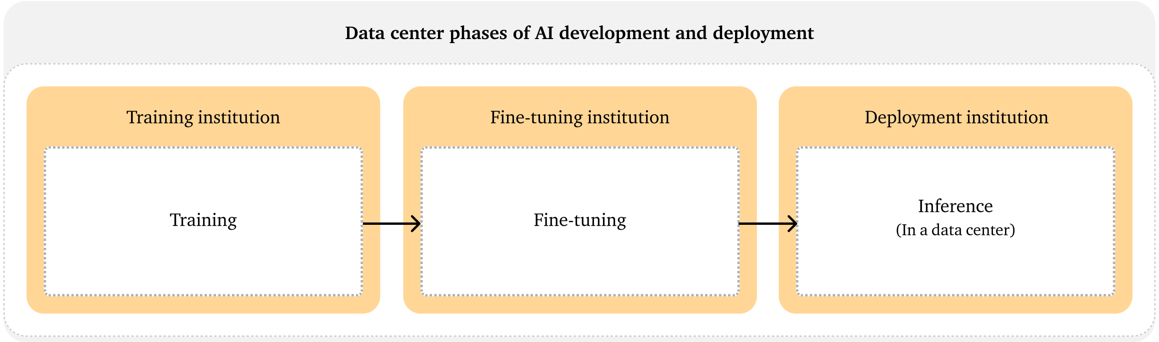

Figure 4.2: Schematic representation of the phases of data center-based AI development and deployment. Institutions are mentioned to highlight the fact that it is possible for different institutions to play different roles throughout the model lifecycle—even though today it is common for single institutions to play multiple (or all) roles in this chain.

4.5.1.1 Training verification concepts

4.5.1.1.1 Training plan

A training plan for a new AI model will include the training data,[316] algorithms, and hyperparameters—including the starting weights (if continuing a prior model run) or the verifiable source of randomness used for initializing the weights.[317] Possession of the training

[315] The terms used for the different phases of training are evolving. For a description of some recent changes, see Toby Ord, 'Inference Scaling Reshapes AI Governance', 12 February 2025, https://www.tobyord.com/writing/inference-scaling-reshapes-ai-governance.

[316] On the usefulness of regulating via data, see Ritwik Gupta et al., 'Data-Centric AI Governance: Addressing the Limitations of Model-Focused Policies' (arXiv, 25 September 2024), https://doi.org/10.48550/arXiv.2409.17216.

[317] A very similar concept to a training plan is discussed in Shavit (2023). The key difference being that a training plan is intended to be scrutinized before training commences, while a proof-of-training transcript is generated while training a model. Shavit, 'What Does It Take to Catch a Chinchilla?'

plan should allow an actor to (re)produce a model, with only relatively small discrepancies that might be ascribed to hardware noise.[318]

Incomplete training plans—such as plans that include data but not the algorithm or hyperparameters—might still be useful for regulation. This report will not explore all potential partial training plans and which kinds of verifiable regulatory commitments those plans might enable. Such an area deserves significant further attention, since if the information that must be exchanged for verification is less sensitive, that might make it much more likely that an agreement is politically acceptable to all sides. For the rest of this report, training plans are presumed to be complete.

4.5.1.1.2 Training transcript

A training transcript is the complete record of the training of a machine learning system, including training data, hyperparameters, and a record of the digital objects that constitute the emerging model and attest to its provenance. This includes model weights and weight shards, as well as potentially gradients or activations depending on the training technique.[319]

4.5.1.1.3 Proof of training

Building on the concepts developed by Shavit (2023), we define a "proof of training" as evidence that is sufficient to prove that a given model was generated using a given training transcript.[320] Provided with such a proof, the Verifier would then have extremely strong evidence that the model was indeed trained in the declared manner.[321]

4.5.1.2 Potential political goals

4.5.1.2.1 Inhibit or ban development of extremely large models

States may make an agreement which verifiably inhibits their ability to produce models above a certain size—measured via quantities such as compute budget or parameter

[318] Shavit 2023 and Choi, Shavid, and Duvenaud 2023 both grapple with this problem. As they note, in theory, it should be possible for the Prover and Verifier to create highly reproducible training techniques, although this may not currently be possible with leading edge AI training hardware. Whether or not highly reproducible training remains a serious technical difficulty—or source of significant training inefficiency—moving forward is a question that deserves urgent work.

[319] Shavit.

[320] Dami Choi, Yonadav Shavit, and David Duvenaud, 'Tools for Verifying Neural Models' Training Data' (arXiv, 2 July 2023), https://doi.org/10.48550/arXiv.2307.00682; Shavit, 'What Does It Take to Catch a Chinchilla?'

[321] Given enough data and compute, it is possible to replicate parts of the training run exactly, thus allowing for targeted or even very extensive examination of the training results. In the limit, this would be as computationally costly as building the model in the first place. Faster alternatives to such a process would therefore be desirable. See Shavit, 'What Does It Take to Catch a Chinchilla?' and 'AICert', Mithril Security, accessed 2 October 2024, https://www.mithrilsecurity.io/aicert.

count.[322,323] This policy goal is premised on the widespread acknowledgement that extremely large models are particularly likely to generate novel capabilities and risks.[324] Depending on the political goals, a continuum of potential agreements are available, ranging from blunt inhibition mechanisms up to fine-grained controls. Blunt mechanisms involve making broad changes to either compute chips or how they are physically arranged to make it more costly and slow to produce large models. By contrast, fine-grained controls are specific mechanisms that could prevent the production of extremely large models in a variety of ways, while having less of an effect on other uses of compute. The efficiency of most AI development would be better served by fine-grained rather than blunt mechanisms. Moreover, a true ban is likely impossible via blunt mechanisms, but is potentially workable via fine-grained controls.

The key advantage of blunt mechanisms is that the information they require to implement and verify is not sensitive—and thus these mechanisms are very likely to be deemed less revealing (or at least less risky) in terms of security details than more fine-grained mechanisms. For example, one speculative mechanism is to limit the admissible hardware configurations that states can employ via "fixed set" mechanisms—implemented via either mechanisms built into AI-specialized chips or monitored arrangements of hardware (this is discussed further later, in Section 4.5.1.3.1).[325] Such an approach would make large models somewhat harder to create, but would otherwise not provide the Verifier with any knowledge of what the Prover was doing.[326] However, this would potentially hamper many activities that are not within the purview of the policy goal. By contrast, a fine-grained ban on large models could be verified via a more complicated verification stack. This would require more information to be made available to automated evaluation systems installed by the Verifier and perhaps even to the Verifier's human agents (see Section 3.5). This more intrusive approach could enable a verifiable ban on non-compliant activities while allowing all other activities to proceed at full pace. The overall design of the verification system for inhibiting or ban-

[322] "Compute budget" refers to the total compute used for a project (e.g., the FLOPs—Floating point operations—used in training an AI model). A related use of this term elsewhere in the literature is the total compute capacity available to a given actor during a given period of time. Lennart Heim et al., 'Governing Through The Cloud', 2024.

[323] "Inhibit" in this sense is raising the cost, time, complexity, or other difficulties of creating a model beyond the threshold set in regulation. The relative scale of this inhibition is not explored in this report. See Scher and Thiergart (2024) for an exploration of how hardware arrangements could be used to impose very substantial costs (approximately one hundred times the cost) on non-compliant activities.

[324] Markus Anderljung et al., 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (arXiv, 11 July 2023), http://arxiv.org/abs/2307.03718; Girish Sastry et al., 'Computing Power and the Governance of Artificial Intelligence' (arXiv, 13 February 2024), http://arxiv.org/abs/2402.08797.

[325] More extreme actions in the blunt category are theoretically possible, including verifiably powering down (or even destroying) portions of compute. The available options of this domain tend to be extremely costly (with effects that would significantly damage a state's AI industry) and are thus not considered further in this report.

[326] A significant further challenge of the blunt approach is that some hardware configurations may be difficult to change if the regulatory model size threshold either rises or falls. Either direction is possible, since algorithmic improvements may make smaller numbers of parameters of computational operations more dangerous, while of course further work may also demonstrate that much larger models are needed for the creation of politically important dangers. The fast-moving frontier of AI development in recent years underscores how rapidly the technical and policy landscapes can change, thus indicating that policymaking in this space should include provisions for rapid revisions to model size thresholds. See also Sara Hooker, 'On the Limitations of Compute Thresholds as a Governance Strategy' (arXiv, 29 July 2024), https://doi.org/10.48550/arXiv.2407.05694.

ning extremely large models may therefore take different shapes depending on the political choices made during its creation.

4.5.1.2.2 Regulation of model inputs

The inputs that are used to create a model may be the subject of regulation. All of these components, which collectively form the training plan (see Section 4.5.1.1.1), can affect the capability and behavior of the resulting system in important ways. Some examples are:

- **Data**: Some training data types could be banned, such as data that might enable the creation of chemical, biological, radiological, nuclear, or cyber weapons.[327,328]
- **Algorithm**: Regulation might ban the use of reinforcement learning methods that are able to create agents with the ability to plan over long time horizons.[329]
- **Hyperparameters**: Model training that is declared to be a new model (and not a continuation of a prior training run) must declare a verifiable source of randomness against which an initial model weight snapshot can be compared.[330]
- **Compute budget**: As noted in Section 4.5.1.2.1, regulation may also limit the compute budget or parameter number of models.

4.5.1.2.3 Regulation of model behavior or attributes

Regulations may also be framed with regards to the actual behavior or attributes of models, such as their performance on certain tasks. These sorts of tests for models are part of a new and rapidly expanding field of knowledge that includes concepts such as model evaluation (or “evals”), audits, and capability elicitation. This report will not detail the content of such regulations, but will explore in the following subsections how the application of such rules could be verified. Some model assessments can be run against intermediate versions of the model that are produced during the model training process, while others may only be viable after the model has completed training (some distinctions shaping these choices are outlined in the following sections).

To make robust verifiable claims about models, one approach is to combine a proof of training with other model assessments. Three pieces that could work well together are:

1. A training plan assessment can demonstrate that declared inputs are compliant.

[327] Data-based regulations are complicated by the challenge that dangerous capabilities might be possible even if only unproblematic training data is used. Widely known and discussed knowledge of nuclear, biological, and chemical sciences might be sufficient for a model to generate insights that states perceive as dangerous. One approach to this challenge is regulation based on model capability, discussed below. See also Reuel and Bucknall (2024), Section 3.1.1.

[328] Jonas B. Sandbrink, ‘Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools’ (arXiv, 23 December 2023), https://doi.org/10.48550/arXiv.2306.13952; John Halstead, ‘Managing Risks from AI-Enabled Biological Tools’, Centre for the Governance of AI (blog), 5 August 2024, https://www.governance.ai/analysis/managing-risks-from-ai-enabled-biological-tools.

[329] Cohen et al., ‘Regulating Advanced Artificial Agents’.

[330] This is one potential rule which can be verified by appropriate use of a proof of training transcript as explored in Shavit, ‘What Does It Take to Catch a Chinchilla?’

2. Section 4.5.1.1.3 can demonstrate that the model was produced in the declared way. It can prove that precisely the declared data, algorithm, and hyperparameters were used to produce the final model.[331]
3. Model evaluation techniques can (attempt to) prove that the model does or does not have certain abilities or attributes (see Appendix G).

4.5.1.3 Verification approaches for AI development

A simplified view of verification for AI development is shown below, illustrating the different kinds of information available for verification at various stages in the process. This diagram is applicable to either the training or fine-tuning steps of the process, though some differences between the two will be discussed below.

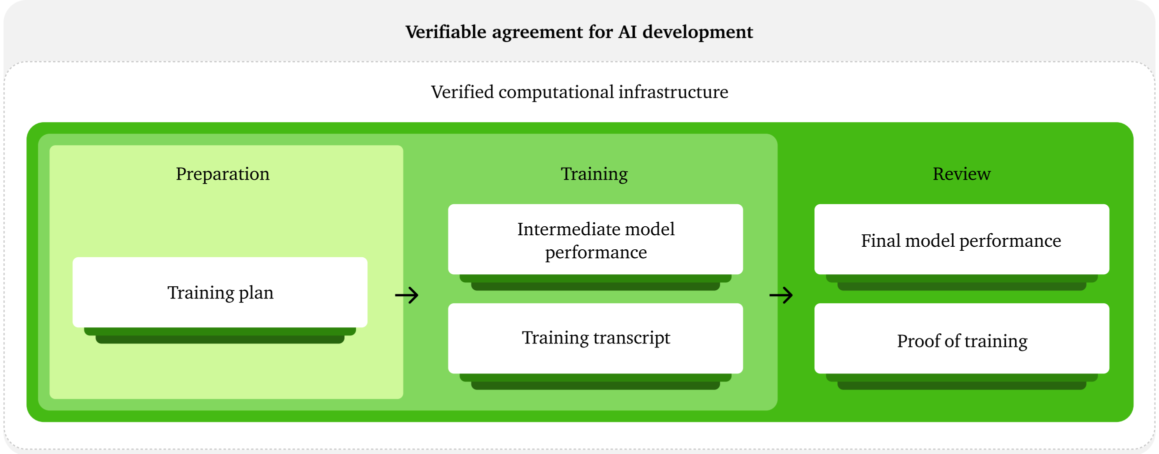
Verifiable agreement for AI development
Verified computational infrastructure
Preparation
Training plan
Training
Intermediate model performance
Training transcript
Review
Final model performance
Proof of training

Figure 4.3: A schematic representation of the information available at different stages of the AI development process. During preparation, all that is known is the training plan. Once training has begun, all of the information from the preparation phase is still available, and further information is available in the form of the intermediate model performance and the training transcript. Finally, during the review phase, all prior information can be made available along with information that only becomes available after training, such as the final model performance.

To verify regulations applied to either training or fine-tuning, there must be verified physical infrastructure coupled with rules applied to at least one of the following phases of AI development: preparation, training, and review. The following subsection will discuss these aspects in more detail, followed in turn by a discussion of how verifying fine-tuning carries some additional requirements.

4.5.1.3.1 Verified computational infrastructure

In order to make robust claims about the AI training being done on computational infrastructure, some kind of verifiable claim must be made about that infrastructure (see Appendix D). Verified computational infrastructure can allow a variety of different claims to be made verifiably, including those claims that are crucial aspects of international agreements and those

[331] Shavit; Choi, Shavit, and Duvenaud, 'Tools for Verifying Neural Models' Training Data'.

claims that allow other kinds of verifiable statements to be made. The following list includes examples of both:

- **Confidential computing**: Verified hardware stacks can make confidential computing a credible way to make verifiable claims about digital objects (see Section 3.5).
- **Disallow digital exfiltration**: Physical enclosure of infrastructure can be combined with other mechanisms to provide guarantees that digital objects such as models cannot leave the premises undetected (see Section 2.5).
- **Digital perimeter**: Relatedly, digital infrastructures can be set up to automatically provide verifiable evidence (such as cryptographic commitments) of the content of digital exchanges within the infrastructure (see Section 2.5.2.4).
- **License-locked hardware**: Verified infrastructure can be demonstrably locked by licensing mechanisms, thus allowing credible claims about the circumstances under which that hardware can be unlocked for use in computations (see Section 2.2.4.5 and Section 3.6.2).
- **Guaranteed encryption**: Similarly, verified physical infrastructure can provide credible guarantees that particular digital objects such as models are guaranteed to be encrypted at particular points in their lifecycle with particular keys—including potentially keys provided by the Verifier, Prover, or both—thus enabling guarantees that the model cannot be meaningfully copied until one or more decryption keys are revealed via governance processes (see Section 2.2.4.7).
- **"Fixed set" hardware**: Verified hardware is placed into a configuration that makes it more costly to create models that would use more compute operations than an amount specified in regulation (see Appendix I).
- **Chip density controls**: Hardware inspections or location mechanisms on chips (see Section 2.2.4.6) could enable the regulation of chip densities. This might reduce actors' ability to train large models, as described above for the fixed set approach (see Appendix J).

4.5.1.3.2 Verification during preparation for development

If development is taking place on verified computational infrastructure as described above, it becomes possible to perform verification during the *preparation* phase—before training (or fine-tuning) commences. Verification during the preparation phase can allow Verifiers to observe the hardware arrangements and the training plan (see Section 4.5.1.1.1), but not the information generated later, such as model snapshots or the final model. Therefore, in addition to the verifiable claims that can be made via verification of the infrastructure (see above), verification in this phase can be used to ensure that regulations pertaining to model *inputs* are applied correctly (see Section 4.5.1.2.2).

Verification during preparation can be applied in a number of different ways for many possible regulatory goals. Here are a two examples of how verification systems could be set up for the preparation phase:

- **Shared data centers**: Data centers that are mutually verified by the Prover and Verifier are used to conduct all phases of training, and thus rules about training plans can be tested and verified before training can proceed (see Section 3.6.1).
- **Greenlighting**: Training hardware can be locked until an appropriate license is provided, thus allowing a governance and verification process to examine the training plan before training commences.[332] One implementation approach would go as follows: a training plan can be securely examined in a mutually verified facility, thus allowing the Prover to demonstrate that it is compliant. Once the training plan is proven to be compliant, a license is provided for that training plan which the Prover can then use to unlock the training hardware and train their model (see Appendix L.1 and Sections 3.5 and 3.6.2). Note that mutually verified licensing systems can be under the physical control of either the Prover or Verifier, which changes the political meaning of such a verification regime (see Section 3.6.2).

4.5.1.3.3 Verification during training

Verification during training allows for in-flight governance of the training process. For example, it could lead to training being discontinued if dangerous capabilities have emerged. In this phase of verification, verification processes can have access to not only the training plan, but also the training transcript and the intermediate versions of the model that have been produced so far.[333] Furthermore, they might even have access to details of training which might otherwise be transient, such as gradient updates.[334] The advantages of verification during training are centered on 1) its access to these normally transient details of the training process and 2) its ability to test for the emergence of dangerous capabilities within the training process itself. Its major disadvantages are that 1) evaluations probably need to be fully automated to be fast enough, 2) only fast, low-cost evaluations are feasible during this phase, since they will presumably be run many times, and 3) there is a tradeoff between development speed and security. If a secure off-site verification system is used, development would be slowed further due to heavy data transmission needs. Alternatively, verification could be done on the same hardware as training. This latter scenario would involve sensitive evaluation content being available on the same machines as model training, thus potentially providing the Prover with a greater chance of extracting that sensitive information.[335] As with verification during the preparation stage described above, both shared data centers and greenlighting are potential ways to implement verification during training.[336]

[332] Lennart Heim, 'The Case for Pre-Emptive Authorizations for AI Training', Lennart Heim (blog), 10 June 2023, https://blog.heim.xyz/the-case-for-pre-emptive-authorizations/.

[333] Here it is presumed that verification during training is taking place on the same hardware—or at least within the same data center—as the main training processes. It is not being done in a separate verification facility.

[334] Storing particularly transient structures might not be desirable under normal circumstances, and it may also not be feasible to both store them and transmit them to a verification facility, depending on the rate at which such structures are produced and their total size.

[335] If verification during training were instead done in a specialized verification facility, this issue would be mitigated. Unfortunately, regularly transferring large amounts of data to a verification facility might be impractical for multiple reasons, including potentially major slowdowns in AI development.

[336] Greenlighting could require that training code also include a structure for running evaluation code provided by the Verifier.

4.5.1.3.4 Verification after training

Verification after training[337] can leverage more data and computing resources than the earlier phases. In theory, it can be used to verify rules about the training plan, the training process, and the final model. The only data that it may lack access to is the ephemeral structures that are used within the training process but not kept, as discussed above.

All of these advantages of post-training verification can be weighed against one key disadvantage. If verification only happens after training, then the Prover could complete a non-compliant training run before they are caught being in violation of the agreement. However, this could be a feature, not a bug, since delaying verification somewhat can be strongly desired by the Prover for various reasons (see Appendix Appendix C.2). Furthermore, to the extent that this disadvantage remains a political problem, it is worth noting that verification schemes are typically not intended to be perfect, and such an imperfect approach could be the starting point for a verification regime that gradually evolves toward covering earlier phases of the AI development process (see Appendix C.3).

Post-training verification can be implemented via the schemes described under verifiable confidential computing (Section 2.5.3) or via a large mutually verified cluster which can be used for both development and verification (see Section 3.6.1). See also Section 3.5 for how security-preserving verification of digital objects can be accomplished in either case.

4.5.1.4 Verifying rules relating to fine-tuning and online learning

Governing the modification of a model after it has been trained introduces new challenges. Fine-tuning is the modification of an existing trained model for new or refined purposes, while online learning refers to models that are regularly and perhaps rapidly updated.[338] Relatively little computing power is required to make small changes to a model, thus making fine-tuning nearly impossible to fully govern if model weights are spread widely. Nonetheless, governance of some hardware might still be useful, even if complete coverage cannot be achieved.[339]

Verifiable regulation of fine-tuning or online learning can be framed in two different ways:

1. Verifying that particular *hardware* is complying with an agreement when it undertakes these operations.

2. Verifying that a given *model* is never modified in a way that violates the agreement.

[337] This is denoted “Review” in Figure 4.3.

[338] Despite the apparent connection in its name, this technique does not relate to the Internet.

[339] At least two political questions are at hand here. First, to what extent should models be allowed to be copied widely if such copying makes it infeasible to track what they are actually used for? Second, how should regulations treat models that come from unverified hardware? Drawing a bright line between the verified and the unverified might be a conceptually simple and politically defensible position, and this would require regulations that both strictly control the locations and copies of models while also disallowing regulated hardware from running inference on unknown models. However, it may prove impractical or undesirable to enforce such a hard line, especially considering the widespread use of open source models today.

While verifying that rules are being followed by governed hardware can use the techniques described above (see Section 4.5.1.3), verifying that a given model is never modified in a non-compliant way requires somewhat different techniques. The remainder of this discussion will focus on this latter challenge.

The verifiable regulation of fine-tuning or online learning for a *model* probably requires verifiable control of trained models. Control of trained models can be accomplished in two general ways:[340]

1. **Model location control**: Verifiably keeping model copies secured within a small number of locations—all employing either strict physical controls and air gaps or a digital perimeter to demonstrate security. Within this approach, the institution that created a model could verifiably control it thereafter, or they could securely transfer it to another institution.

2. **Models that cannot be fine-tuned**: It may be possible to construct models that are extremely difficult to fine-tune—requiring approximately as much compute to fine-tune as it would require to fully pre-train a model of similar scale. Exploratory work on this subject has had some success, but much more study is needed to understand what potential this approach holds for making fine-tuning infeasible even when highly capable adversaries attempt to circumvent these protections.[341] Presuming that such techniques can be developed, this approach could be paired with a governance apparatus that can regulate model creation to ensure that important open source models are broadly safe according to regulated dimensions, thus reducing the overall risks of fine-tuning.[342] Paradoxically, however, similar technologies could also enable certain categories of verification circumvention.[343] A further concern for this approach is that downstream users might want to fine-tune their models to avoid safety hazards that they discover within them, which they wouldn't be able to do if the models couldn't be fine-tuned.

Models kept under physical control can be modified according to governance rules, and all of the techniques detailed in Section 4.5.1.3 can be used to prove adherence to those rules. For example, hardware controls and code attestation can be used together to credibly demon-

[340] Many other factors affect the relative danger of broadly available models. Most were discussed previously as model inputs. For example, if data related to CBRN weapons were very tightly controlled, it might be less feasible for a non-state actor to fine-tune a widely available model to provide them with information about such weapons. Not only would it have been more difficult for that actor to find fine-tuning examples to work with, but the model itself would likely also contain less CBRN-related knowledge. Realistically however, meaningfully changing the availability of data in the public sphere and dark web would be very challenging.

[341] Rishub Tamirisa et al., 'Tamper-Resistant Safeguards for Open-Weight LLMs' (arXiv, 8 August 2024), `https://doi.org/10.48550/arXiv.2408.00761`; Jiangyi Deng et al., 'SOPHON: Non-Fine-Tunable Learning to Restrain Task Transferability For Pre-Trained Models', arXiv.org, 19 April 2024, `https://arxiv.org/abs/2404.12699v1`.

[342] Furthermore, there may be a deep tradeoff with these techniques since it may or may not be possible to make models infeasible to fine-tune on a topic-sensitive basis. A model that cannot be fine-tuned in any way may be of drastically limited utility, but a model that cannot be fine-tuned in only a small number of key regulatory dimensions might be able to retain much of its broad usefulness while also staying safe.

[343] If models can be made that cannot be fine-tuned, then entire categories of evaluations—capability elicitation evals—cannot be used on such models. Furthermore, it might not be possible to unlock password-locked models via fine-tuning to understand what the true capabilities of the model are.

strate that a particular model is being modified. Other techniques detailed above would allow for verification of the training plan and the training transcript for the fine-tuning process. While the fine-tuning algorithms will look somewhat different from the training algorithms, the general verification approaches described above should still be workable.

4.5.2 AI deployment

Regulating model deployment is generally more difficult than regulating model development. A full account of the difficulties of deployment governance is outside the scope of this report. Nonetheless, four major factors are crucial to any conversation about deployment governance. First, models can be used in diverse ways in diverse locations on diverse equipment—thus making it very difficult to provide governance that is sensitive to all these various contexts.[344] Second, governing model usage is likely to require access to much more private or security-sensitive information than analogous governance of model development—thus raising legitimate fears about information exposure, both for individuals and for institutions. Third, even a relatively small amount of model usage might allow dangerous actions such as the creation of a biological or cyber weapon—thus requiring the close examination of a large proportion of all post-deployment model usage. Fourth, most existing techniques for verifying the integrity of inference are very slow—many times slower than inference itself—thus raising the concern that some forms of deployment governance may be computationally infeasible without significant technical advances.[345]

4.5.2.1 Inference governance and strict model control

Inference governance depends on strict control of hardware, and may also depend on strict control of created models. As noted above for the verification of rules regarding fine-tuning, downstream governance may be very limited if models can be copied without controls. While some hardware might have inference governance in place, other hardware could be creating, modifying, or running non-compliant models. For the same reasons noted in Section 1.5.2, proving that rules are being followed about AI deployment requires that there be a credible way to limit the space of things that must be examined. Two categories of solutions to this problem are explored briefly in this report.[346] First, strict control of both model creation and post-creation modification for models (as described above for the verification of fine-tuning, see Section 4.5.1.4) allows actors to be confident that the model will behave correctly and that it cannot be copied and changed without detection. Inference conducted with a controlled model can therefore be tracked, governed, and verified (see Section 4.5.1.4).

[344] Consider that the same inference in different contexts could be compliant or non-compliant. For example, using a model to find cyber vulnerabilities could be fine if done by a white hat hacker, but misuse if done by a black hat hacker. Precisely the same inference content might be employed in each case, thus making the context the crucial dimension.

[345] For more on verifying deployed models see sections 5.3 and 5.4 of Anka Reuel et al., 'Open Problems in Technical AI Governance', 2024.

[346] Future work on AI deployment rulemaking and governance could explore whether there are other ways to simplify this space, such as through a provably limited supply of relevant deployment hardware (such as data centers above a certain size, inference-specialized chips, or robots of a given category), provable limitations on electric power availability, or extremely robust hardware mechanisms for deployment hardware.

Second, control of inference compute allows for fingerprinting of models and thus can be combined with similar fingerprinting within development governance (see Section 4.5.1) to prove that the deployed model is in fact identical to the model that was tested earlier. Such a scheme could also be extended to limit or even ban inference using models with unknown fingerprints.[347] Neither of these approaches are designed to prove anything about the ways that ungoverned compute is being used, but they do provide a way to verify the rules enforced on governed compute.

4.5.2.2 Inference verification concepts

4.5.2.2.1 Inference plan

Analogously to the training plans described earlier (see Section 4.5.1.1.1), an inference plan is all of the information needed to deploy a given model in a way that produces replicable inferences. This information could include all or some subset of the following kinds of information:

- the model's fingerprint,
- the full model,
- precise identifiers for the type and identity of the hardware that the model will be deployed on (perhaps including information about networking hardware, enclosures, and AI-specialized chips),
- a codification of the kinds of inference allowed, including potential limitations on:
    - the topics of inference (e.g., prohibiting disallowed topics such as CBRN weapons).
    - compute budget, number of tokens, number of iterations of test-time-compute, and memory consumption.[348]
- rules regarding how queries from different categories of users will be limited.[349]

In theory, an inference plan provides a description of what will be done that is granular and clearly operationalized enough to be used as a regulatory (and verifiable) declaration. Ideally, it will be specified in a way that allows it to be used as a rubric against which the actual deployment can be tested using automated checks. For example, all inference requests could be required to pass a set of automated tests (including AI-powered evaluations) which can apply to the input, output, or detailed internal operations of the inference process. As

[347] See also Appendix L.3.

[348] Test-time compute is the use of additional compute at inference time to arrive at better answers. Key early products employing this feature are OpenAI's o1 and o3 models. See also Charlie Snell et al., 'Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Model Parameters' (arXiv, 6 August 2024), https://doi.org/10.48550/arXiv.2408.03314.

[349] This would require having different inference plans for different categories of users, who would be identified via standard digital authentication and authorization techniques. Differences in inference plans could be for differing tiers of service (e.g., premium accounts), different types of service (e.g., internal, corporate clients, individuals, government agencies), and for different security levels (e.g., white hat cybersecurity companies). Mixing these many categories via a single API might be inadvisable for security reasons, but providers may nonetheless choose to provide a set of capabilities via one API.

will be explored below, the total desirable suite of tests might be substantial, thus requiring non-trivial computational resources.

Incomplete inference plans might also be useful for regulation. For example, an inference plan stipulating only a model fingerprint along with the unique identifiers for the hardware it will run on might be sufficient for a Prover to meaningfully declare their compliance with simple rules such as "only models that have been shown to be compliant with training regulations are allowed to do inference". While this kind of rule is very limited, it might be sufficient for some governance goals.

4.5.2.2.2 Inference transcript

Analogously to a training transcript described earlier (see Section 4.5.1.1.2), an inference transcript contains the complete record of an inference interaction. When a complete record is not needed, a minimal version of the inference transcript could include the model fingerprint as well as the input and output tokens and metadata.

4.5.2.2.3 Proof of inference

In combination with access to the inference plan, a Verifier with access to the inference transcript and sufficient computational resources should be able to fully verify that the inference transcript is precisely correct and complete. A fast (but less robust) version of this could be accomplished via a remote attestation signature similar to that available via confidential computing (see Section 2.2.4.4 and Section 2.2.4.2). Slower (but more robust) versions might use other techniques, such as zero-knowledge proofs, to demonstrate that the model performed this inference.[350] Today, only the former category appears to be performant enough for usage at scale.

4.5.2.3 Regulating data center inference

Regulation and verification of inference must balance four important factors: 1) sensitive inference data, 2) sensitive evaluation data, 3) computational overhead, and 4) time delay between the operation and verification. To protect inference data, the Prover must be sure that all relevant infrastructure is safe from data exfiltration by the Verifier or any other actors—and the Verifier mirrors these concerns with their evaluation data. Solutions are available to manage both sets of security needs, but at least some of those solutions call for separate infrastructures for AI operations and verification. These separate infrastructures might make time delays too substantial (as in the case of an air-gapped neutral verification facility), raising worries that non-compliant inference might be undertaken for some time before the Verifier would notice. Furthermore, verification processes must be computationally lightweight

[350] Haochen Sun, Jason Li, and Hongyang Zhang, 'zkLLM: Zero Knowledge Proofs for Large Language Models' (arXiv, 24 April 2024), https://doi.org/10.48550/arXiv.2404.16109.

enough to either be rapidly completed in-line on the production hardware or run on the verification data center.[351]

Given the expected low latency of inference, options for inference verification might be limited to two: First, inferences could be subject to verification checks on the production hardware itself as the inference is being completed.[352] Second, production inference processes could verifiably save all the relevant data needed to verify adherence to the rules, and that data could be processed in a separate verification facility. Given that this data might be of substantial size, and processing it fully might be computationally weighty, one workable approach might be a probabilistic inference verification system. Rather than attempting to verify that the rules were followed in every inference exchange, a random subset could be selected for verification. If this system were arranged in a way such that the Verifier could select the random inferences to be tested after those inferences had already been attested to with cryptographic commitments, then the Prover would have limited ability to circumvent verification without being caught.

4.5.2.3.1 Potential political goals of regulating data center inference

Data center inference governance could be aimed at a number of political goals. Here are a few rules that could be enforced and verified via inference governance on specified hardware:

- Only models that have been demonstrated to be compliant with training rules can run inference.
- Inferences can be attributed to governed models.[353]
- Only certain amounts of test-time compute can be used by any actor.[354]
- Permit only certain categories of inference prompt or output tokens (e.g., disallowing CBRN-related topics or stipulating one operating language to avoid low-resource language attacks).
- Limit *sensitive* queries to specific users on specific deployments. If states are cooperating to prevent the unauthorized use of potentially dangerous AI capabilities, they may also deem some uses of potentially dangerous capabilities to be acceptable, such as white hat cybersecurity work—but they may also want to verify that they are both limiting access to such capabilities.

[351] As will be explored below, more costly verification operations can be run on a random subset of inference to provide evidence of compliance even if running the costly checks against all inference operations would be infeasible.

[352] This would slow down inference somewhat. It would also mean that evaluation data must be available on the production hardware.

[353] This can only be enforced imperfectly in the wild, such as attributing text on the internet, since a nefarious user can make further changes to inference results before they employ them. So even if you know every bit of text produced by all governed models, it could still be difficult to match observed text with the generated text. Note that this requires having some kind of database of searchable inferences, which might be a further challenge to verify if its functionality is important. This challenge is not explored further in this report. See also Reuel and Bucknall (2024), Section 3.1.3.

[354] This requires tracking actors themselves via a verified authentication and authorization system—a topic outside of this report's scope.

- Stipulate a combination of different rules, such as rules that diverge based on model deployment location, hardware capabilities, token number, prompt/output rules, compute used per minute, identity of users, etc.
- Allow militaries to demonstrate compliance with rules while permitting the Verifier to know nothing else about their activities—including the location of their compute (see Section 2.5.4.2).

4.5.2.3.2 Inference plan assessment

Inference plans (see Section 4.5.2.2.1) can be subject to evaluation to ensure that they are compliant with regulations. Such evaluation might be trivial: for example, confirming that the plan indeed disallows the various forms of prohibited inference. However, it can also involve much more substantial techniques, such as privacy-preserving checks on inference context metadata (e.g., system prompts) or inference algorithms (see Appendix E).

4.5.2.3.3 Automated enforcement of an inference plan

Hardware-enabled mechanisms can enable the enforcement of inference plans. Most likely these mechanisms will need to operate at the pod level (see Section 2.5.4.1) to ensure that rules about the model itself can be enforced.[355] Mechanisms such as confidential computing, speculative licensing schemes, or flexHEG secure processors could be used to implement such enforcement.[356]

4.5.2.3.4 Verifying adherence to an inference plan

Verifiers could verify that a Prover is adhering to an inference plan in at least three ways:

1. Before inference, the Prover could demonstrate that their hardware is in a compliant configuration and will only accept correctly signed code—which itself is also verified—via hardware-backed mechanisms (see Section 4.5.2.3.3).
2. During inference, verification code can be run alongside inference code to check that the rules are being adhered to.[357]
3. After inference, inference data can be analyzed in a verification facility to examine whether it adheres to all rules.[358]

4.5.2.3.5 Verification approaches for inference

Finally, we can summarize some of the ways that rules about inference can be verified.

[355] If this were attempted at the chip level, then each chip might only see a fraction of the model and other associated information, making it difficult to decide what enforcement is needed.

[356] See Section 2.2.4.4, Appendix L.2, and Petrie et al. 2024.

[357] Verifying this scheme would likely depend on the same hardware-enabled mechanisms as described in Section 4.5.2.3.3.

[358] See Section 2.5.3.

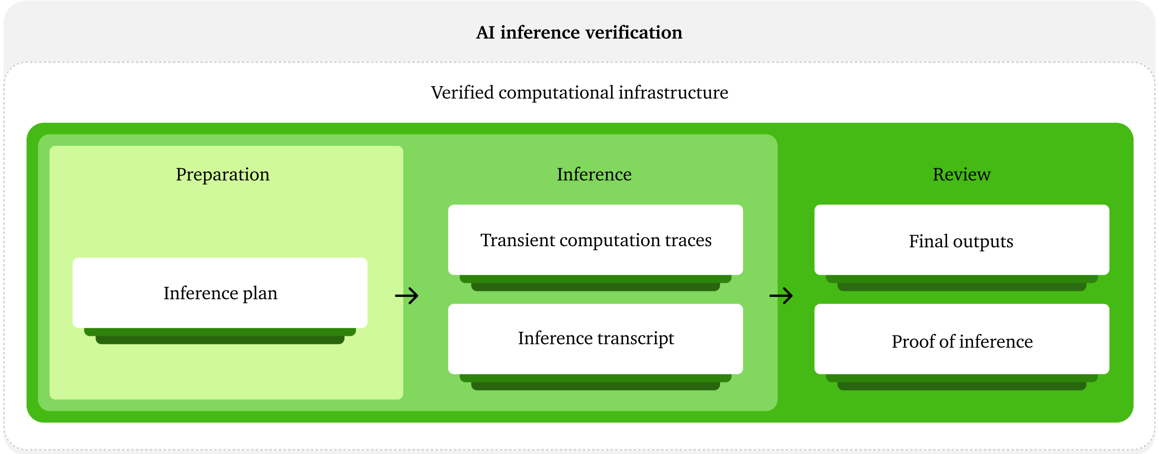

Figure 4.4: A schematic of the different kinds of information available within various phases of data center-based AI inference. During preparation, all that is known is the inference plan. Once inference has begun, further information can become available, including both transient computational traces and the inference transcript. Finally, during the review stage, all prior information can be available in addition to the final outputs of the inference as well as the proof of inference. A discussion in the text engages with the potential difficulty of retaining transient computational traces for later review.

4.5.2.3.5.a Verified computational infrastructure for inference

Verified hardware controls allow a wide variety of verifiable claims to be made. In addition to the list of verifiable claims described in Section 4.5.1.3.1, a number of further claims can be made which pertain to inference in particular. Verified hardware controls can prove that:

- a particular model is the one providing inference (see Appendix L.4).
- only a particular set of identified models have ever been loaded into memory (see Appendix L.5).
- extraneous copies of inference exchanges cannot be made.[359]
- inference plans will be adhered to (see Section 2.2.4.2 and Section 4.5.2.3.4).

4.5.2.3.5.b Verification during preparation for inference

Once an inference plan has been provided, further checks are possible. At this stage, verification processes have access to the hardware arrangements (see above) and the inference plan (see Section 4.5.2.2.1). Verification during the preparation phase is likely centered on ensuring that the inference plan has been tested for compliance (see Section 4.5.2.3.2), and ensuring that the outcome of compliance testing controls whether inference actually begins. Inference plan assessment could be done on mutually verified hardware (see Section 3.5) if the plan contains sensitive information. Once the training plan has gone through testing, it can be implemented on production hardware.

[359] One way that this could be done is by proving that the only copy of the data goes to a verification data stream that is encrypted by both the Prover and the Verifier (see Section 2.2.4.7).

States can ensure that the correct training plan is loaded onto production hardware in at least three ways. First, the Verifier might have access to cryptographic commitments about all data exchanged with the production hardware (see Section 2.5.2.4), thus allowing the Prover to demonstrate that the correct data has been loaded. Second, confidential computing techniques might be used to demonstrate that the correct code and data has been loaded (see Section 2.2.4.4 and Section 2.2.4.2). Third, the production inference hardware might be locked until a signed inference plan is provided—thus allowing the inference plan verification process to "greenlight" the inference plan.[360]

4.5.2.3.5.c Verification during inference

During inference, regulation and verification processes can access computational details that will likely be unavailable later due to their size or sensitivity (see Section 4.5.2.2.1). Furthermore, regulation can apply directly to the output of inference, *before* it is returned to the caller. During this phase, the inference system can enforce inference plan rules (see Section 4.5.2.3.3) which can only be enforced at this time. For example, it can ensure that the model is not reasoning in dangerous and proscribed ways.[361]

4.5.2.3.5.d Verification after inference

Finally, after an inference call has been completed, the Prover can:

- Prove that actual inferences completed align with the rules of the inference plan (see Section 4.5.2.3.4 and Section 4.5.2.3.3).
- Prove that the inference completed was indeed completed with precisely the declared model and input data (see Section 4.5.2.2.3).
- Provide verifiable fingerprints of all inference outputs, thus allowing inference-generated content to be recognized later.[362]

Post-inference verification has access to almost as much information as verification during inference (see above) and has two other advantages. First, post-inference verification can run much more thorough checks on the entire inference process because it is not limited by a strict time horizon, unlike tests that take place during the inference process. These more thorough checks could even allow a full copy of the model to be run with the same random seed as the original, thus allowing perfect replication of the inference results.[363]

[360] See the related discussion in Section 4.5.1.3.2.

[361] The term "reasoning" here applies in particular to the new family of models available since late 2024, termed "reasoning models", which use various techniques to think longer and more expansively about topics before giving their final answers. These techniques in theory allow the "thoughts" of the AI system to be inspected even as they are happening. Some existing techniques use chains of thought that are written in natural language while other techniques might employ the model's internal representations and thus require much more complicated techniques to inspect. The latter category requires "interpretability" concepts that are beyond the scope of this report.

[362] The technical details of this technological challenge are out of scope for this report. In general, a complete system of this sort may allow recognition of whether data in the wild was generated by a particular model, even if that data has been changed somewhat.

[363] Perfect replication is certainly not technologically trivial, but it should be possible with effort. See Shavit, 'What Does It Take to Catch a Chinchilla?'

Second, political agreements might allow inference to be verified only after a specified period of time, which could also evolve during the lifetime of the agreement (see also Appendix C.2). The Prover might desire a gap between inference time and inference verification for security reasons. In such a scenario, the Prover has a deep interest in both a) proving their compliance and b) guarding against the possibility that the verification processes will reveal important security-relevant information. Meanwhile, the Verifier would similarly be balancing their interest in rapid verification to ensure that no non-compliant behavior is taking place against their desire to offer an agreement that is acceptable to the Prover—since the alternative to a time lag might be no agreement at all.

4.5.2.3.6 Regulating sensitive mobile AI-enabled devices

The verification of mobile AI-enabled devices (see Section 2.2.5) such as autonomous weapons differs sharply from the verification of data center operations. Regulating highly sensitive AI-enabled devices such as weapons is a domain with particularly extreme transparency-security tradeoffs. This makes it very difficult to find verification schemes that have a good chance of being politically acceptable even if a form of regulation or mutual restraint is desired by a set of states. Due to the breadth of this report, this section will only be able to provide a high-level overview of the potential problems and approaches in this domain.[364] This is the most explored subfield of AI verification, due to more than a decade of research efforts to understand the possibilities for regulating lethal autonomous weapons.[365]

The analysis below presumes that there are three roughly separable processes involved in creating mobile AI-enabled devices:

1. model development and testing, which is discussed at length in Section 4.5.1.
2. a pairing operation, where a model is installed into a hardware device.
3. activation of the model in a real usage context.

This is a highly simplified view of how AI-enabled devices might be managed. The goal of this separation is to highlight how regulation and verification can take place at different times and at different parts of the AI-enabled device lifecycle.

There are many potential types of rules about mobile AI-enabled devices, but this section will only engage with rules that pertain relatively directly to the embedded AI—not general features about hardware capabilities or deployment.

Before discussing the potential ways that these devices might be verifiably regulated, a few notes about the scope of this analysis are worth emphasizing:

1. The primary goal of this report is to map out the technical-political frontier for verification for various AI-related agreements. It is not a proposal for an agreement, advocacy for an agreement, or a claim about the desirability, legality, or likelihood of such agree-

[364] In particular, this subsection is likely missing approaches that might be reasonable in the civilian realm, where transparency-security tradeoffs are less severe.

[365] A vast number of works exist in this space. One review of particular interest is Ronald Arkin et al., 'Autonomous Weapon Systems: A Roadmapping Exercise', 9 September 2019.

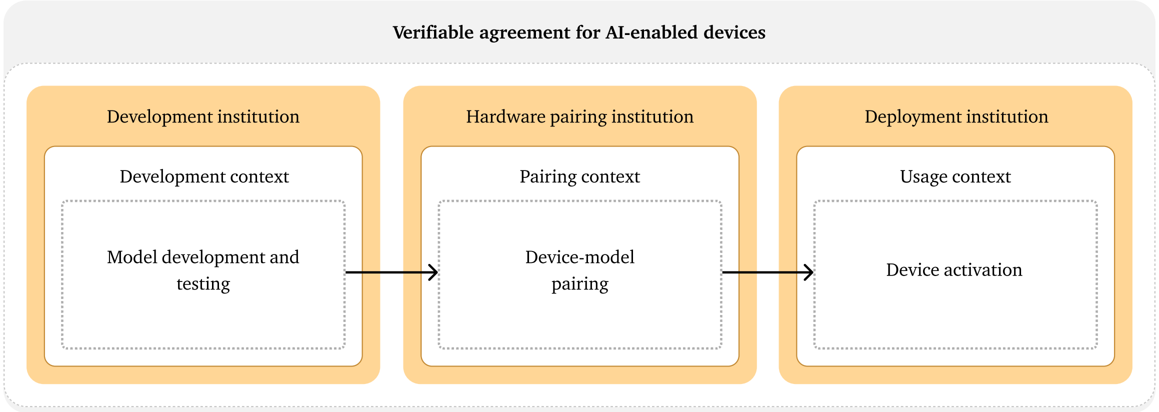

Figure 4.5: A schematic representation of the lifecycle of an AI-enabled mobile device up until it is employed in its usage context.

ments. At most, this is a sketch of a menu of technical verification options that states might choose from if they face scenarios wherein effective verification would open up mutually desirable political options.[366]

2. The emphasis on AI-enabled weaponry (as opposed to other kinds of AI-enabled devices) is deliberate, as weaponry potentially has the most severe transparency-security tradeoff. Thus, focusing on weaponry biases the discussion toward the most robust and secure verification measures that appear to be available now or in the near future.

3. The subsections below emphasize scalable technical verification measures and thus deemphasize verification measures based on personnel, processes, and institutions.[367] This is not intended to sideline the important value that could be added via such measures. Rather, this report focuses on the *technical*-political frontier of verification, and correspondingly prioritizes verification schemes centered on technology and those which have the potential to be scaled up as much as political authorities would realistically desire.

4. This section explores *direct* rather than indirect verification (see Section 3.7.1). It should be noted that current discussions on this topic among states are emphasizing (weak) verification that is extremely indirect. Generally, states are implementing their own systems to ensure that they follow their own interpretations of rules about AI-enabled weapons. International visibility into the details of these implementations is minimal.

[366] Relatedly, it is presumed that if states chose to implement one of these techniques, they would do so in a way that is targeted at their specific political needs. So, for example, they would be judicious in defining which devices would be covered under the agreement. Since states can face particularly intense transparency-security tradeoffs with regards to arms control, it is reasonable to expect that any future efforts to create international agreements to regulate weapons will focus on those weapons that pose the greatest risks. Weapon systems with the potential to create major issues due to mistakes or misuse are one category where controls are more likely. From this vantage point, it makes sense why even rival states have successfully placed important limits on nuclear weapons but have not placed similar limits on most conventional arms. The broader debate about which AI-enabled weapon systems might have this potential is beyond the scope of this report, but some aspects of it are summarized in Section 1.3.

[367] Reviewers of this work emphasized that the trend in international discussions regarding limits on autonomous weapons is toward softer mechanisms such as measures and processes that are publicly adopted by states.

Each of the subsections below briefly analyzes an approach that could allow actors to make verifiable claims about mobile AI-enabled devices employed in sensitive contexts, using the example of AI-enabled weaponry as the central challenge. In the closing subsection, these findings are summarized and their significant political challenges made more clear.

4.5.2.3.6.a Prove that AI control of the device is strictly limited

The hardware design of the AI-enabled device (and associated documentation) is revealed in sufficient detail to a Verifier's agent (e.g., a neutral party) to show that the AI would be unable to access certain hardware functions (such as firing a weapon).[368] Such a demonstration may also need to be conducted physically by the Verifier's agent, and would have to be repeated in some way for all devices of a class.[369] This kind of inspection is very likely to reveal information about the weapon system that is unrelated to the agreement and thus run afoul of the transparency-security tradeoff. The security concerns raised by the inspection process make it unlikely that states would agree to it. To mitigate the severity of these security concerns, hardware-based verification systems could potentially be used, such as those employed for the INF Treaty,[370] or those that have been proposed for nuclear warhead verification.[371]

While hardware can be changed after verification, such changes could potentially be revealed through reinspection, whistleblower action, intelligence collection, or capture of a device or its wreckage. Furthermore, if regulations limit the functions that can be automated, this might be partially circumvented by having a remote automated system send the signals that provide access to the limited hardware functions (e.g., telling the weapon to fire). This would allow automated systems to be in command of all functionalities even if the governed device is not fully automated on its own. However, this could still meaningfully limit the functioning of the device, especially in terms of range and autonomy, since it would need to remain in contact with a remote command system, and such continuous contact is difficult to guarantee for some systems.[372,373]

4.5.2.3.6.b Prove that the device is incapable of undertaking prohibited actions

The Prover could also demonstrate that the device is incapable of certain things. For example, a Prover could credibly demonstrate that the computational abilities of their device

[368] Ronald Arkin et al., 'Autonomous Weapon Systems: A Roadmapping Exercise', 9 September 2019.

[369] Another approach is to focus on the hardware supply networks in order to make claims like this. See also Miles Brundage et al., 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims', 2020, Appendix IV.

[370] Toivanen, 'The Significance of Strategic Foresight in Verification Technologies'.

[371] Sébastien Philippe et al., 'A Physical Zero-Knowledge Object-Comparison System for Nuclear Warhead Verification', Nature Communications 7, no. 1 (20 September 2016): 12890, https://doi.org/10.1038/ncomms12890.

[372] Consider the fact that electronic warfare substantially affects the range and usefulness of remotely piloted weapons if they require electromagnetic signals in the open air (as opposed to over a fiber optic cable as some variants employ).

[373] Further technical and operational challenges with demonstrating "human control" are discussed in 'Verifying LAWS Regulation - Opportunities and Challenges' (International Panel on the Regulation of Autonomous Weapons, 2019), https://nbn-resolving.org/urn:nbn:de:0168-ssoar-77413-1.

are guaranteed to fall short of the computational abilities needed for full autonomy.[374] This could be accomplished by demonstrating that the embedded computational power, memory, internal bandwidth, or electric power source are insufficient to achieve full autonomy.

4.5.2.3.6.c Prove that an AI model is compliant and it is the one embedded

If states make an agreement that includes *rules* about the *behavior* of devices with embedded AI, demonstrating compliance with these rules could in theory be accomplished via a two-part process: 1) demonstrate that the AI model is compliant with the rules and 2) demonstrate that the compliant model is the one embedded in the device. This approach must address the *software update problem* for mobile AI-enabled devices such as autonomous weapons: that the Prover can modify a device into a non-compliant configuration after it's been verified.[375]

Section 4.5.1 details how a model could be made demonstrably compliant with regulations. The remaining challenge, therefore, is demonstrating that the compliant model is the one that is actually embedded into the device. A few avenues are speculatively possible for solving this problem:

- **A hardware mechanism shows the loaded model's fingerprint:** A speculative hardware mechanism akin to that described in Appendix L.4 allows the device to demonstrate which model it has loaded.[376] Physical access to the device is likely required for such checks, thus necessitating a neutral setting for hardware verification (see Section 2.5.4.3).[377] A major security concern for this approach is that a plaintext model installed on a device is subject to theft. One somewhat speculative potential approach for solving this problem to the satisfaction of both the Prover and the Verifier is a hardware-

[374] As with all agreements in this section, an agreement with this goal is currently a distant hypothetical, since the international debate on weapon autonomy has shifted from a technical focus to a socio-technical focus. More recent discussions frame autonomy from the perspective of whether the operation of the systems remains within a responsible chain of human command and control. See 'Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System (Annex III)' (United Nations Office of Disarmament Affairs: Convention on Certain Conventional Weapons (CCW) - Group of Governmental Experts on Lethal Autonomous Weapons Systems, 2019).

[375] Paul Scharre and Megan Lamberth, 'Artificial Intelligence and Arms Control' (The Center for a New American Security, 12 October 2022), https://www.cnas.org/publications/reports/artificial-intelligence-and-arms-control; Vincent Boulanin and Maaike Verbruggen, 'Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies' (Stockholm International Peace Research Institute, 2017), p 17–24.

[376] Very speculatively, the running device will emit Van Eck radiation in a way that allows the specific running software to be verified. It remains unclear whether this mechanism can be used to verify running models in sensitive applications such as AI weapons (especially if a state is trying to trick you), and it also requires the hardware to be active in order for the radiation to be detected. See Matthew Mittelsteadt, 'AI Verification: Mechanisms to Ensure AI Arms Control Compliance', February 2021, https://cset.georgetown.edu/publication/ai-verification/.

[377] An ideal mechanism would allow the retrieval of a verifiable fingerprint for a model even if the device has been damaged somewhat (e.g., battlefield wreckage). Note in this case the model itself should not be retrievable from the device wreckage—thus requiring some kind of ledger (see Appendix L.5) or an encrypted model as discussed under the heading "Device mating with an encrypted model".

backed private key on the device (perhaps backed by a battery and PUF) coupled with a verifiable process for loading the model onto the hardware.[378]

- **Greenlighting:** Chips used in the devices can be of a type that require licenses (see Appendix L.2), thus allowing the Prover to demonstrate to the Verifier that the licensed model is compliant, and that the license pertains to precisely the declared chip that is installed in the device. For more about greenlighting model inference, see Section 4.5.2.3.5. Note that reliable licensing systems are not yet available for chips, making this a speculative approach.
- **Device mating with encrypted model:** Speculative hardware features allow for a model to be installed onto a device in a way such that 1) the model cannot be feasibly copied to run on other hardware, and 2) the hardware cannot be repurposed to run a different model. In theory, such a mechanism would allow more certainty that models and hardware are not being repurposed outside of the verification setting. Note that this goes beyond the licensing proposal described above, because the device-model mating approach presumes that copying the model to use elsewhere is disallowed within the agreement—even for the Prover.[379] Note that this approach would disallow model updates, which is a major technical and political issue, given that software updates are commonly employed and can also be frequent.

4.5.2.3.6.d Prove that device usage is compliant

This category of approaches for regulation and verification presume that rules about the *usage* of the device are the central concern for regulation. This differs from the above sections, which discussed constraining the *capabilities* of the device, either through hardware limitations or model checks.

- **Rules-based on-chip governance**: On-chip mechanisms enforce (simple) rules about model behavior (see Section 4.5.2.3.3). The chip could cease to function or function with lower capabilities if the rules are violated.[380] There are three major challenges with this approach: 1) evaluations increase computational load, which might be intolerable on

[378] To safeguard the model from the Verifier, it must be encrypted in a way that renders it unavailable to the Verifier even if they gain physical access to the device (e.g., battlefield wreckage). It may be possible to use a verified physical process—conducted within a hardware verification facility—to load the model (and confirm the model fingerprint is the expected value established via previous governance processes—see Section 4.5.1) and then record the new model fingerprint generated by the device hardware after the model has been encrypted using the device's private key. The device hardware private key could be generated by a PUF mesh and potentially backed by other logic run by a battery, thus allowing the Prover to know that the key will be destroyed if the device is damaged or loses power. Overall, this allows the Verifier to be certain that the correct model was loaded and to later check whether the fingerprint of the loaded model matches the expected fingerprint—all without revealing the plaintext model to the Verifier.

[379] Such a scheme might be implemented in something like the following way: 1) A model is created and doubly-encrypted by the Prover and Verifier (see Section 2.2.4.7.1) before then going through verification checks (see Section 3.5). 2) The model is embedded into the device in a secure facility that is mutually verified by the Prover and Verifier, thus demonstrating that no other copies were made (see Section 2.5.4.3). The actual hardware design of such a device is highly speculative—see Appendix K for a more in-depth exploration.

[380] One related example is how consumer-grade global positioning system (GPS) devices must shut down if they violate speed and/or altitude rules that prevent them from being used in weaponry such as intercontinental ballistic missiles.

military hardware, 2) evaluation content would be on the device itself, thus revealing it to the Prover and allowing the Prover to circumvent its safeguards, and 3) reduced function of devices might not be a tolerable outcome for high-reliability organizations (e.g., critical infrastructure) or security organizations such as militaries.[381]

- **On-chip logging with neutral offline analysis**: The chip logs key information, which allows a neutral analysis to later verify that all rules were followed (Section 3.6.1). One variant of this idea employs cryptographic commitments to allow the Prover to commit to information long before they reveal it at a time and place of their choosing.[382] It is unclear how much storage would be required for this data on the device itself, and it is very possible that full storage requirements would be impractical.[383] Other proposed implementations of this approach have data storage centralized within military facilities, thus solving the data size challenge but raising further questions about whether this data represents the true operational data for the machine.[384] Overall, these approaches presume that the Prover wants to demonstrate compliance *eventually*, but is unwilling to do so during the operation of a device or even during the duration of a conflict (see also Section 4.5.2.3.5.d).[385] Finally, it should be noted that even though the plaintext information is not revealed until some later time, the cryptographic commitments could in theory be sent much earlier, perhaps even automatically as a continuous part of the operation of all devices.[386] Therefore, while full scrutiny must wait, the Prover is able to significantly constrain their own ability to circumvent compliance checks without being noticed.

4.5.2.3.6.e Summary: Governing AI-enabled weapons

This section has considered four general approaches for verifiably governing mobile AI-enabled devices, with an emphasis on their workability in extremely sensitive contexts, such as the governance of AI-enabled weapons. The approaches discussed in the first two

[381] If automated shutdown is politically unworkable for key institutions, it is unclear whether such rules could have automated enforcement of any kind for those institutions. Therefore, this approach might only work for other less sensitive kinds of devices and completely different approaches must be used for devices employed by these more sensitive organizations.

[382] Marc Gubrud and Jürgen Altmann, 'Compliance Measures for an Autonomous Weapons Convention' (International Committee for Robot Arms Control, May 2013), https://www.icrac.net/wp-content/uploads/2018/04/Gubrud-Altmann_Compliance-Measures-AWC_ICRAC-WP2.pdf.

[383] Consider in particular the idea of systems with substantial endurance, with autonomous deployments ranging from months to years.

[384] A key concept for all of these storage schemes is ensuring that the Prover would not be able to easily change the cryptographic commitments after they have been laid down in storage. The same does not apply to the full-text storage, which can take place on commodity hardware since the verification process flows through the cryptographic commitments and thus it is the responsibility of the Prover to ensure that they can bring the true data to match the cryptographic commitment when they demonstrate their compliance.

[385] Further concerns about this kind of proposal can be found in 'Verifying LAWS Regulation - Opportunities and Challenges' (International Panel on the Regulation of Autonomous Weapons, 2019), https://nbn-resolving.org/urn:nbn:de:0168-ssoar-77413-1.

[386] To avoid the security issues inherent in sending a data stream that tells others how many devices you are operating, transmissions could either be continuous for all devices other than their downtime or could be bundled and delivered on a day or week basis to avoid revealing information about the deployed fleet. Furthermore, cryptographic schemes also easily allow for complex data structures such as a Merkle tree to be committed to, thus providing commitments for many devices without either the plaintext information or *the number of devices* being revealed initially to the Verifier.

subsections—limiting onboard computational control (Section 4.5.2.3.6.a) and limiting hardware capabilities (Section 4.5.2.3.6.b)—appear to be technically possible but it is unclear whether they can be accomplished in a way that states will tolerate. The crucial problem with both approaches is that nearly any possible detail about the design of a weapon can in theory be used by an adversary to devise efficient countermeasures. If privacy-preserving methods could be created for both the *evaluation* of sensitive weapons designs and the *inspection* of hardware, then these categories of verifiable governance might be workable. At present, such methods appear to be speculative but theoretically possible.

By contrast, only speculative approaches appear workable for the dual challenge of 1) proving that an AI model is compliant with governance rules and then 2) *proving that the compliant model is actually embedded in the device*. The least speculative approach described here involves verifiably loading the model onto a device in a way that allows the device to encrypt the model with its hardware-specific key, and then fingerprinting the loaded model. Such a process would allow the Verifier to be sure that the model was loaded and to later test devices (including potential device wreckage) to prove that the loaded model was unchanged—all while disallowing the Verifier from seeing the plaintext model. The specific hardware mechanisms and verification facilities required for this approach are not known to have been created yet, and their feasibility remains very uncertain. Optimistically, it might be possible to create these with several years of focused effort.[387]

The final category of regulation is the most challenging: proving that device usage is compliant. While highly speculative future technologies might allow rules-based on-chip governance, the technical and political issues with that approach are substantial. A logging-centered approach has also been proposed for allowing states to demonstrate that their weapons are used in ways which are compliant with widely-held interpretations of applicable rulesets or legal regimes, potentially including international humanitarian law and international human rights law.[388] This proposal involves no speculative technologies but would likely require some minor reworking of data flows. The political crux for this form of governance is whether the evaluation of compliance could be privacy-preserving or otherwise security-preserving for the state. An ideal governance system would employ privacy-preserving computational operations on the logged data (including the full context for the weapon's use) to demonstrate compliance with the rules. No other states or their human agents would be able to see the plaintext data. If such computational verification approaches could be devised, they could be a key part of a robust verification regime built around governing the use of AI-enabled weapons. This is likely to be an extremely challenging technical problem due to the difficulty of operationalizing legal requirements and for the technical and epistemic reasons explored elsewhere in this work (e.g., Appendices D and F). If states

[387] The claim about several years of effort is a very rough estimate based on the analysis provided in Section 4.5.2.3.6.c and a roughly similar timeline described for a sensitive PUF-protected digital system described in James Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024.

[388] Marc Gubrud and Jürgen Altmann, 'Compliance Measures for an Autonomous Weapons Convention' (International Committee for Robot Arms Control, May 2013), https://www.icrac.net/wp-content/uploads/2018/04/Gubrud-Altmann_Compliance-Measures-AWC_ICRAC-WP2.pdf.

can agree on rules which are amenable to this form of verification, they may be able to avoid the need for human evaluators to see the data (see Section 3.5).

However, in lieu of a fully privacy-preserving technical stack of this kind, human evaluators would be needed and would therefore expose the Prover to at least some security concerns, since the human evaluators would certainly learn things about the Prover's operations, personnel, and equipment which are not strictly required for demonstrating compliance, and such knowledge could be used against the Prover. The political ramifications of this are unclear and may be contingent on other aspects of the proposed verification regime, such as the time delay between operations and verification. Verification taking place even years after the use of an AI-enabled weapon might serve a political purpose, such as deterring some inappropriate uses of the technology, or helping solidify emerging legal norms. However, it is unclear whether the revelation of sensitive information would be tolerable to states even in such circumstances. Equally, other states might have much less trust in a verification system that allows a Prover to significantly delay their demonstration of compliance. The answers to political questions like these will shape the viability of this form of arms control.

5 Conclusion

This report reviewed several families of potential international agreements regarding AI. It found that some of these agreements appear to be highly verifiable today while others face significant verification challenges. Those agreements which resemble prior or existing international agreements tend to be at least somewhat verifiable, even if they were implemented immediately. Agreement types with few prior analogues—such as regulation of AI development or deployment—face more significant challenges. For this latter category, near-term verification mechanisms might be workable for lower-sensitivity domains such as civilian AI, but these are less able to address the more severe transparency-security tradeoff for high-sensitivity organizations such as militaries. Concerted effort by key states and other actors over the next several years may be able solve these problems, thus opening up political options for states to make deals over AI.

Agreements that are at least moderately verifiable today include agreements which relate to the transfer or pooling of knowledge or resources. The verifiability of most of these agreements is primarily limited by the receiving state's ability to credibly demonstrate to the sending state that resources will be appropriately used after transfer. Outside of general political alignment, it is very difficult for the receiving state to credibly demonstrate that their personnel and equipment controls are systematic and sufficient, thus making it difficult for sending states to believe that their transfers will not be copied, sold, or diverted to undesired uses. Other aspects of these agreements can be built atop more robustly verifiable hardware-enabled commitments in key infrastructure such as data centers. With a few years of effort, a wider array of states might plausibly be poised to make hardware-centric credible claims of this kind, thus opening new economic and political opportunities for these states as well as the AI-exporting states that might seek to work with them. In the long run, more aspects of these agreements can be subjected to privacy-preserving digital verification as that infrastructure matures, thus allowing the technical possibility of increasingly robust verification over time.

The three remaining categories of agreements—preparing for emergencies, regulating AI development and inference, and regulating AI-enabled mobile devices—are significantly more challenging. There does not appear to be an immediately workable plan for robust verification of any of these agreements. However, the prospects for these different domains differ in the coming years. While the verification of regulations for AI-enabled mobile devices such as weapons should be expected to remain very politically difficult even with technical advances, the verifiability of preparing for emergencies and regulating AI development and inference can be advanced enormously through the development and deployment of a set of hardware-enabled mechanisms. Counterintuitively, the technology needed for verifying these latter kinds of agreements appears to be broadly available today, but deploying these capabilities in a cooperative manner to build verification systems will likely require major political and logistical efforts. A "crash program" to provide politically workable levels of verifiability for a small number of key data centers might take about one year to complete, and more robust verification systems that cover more compute will likely take between one and three years of intense effort to implement. On a positive note, while regulatory verification of this kind seems difficult to build, once it has been built it then has great potential to enable scalable, privacy-preserving, and fine-grained verification of a large category of computations. This opens up new frontiers in verification *in general,* since it is often possible to transform verification questions centered on physical objects into verification questions that are represented digitally (e.g., via sensors and cameras).

Preparation for verifiability is a key technical and political priority for most of the agreements described in this report. Given the potentially large impact of near-term research, development, and policy action, the following subsections will highlight the areas that appear most valuable for further work.

5.1 Research and development of particular importance

Further (and urgent) development of nascent and prospective technical verification mechanisms may be crucial for the success of international AI governance. It is important for the reader to realize that work on AI verification is in its infancy. While AI verification depends on several mature fields of study, the overlapping challenges of the verification problem mean that further work on AI verification should be expected to significantly improve our understanding. It is helpful to recall that technical work on nuclear arms verification was undertaken at significant scale for many years before it was employed in key agreements such as the INF Treaty.[389]

While verifiable methods for aggregate and *approximate* measures of strategic variables such as compute are available and relatively mature, this report mainly explored verification schemes that could potentially examine much more detailed information in a way that is still compatible with state security.

The central technological crux of this report is *the realistic viability of privacy-preserving methods for verification*. In particular, this report focused heavily on verification mechanisms with the realistic ability to verify that digital objects follow stated rules—*precisely* and *completely*. "Precisely" here means that the granularity of governance rules can apply all the way down to each byte of information. "Completely" here means that governance rules might apply across many computations on many different pieces of hardware, potentially even located in different countries. In sum, this report emphasized approaches that could possibly claim to achieve high-fidelity visibility of regulatory compliance at vast scales—all while remaining technically, economically, and politically viable.

To this end, two technical approaches for achieving *verifiable confidential computing* were proposed, each aiming to allow the Prover to protect their private data while also allowing the

[389] Toivanen, 'The Significance of Strategic Foresight in Verification Technologies'.

Prover to demonstrate precisely what they did in their computations—at a time and place of their choosing. First, hardware-enabled mechanisms similar to confidential computing appear to be capable of embodying all of the information exchanges needed. The crucial areas of further work on confidential computing relate to its robustness and security. In particular, are the hardware roots of trust in existing chips sufficiently robust to support governance needs, or will additional mechanisms need to be added to shore up crucial cybersecurity or physical security? Moreover, can a neutral mutually verified data center be realistically built, maintained, and used for extremely sensitive verification computations? "Crash program" versions of the approach might be viable, since confidential computing mechanisms already exist on some of the most advanced AI-specialized computing chips. Further exploration of this possibility and its limits appears warranted.

Second, hardware-enabled mechanisms can potentially be installed on networking hardware and enclosures (particularly at the pod scale, where dozens to hundreds of chips work together on tasks) to provide cryptographic commitments which in turn allow for credible verification computations to take place within a neutral mutually verified verification facility. Can relatively simple hardware allow for a credible flow of cryptographic commitments to the Verifier in a way that robustly protects the security of the Prover? Key questions remain about how the cryptographic commitment scheme can be made robust within either existing or newly built hardware. The relative maturity of both core technologies—networking hardware and cryptographic commitments—should provide a strong basis for work on this front, but the abilities of this type of stack remain speculative, as it does not yet exist and will not come to exist without further research, engineering, and policy effort.

Both of these schemes depend on a few fundamental capabilities, including 1) a neutral mutually verified data center that is running something like confidential computing and 2) the ability of each side to continuously monitor the activities within these facilities in a way that adequately convinces them that no attacks on the hardware in the facility are taking place locally (launched by either their counterpart or a third party). It is not yet fully clear that building such a facility is possible, or on what timeline it could be accomplished. It is also not clear how a highly secure facility can also be continuously monitored in a way that provides extreme levels of cyber and physical security. In particular, the advantages and limits of various parallel information streams need to be explored, including monitoring via video cameras, electric power systems, acoustic monitors and accelerometers, and other electromagnetic spectrum sensors.

Given that this discussion is in the early stages, further work should also explore whether there are realistic options for verifiable confidential computing beyond the two described above.

More broadly, further research appears warranted in five other domains. First, containerized data centers have been proposed here as a way to address some of the security concerns of particularly sensitive institutions such as militaries. Further work should examine this understudied category of ideas. Second, further work should be done on privacy-preserving verification for hardware developers such as NVIDIA and TSMC. Is it possible for them to credibly demonstrate that their work adheres to strict and testable rules without revealing any trade secrets? More narrowly, could even an open source chip design be fabricated on sensitive hardware in a way that is verifiable so that all actors can be assured that no changes were made to the design? Presuming that all other efforts fail, what are the prospects and limits for cooperative hardware design and construction employing open source designs and only trailing node fabrication hardware?

Third, presuming that a system like confidential computing is workable, can such a system be used to provide robust answers for algorithm governance and the "Who watches the watchers?" problem? Algorithmic assessment seems necessary in order for the Prover to demonstrate that they are not circumventing an agreement via techniques such as password-locked capabilities. The "Who Watches the Watchers?" problem refers to the challenge of verifying the code that purports to be doing verification for another actor. While the full details of evaluation data should not be revealed to the Prover, they are likely to demand to know that evaluations are actually aimed at their declared purpose and not some secret other purpose. In sum, verification will be needed for the tools of verification themselves, including aspects of those tools which cannot be shown to the counterparty. Can the techniques of privacy-preserving computation—such as mutual code review, selectively hidden evaluation content, and arbitrary numbers of nested evaluations—solve this problem in general? One specific idea worth exploring is whether open source and open data evaluations can be developed which can be regarded as a "ground truth" for claims about what other evaluations are actually aiming to accomplish. These open and thus mutually verifiable evaluations (or meta-evaluations) might then be used to provide credible checks on the behavior of other evaluations which have hidden content.

Fourth, can techniques like confidential computing allow for the privacy-preserving verification of declarations about electrical power networks? It would be ideal if Provers could demonstrate that their electric grid is compliant with a set of claims—for example, that there are no hidden data centers—without revealing the details of their grid to the Verifier. Meanwhile, it would be ideal for the Verifier to be able to test whether the Prover's declarations actually comport with the Verifier's best understanding of the Prover's grid. Presuming honest actors, this verification computation might be trivial; but honesty cannot be assumed. Can a mechanism be designed to accomplish this governance goal without revealing to either party what the other knows?

Fifth, zero-knowledge proofs for AI verification should continue to be explored. Major advances in this domain have the potential to completely upend the technological frontier of verification and thus provide far more political options to states.

5.2 Policy action of particular importance

When states are better able to verify agreements, this can open up political space for mutually beneficial bargains. Equally, however, failure to develop verification abilities can lock states into equilibria that they dislike. The recommendations described below assume that the AI industry will continue to rapidly evolve, with significant ramifications for all states. It is further presumed that state governments are interested in keeping some room to maneuver in this rapidly changing world, and thus will be interested in developing their ability to both verify the behavior of their peers and be verified in turn.

Paradoxically, cooperative verification requires that some policy actions be taken unilaterally, while others must be undertaken in collaboration with the other states who might wish to be part of a governance regime together. Generally, the structure and infrastructure of verification interactions must be developed cooperatively, while other preparations must be developed unilaterally—and some components of verification systems should be kept secret, such as the detailed content of digital evaluations.

Unilateral policy action can support the development of verification in at least five important ways. First, any institution can support research and development regarding the key issues identified in the section above—thus improving the general level of knowledge in this space and allowing policymakers to be better informed in their future decisions. Second, any institution can support the open source development of standards, evaluations, and technical stacks which can enable verification. States can accelerate development and rollout of standards by ensuring that their domestic AI regulations are designed to integrate with a standard. Industry players may want to support the rapid design and rollout of such a scheme to increase the probability that governance rules will be interoperable among the different jurisdictions they do business within. Third, any institution can create incentives for serious scrutiny of the technical foundations for verification, including the funding of large "bug bounty" systems to incentivize external scrutiny of proposals. Fourth, states should develop their own evaluations and keep at least some of their detailed content secret. These evaluations should not be limited to those intended for examining completed models; they should also include techniques for examining other digital aspects of the AI ecosystem, including training data and inference exchanges. Fifth, states should avoid co-locating AI development and inference infrastructure with other sensitive assets such as cutting-edge weaponry or the infrastructure for producing or maintaining that weaponry. In the future, states may want to mutually verify each other's data center-scale AI hardware, just as the Cold War superpowers found that they needed to be able to inspect specific sites as they implemented the INF Treaty. If AI hardware and other sensitive assets are located together, otherwise workable verification plans might be rendered infeasible due to security sensitivities, or rendered highly inefficient due to the need to rapidly move infrastructure to less sensitive locations. To avoid these potentially significant costs, it seems prudent to embed this as a policy today and apply it as broadly as reasonably possible.

Cooperative policy action can begin in three ways. First, states should seek to track crucial inputs such as AI-specialized chips. Even if this knowledge is kept siloed for now, collecting it is robustly valuable for future agreements which might need to broadly account for AI-specialized compute. Second, states should allow academic and civil society actors from all states to discuss the technical aspects of verification with an eye toward achieving common understanding of the technical outlines of the problems and their possible solutions. Such conversations can take place more robustly if states explicitly carve out political space for them. Third, it is advisable for states to begin monitoring the emerging discussions about AI verification and consider how and when they want to engage with other states on these topics. The rapid changes in the AI ecosystem mean that progress in both the political and technical spheres may be rapid—so states may have to make meaningful efforts to stay informed about these issues.

Acknowledgements

This report benefited from the insights of many people. In particular, the authors would like to thank Mauricio Baker, Michael Horowitz, Thomas Simpson, José Jaime Villalobos, Aidan O'Gara, Elisabeth Siegel, and Zac Richardson.

We thank Wes Cowley and Amber Ace for copy editing portions of the report. We also thank José Medina for his typesetting assistance and Bernard Barsley for his work on the figures.

Midjourney was used in the creation of the report’s cover image.

Appendix A: Verification of personnel controls

Verifiable controls on personnel include verifiable processes, physical controls, digital controls, and legal controls. Verifiable processes include personnel recruitment, onboarding, and management techniques such as institutional hierarchy, background checks, and exclusion from parallel employment. Physical controls include location, travel, and contact limitations.[390] Digital controls include rigid rules on usable electronics and software.[391] Legal controls include laws against the sharing of particular kinds of sensitive information or laws requiring particular kinds of action.[392]

For each of these types of controls, verification would require a Prover to make itself legible enough to a Verifier to be able to claim that these controls are actually implemented. In general this is an extremely difficult problem, since complex institutions like states have the capacity to affect outcomes in many ways and they typically do not have available means for disavowing the use of such capacities. For example, a Prover might provide voluminous details about a personnel management technique to a Verifier, but then covertly influence the behavior of personnel though additional measures such as loyalty tests and threats (see Section 1.5.2). In some cases, techniques for such restraint are possible, but politically difficult—such as extreme limitations on personal liberty within liberal countries. In other cases, such techniques would provide too much information to the Verifier and thus run afoul of the transparency-security tradeoff (see Section 1.5.1.1).

A more speculative idea is using complex digital systems such as AIs to monitor personnel. In theory, appropriately trained AI systems with extensive access to the activities of the personnel in question might be capable of assessing whether or not those activities are in compliance with an agreement. The key problem with this approach is similar to the approaches described above: while it is easy to provide information to a verification process, it is difficult to prove that the information is *complete* (see Section 1.5.2). The potential answers to this problem—such as continuous AI-powered monitoring of all key personnel and systems—still require proof that these particular personnel and systems are the only ones that could reasonably violate the agreement. Furthermore, given the potentially drastic security costs of continuous monitoring of personnel, it is unclear whether such a scheme could ever be capable of successfully navigating the transparency-security tradeoff. Automated systems with complete access to the sensitive activities of key personnel would be an extraordinarily

[390] All of these have been employed in government programs such as the Manhattan Project. Mario Daniels, 'Controlling Knowledge, Controlling People: Travel Restrictions of US Scientists and National Security', Diplomatic History 43, no. 1 (2019): 57–82.

[391] Limitations on which devices can be used for sensitive work are commonplace in both industry and government. Similarly, many organizations have the ability to monitor or automatically enforce limitations on software on managed devices.

[392] Unauthorized sharing of state secrets is often illegal. Hitoshi Nasu, 'State Secrets Law and National Security', International & Comparative Law Quarterly 64, no. 2 (2015): 365–404.

valuable target for cyberattacks and espionage, and the security consequences of a breach could be substantial.

The prospect of widespread monitoring of key humans also raises the question of whether verification effort might be better spent building comprehensive monitoring of *digital* systems such as those that are used to create or run AI models. While a complete accounting of all humans who could know a fact is extremely difficult, a similarly complete accounting of all cutting-edge AI hardware is comparatively easy. Moreover, many of the systems employed for AI development or deployment—such as power and networking hardware—are very simple compared to a human. Meanwhile, open-ended computational systems such as processors and the software running on them can have some forms of simplicity imposed on them through rules. Most of this report expands on these themes and therefore emphasizes the potential for verification via the monitoring of digital systems rather than humans.

Appendix B: Verification of claims centered on access to personnel

Two kinds of access to personnel are discussed here: interviews and whistleblower programs. The related topic of personnel controls is discussed above, in Appendix A.

Interviews have the potential to provide information about the compliance of the Prover's organization. The usefulness of interviews hinges on whether the people being interviewed actually know compliance-related knowledge and whether they want to reveal it to the Verifier. A verification agreement might stipulate that personnel are obligated to tell the truth to the Verifier's agents, but that alone is not sufficient reason to believe that the truth will indeed be told. As explored in the discussion about personnel controls above (see Appendix A) and the discussion of whistleblower programs below, the Prover typically has significant power to ensure that most people who would know about agreement violations would not be available for the Verifier to interview—and equally ensure that those who do know would be loyal. Furthermore, interviews are a relatively blunt instrument that is likely to run afoul of the transparency-security tradeoff in high-stakes circumstances—and therefore a robust interview-based mechanism would never be agreed to. In a high-stakes domain, the accidental or inadvertent revelation of security-relevant information would potentially be very damaging to the Prover's perceptions of security. However, in low-stakes environments, the ability to interview people across a relevant organization could be more than sufficient for the Verifier to achieve relatively high certainty that actions are being taken in compliance with the agreement. This dichotomy hinges on the relative stakes, since extraordinary circumvention efforts can be expected only in high-stakes scenarios.

Whistleblower programs have been previously discussed as one of the ways that a complex institution such as a state can make verification credible.[393] For example, consider a scenario where a Prover designs a whistleblower program that ensures that a designated group of people are a) guaranteed to know information that relates to certain verification questions, b) each, separately has a regular opportunity to "blow the whistle" via travel to a credible neutral facility where they can provide credible information about the Prover's compliance.[394] Such a scheme could in theory ensure that the people who know crucial things about the Prover's compliance will regularly be put into a neutral secure location in which they will

[393] Akash R. Wasil et al., 'Verification Methods for International AI Agreements', arXiv.org, 28 August 2024, https://arxiv.org/abs/2408.16074v1; Akash Wasil et al., 'Understanding Frontier AI Capabilities and Risks through Semi-Structured Interviews', SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 1 July 2024), https://doi.org/10.2139/ssrn.4881729.

[394] One approach would be to allow the whistleblower and their family to gain permanent sanctuary in a non-aligned third party state. Another approach would be to allow large groups of whistleblowers to interact with a digital system over many days, with the resulting data only produced at the end in an anonymized fashion, so that the specific whistleblower could not be directly identified from the resulting data.

personally be able to tell the truth about the Prover's compliance without fear of retaliation from the Prover.

While such a whistleblower scheme may indeed be possible and useful, to be effective at verifying activities that are central to a state's security, the scheme would have to overcome a set of interrelated challenges. Presuming that the transparency-security tradeoff is not violated by the goals of the whistleblower program (see Section 1.5.1), the program must ensure that the Verifier has a high enough probability of catching a violation while simultaneously ensuring that as little extraneous information as possible is transmitted to the Verifier. Minimizing information transfer might lead to a theoretical ideal such as a single bit being transferred from each individual who enters the neutral secure location—with many such bits aggregated together before they are reported to both the Verifier and Prover.[395] However, even in such a scenario, the Prover has plenty of opportunity to circumvent verification. It can hide problematic activity from the set of people who are part of the whistleblower program.[396] It can ensure that potential whistleblowers are loyal (and would therefore be unlikely to blow the whistle). It can (privately) intimidate whistleblowers into silence by threatening them or their loved ones.[397] The Prover's open-ended action space with regards to its own personnel and their connections means that verifying the lack of such actions is likely intractable (see Section 1.5.2).[398] In short, the Prover's room to maneuver with regards to its own personnel might put nearly any whistleblower program at risk.

Addressing the potential shortcomings in the design of whistleblower programs is a worthy goal.[399] This brief discussion outlined only a few of the key challenges they face. For the purposes of the remainder of this report, it will be assumed that whistleblower systems can be useful for verification in relatively open domains such as academic research in open societies but not for closed domains such as the regulation of militaries or secret government projects.

[395] Other proposals tend to have a much more severe transparency-security tradeoff. Interviews might reveal a lot of extraneous information to the Prover even if the interview questions are structured relatively tightly. Physical defection of a whistleblower to the Verifier's institution could lead to the revelation of every key secret known by that person. The risks of these outcomes would shape whether the states involved would agree to these mechanisms.

[396] Addressing this might be possible via extensive controls on—and verification of—digital infrastructure, which could allow high certainty that specific personnel were shown specific information at a specific time. Such digital controls might be highly compatible with the regulatory verification proposals discussed later in this report.

[397] Potential whistleblowers might be led to believe, for example, that the Prover would be able to figure out who specifically blew the whistle or would simply retaliate against all potential whistleblowers. The former problem is a risk even if the Verifier employs very strict information security protocols since the Prover might be able to learn some of that information via its intelligence apparatus and therefore reveal the identity of the whistleblower.

[398] Moreover, drastic efforts to corral this complexity—such as through the revelation of large amounts of information about the activities and connections of all relevant personnel—are likely to run into a severe transparency-security tradeoff. For example, even revealing the specific location of each person's home or workplace via such a process might be a concern for national security, as these may be security-relevant facts for people in important roles or working in important institutions.

[399] Forthcoming work ("Verifying International Agreements on AI" by Mauricio Baker and others) examines this question in detail.

Appendix C: Best practices in verification

C.1 Agreement content that makes verification easier

The content of agreements can also facilitate their verification. Three examples of such content are worth noting. First, agreements can ban activities that would make verification generally more difficult. One example of this from the START arms control treaty is the requirement that all parties broadcast telemetric information from test launches of missiles in the clear, thus allowing their adversary free access to that information.[400,401] In the AI verification domain, one can imagine that guardrails could be installed in agreements to mitigate particular information problems or other issues. For example, while it might be possible to saturate the bandwidth of a verification system by submitting too many declarations with too much complexity, the agreement itself might reasonably prohibit abuse of the system by rate-limiting declarations or limiting the declaration complexity—or both.[402]

Second, agreements can explicitly legalize things that enable better unilateral verification. For example, the Open Skies Treaty legalized unarmed reconnaissance flights over sovereign territory with the express purpose of providing a means for unilateral efforts to provide better and more complete information than would otherwise be available.

Third, agreements can be centered on declarations from each state, thus drastically simplifying the information problems associated with the agreement. Declarations allow verification efforts to focus on checking whether declarations are correct and complete. As described in Section 1.5.2, this makes discovering non-compliant behavior significantly easier, since all that needs to be discovered is a discrepancy with the declaration.

C.2 Gradual implementation

Implementation phases of political deals are often crucially important for the success or failure of the deal. Examples abound in the history of arms control and the negotiated ends of civil wars.[403] Actors in such negotiations particularly fear sharp shifts in power against them (see Section 1.5.1). Ensuring that deal implementation phases are sufficiently granular and reassuring to all sides is a challenging problem that this report does not grap-

[400] "The Treaty requires the Parties to broadcast telemetric information during flight tests of ICBMs and SLBMs and generally to refrain from any activity that could deny the other Party full access to such information."

[401] 'Strategic Arms Reduction Talks (START) Treaty', United States Office of the Assistant Secretary of Defense, 20 November 1991, https://www.acq.osd.mil/asda/ssipm/sdc/tc/start1/start1-aaa/START1lett-sub.html.

[402] This might be particularly relevant to those parts of the verification system that do not scale up easily, such as inspectors and neutral mutually verified compute capacity.

[403] Barbara F. Walter, Committing to Peace: The Successful Settlement of Civil Wars (Princeton University Press, 2002).

ple with. The remainder of this section describes a few techniques that are useful for gradual implementation.

Cryptographic escrow allows a party to commit to a piece of information (such as the location and composition of sensitive facilities) without revealing that information to anyone.[404] Later, they can reveal parts of the information at times of their choosing—or in response to specific challenges. This allows for the sequential / gradual revelation of specific information to relevant parties.[405] Furthermore, due to its digital nature, the time lag between commitment and revelation can change over time, thus allowing an agreement to begin with delayed revelation and move very gradually toward immediate revelation. Used appropriately, this can alleviate security concerns of a rapid power shift due to revealed information. The political danger of an escrow without any (ideally random or adversarially targeted) revelation of its contents is that it could be some kind of lie—and the time delays in the escrow process might mean that the lie is not detected until much later.

Initial exemptions allow certain activities to be exempted from scrutiny, at least initially. Perhaps a small number of AI workloads or sites can be exempted from some kinds of scrutiny by request of the state. Only a certain number of such exemptions should be allowed, and the relative frequency and size of these exemptions may also provide information to other actors about what is being done (e.g., the proportion of compute going to military AI). Perhaps these limits can be evolved over time to gradually tighten the regime.

Delegation can allow verification processes to be shifted between second-parties and third-parties (see also Section 3.7.1). Verification could potentially begin with third party verification and transition over time to second-party verification. For example, credible third parties at arms length from the Verifier and Prover might be the first parties with a direct hand in verification processes such as inspections. After a spin up period, the agreement could shift toward more direct verification. To make this transition smoother, a cryptographic escrow (see above) could also be used to control when the second-party information (e.g., inspector reports) is revealed, and that time delay can gradually be changed until the second-party verification has fully replaced the third party. Alternatively, verification could potentially begin with second-party verification if a third party such as an international institution is not yet available. Later, once that institution is available, verification responsibilities could be shifted to it.

C.3 Learning and iteration

Verification mechanisms (and their associated agreements) are often imperfect to begin with, but can improve if iterated upon. For example, crucial parts of nuclear arms control agree-

[404] Sébastien Philippe, Alexander Glaser, and Edward W. Felten, 'A Cryptographic Escrow for Treaty Declarations and Step-by-Step Verification', Science & Global Security 27, no. 1 (2 January 2019): 3–14, https://doi.org/10.1080/08929882.2019.1573483.

[405] This scheme typically employs cryptographic commitments for the escrow.

ments required iteration in order to successfully fulfill political objectives.[406] Since the AI industry is largely private, there is potential for companies to resist verification via lobbying, moving their supply chains, and tweaking their standards to manipulate what falls under the verification requirements. For example, NVIDIA built multiple new kinds of GPUs specifically for the Chinese market after the United States placed limits on their ability to sell to China.[407] Actions such as these by market actors should be expected, as they are incentivized to take advantage of economic and political opportunities provided to them by states. From the vantage point of governance processes therefore it makes sense to ensure that governance and verification regimes can iteratively evolve their rules as the contours of the challenge evolve and as new issues are discovered.

C.4 Keep AI infrastructure away from sensitive sites when possible

The computational hardware at the heart of AI-specialized data centers is sensitive in some regards but not in others. One domain in which it is not sensitive is to a human inspector walking by as they assess the facility for compliance. By contrast, other kinds of assets—such as high-tech weaponry or cutting-edge industrial equipment—are extremely sensitive to even this kind of inspection.[408]

Since physical inspection of AI data centers may be a key part of a hardware-enabled verification scheme, it therefore makes sense for states to avoid co-locating AI infrastructure with sites or assets that are too sensitive to allow inspectors near. While some proximity might be inevitable, it makes sense to minimize this issue to the extent possible, since this is one of the most straightforward ways to make AI hardware more inspectable and therefore open up more room for political deals via verification.

C.5 Additional certainty via compound or parallel verification

There are two major ways to make the overall verification system more robust than any of its constituent parts:

1. composing verification mechanism components together to cover for the weaknesses of each component.

[406] Mauricio Baker, 'Nuclear Arms Control Verification and Lessons for AI Treaties' (arXiv, 8 April 2023), https://doi.org/10.48550/arXiv.2304.04123.

[407] Lennart Heim, 'The Rise of DeepSeek: What the Headlines Miss' (RAND Corporation, 28 January 2025), https://www.rand.org/pubs/commentary/2025/01/the-rise-of-deepseek-what-the-headlines-miss.html.

[408] See also Coe and Vaynman (2020); Scher and Thiergart (2024), p 54.

2. employing parallel verification mechanisms that are independent of each other, thus making it more difficult for a covert adversary to successfully defect without being noticed.

The following subsections expand on each of these in turn.

C.5.1 Compound verification

If a component of a verification approach has particular weaknesses, it may be possible to combine it with other components that can address those weaknesses. When combined together these components (each with its own flaws) can become a broader mechanism that has fewer vulnerabilities. For example, confidential computing (Section 2.2.4.4) is vulnerable to attacks that can violate the hardware root of trust within the chips involved. To mitigate this risk, the chips could be monitored from fabrication through to their installation and usage. While hardware attacks on the chip's root of trust would certainly still be theoretically possible, it would be more difficult for a covert adversary to complete such an attack with physical monitoring in place.

A necessary complication to this story is that each component may also have a different implied transparency-security tradeoff (Section 1.5.1) and thus a combination of mechanisms might impinge on state security more than any one of them alone could have. However, this effect can also be mitigated through privacy-preserving approaches, such as the method described in Section 3.5 for verifying properties of digital objects.

C.5.2 Parallel verification

Combining a set of independent mechanisms can also make it more difficult for a covert adversary to circumvent verification. Presuming that each mechanism is fully independent of the other, their probability of successful circumvention might be estimated as $p = p_1 p_2$, where p_1 and p_2 are the probabilities of successful circumvention for each parallel mechanism separately, and p_c is the probability of an overall circumvention success presuming that both mechanisms must be defeated. For example, if we have $p_1 = 0.2$ (20%) and $p_2 = 0.05$ (5%), we could infer that $p_c = 0.01$ (1%). Therefore, two somewhat leaky mechanisms can be used in parallel to achieve more robust overall verification, presuming that their weaknesses and strengths are independent of one another.

C.6 Transparent infrastructure; secret test content

An ideal verification stack for AI-related digital objects (such as training data, algorithms, models, and inference exchanges) would involve a combination of fully public and fully private components.[409] Verification processes, stacks, and infrastructure should be designed

[409] Ben Bucknall, Robert F. Trager, and Michael A. Osborne, 'Position: Ensuring Mutual Privacy Is Necessary for Effective External Evaluation of Proprietary AI Systems' (arXiv, 3 March 2025), https://doi.org/10.48550/arXiv.2503.01470.

and built in the open to the extent possible, so that they can benefit from mutual verification, open scrutiny, and perhaps even substantial “bug bounty” programs intended to uncover issues. By contrast, at least some of the actual content of evaluations—such as the data that model evaluations use in their tests for AI model behavior—should be kept private. Public data evaluations might be useful for making some evaluation capabilities available broadly, but there is a significant danger that an evaluation with public data will become useless as it allows actors to “teach to the test” as they design circumvention attacks. Therefore, it makes sense for major actors such as states to have their own privately held evaluations that are never revealed to any other state or to the public.

Appendix D: Reliable digital verification typically requires looking at the whole stack

Making verifiable claims using digital stacks often requires verifying the stack from "the ground up". The "stack" is the hardware and software that enables computations to happen in the desired way. In this report, key portions of the stack discussed include a) data center infrastructure, b) networking hardware, c) general-purpose computational devices such as CPUs, d) specialized computational devices such as AI-specialized chips or GPUs, e) all associated components of these digital systems, including their memory and internal connections, f) all software running on the system to be verified, including its operating system and all applications.

Perhaps the central problem with making claims with digital systems is that the credibility of the claim typically depends on your trust in all the layers of the stack that undergird the claim.[410] For example, if a web page says something, not only do you have to trust the website software involved and the Internet infrastructure intermediating your exchange, you also have to trust every layer of the stack listed in the paragraph above. Transport security techniques can make it possible to largely ignore parts of the "top" of the stack, including communication systems in between the two systems.[411] Furthermore, *presuming robust hardware integrity*, hardware keys can enable remote attestation and confidential computing (Sections 2.2.4.2 and 2.2.4.4)—though it should be acknowledged that robust hardware integrity also depends on a stack of its own, including not just physical access to the hardware but its supply chain (see Section 2.2.3).

Given the emphasis on hardware enabled governance for AI verification (including in this report), it is instructive to realize that when talking about competent states such as the great powers, *unrestricted physical access to hardware undermines all potential hardware governance*. There is no guarantee that a given state can break the governance systems installed in hardware provided to them, but there is certainly no guarantee of the opposite. Given that most hardware and software have been repeatedly demonstrated as having major security issues in *public*, it is best to assume that any system that has not undergone extreme levels of scrutiny—for years to decades—has security vulnerabilities. A corollary to all of this is that

[410] The one major exception to this claim are some kinds of cryptographic proofs. Some of these can be verified as true or valid with absolutely no requirement that you know anything about the system that generated that proof. In this scenario, you must only trust your own computational system and its ability to check the proof.

[411] Presuming that it is possible to recognize the appropriate interlocutor (e.g., via a public key or similar technique), perfect encryption would mean that data cannot be tampered with although it could be omitted. More sophisticated schemes can provide some guarantees of both correctness and completeness. Finally, the prospect of quantum computers has raised the possibility of "harvest now, decrypt later" attacks, where encrypted data can be saved now and then later (speculatively, years or decades later) decrypted using quantum computers that exist at that later date. Encryption schemes are now available which should be robust against this attack, but the most commonly employed schemes today are not.

supply chain security for hardware is also required for any piece of hardware that cannot be verified downstream, such as advanced semiconductors (see Section 2.2.3.3).

This is not to say that security is hopeless. It is certainly possible to achieve improving security over time with attention, iteration, incentives, and layered mechanisms—as discussed in Section 2.5.1.

Appendix E: Algorithm verification

When evaluating the compliance of computational activities such as training or inference, one of the central concerns is the algorithm being used. Algorithm code can plausibly be the most sensitive information in the declaration because it embodies many of the AI discoveries made by the Prover. However, algorithm code is also a pathway by which the Prover might circumvent regulation to produce a non-compliant model without failing any automated checks.

One example of a circumvention attempt is algorithm code that makes a model substantially underperform unless provided with a specific password in the query. Coupled with an algorithmic insight that improves the quality of the model, a training process that locks most model capabilities behind a password would allow a Prover to demonstrate compliance (since the capabilities of the locked model might look reasonable to external tests) while in fact training a model with capabilities that in fact go far beyond what the regulations expect. Discovering such a password may be possible via inspection of the algorithm code and automated inspection of the data.

As discussed in Section 3.5, having a human assessor look at algorithm code could be perceived as a substantial security problem by states and should therefore be done sparingly. Automated assessment tools should be able to check for obvious circumvention attempts, but it is unclear whether they can be the sole answer to the challenge of algorithm verification. One particular challenge is that the Prover might also fear that an evaluation of their algorithm could provide important information about their capabilities. For example, imagine that the Verifier writes an evaluation that checks whether the Prover is employing a certain algorithm in their code. If such an evaluation were run, the Verifier would learn something about how the Prover is doing their machine learning. If a suite of such evaluations were run, the Verifier might be able to learn quite a bit about the Prover's code.

However, presuming as we do in Section 3.5 that the Prover always retains the ability to say no to a given evaluation given what they see in its code, we can also suppose that the Prover will say no to evaluations that would reveal important secrets. Problematically, this also means that the Verifier might also struggle to convince the Prover that their evaluation code is reasonable without revealing their code in its entirety—with the knowledge that revelation of their evaluation code could allow the Prover to submit code that passes inspection checks while circumventing the intent of the checks. Therefore, in the absence of simplifying assumptions, we can expect a kind of dance between the Verifier and Prover, where they propose different ways of demonstrating that the algorithm code is compliant without revealing sensitive secrets about it to each other. It is unclear where such a process would end up. The purpose of the rest of this section is listing some ideas that could be used to begin exploring ways that a broader ecosystem of algorithm governance could contribute to a solution to this problem. See also the related Appendix F which grapples with a similar problem relating to evaluation content instead of algorithms.

Strict coding practices: Strict and modular coding practices for both evaluations and algorithms might allow the potentially rampant complexity of these objects to be brought down to a more manageable level. Decades of progress in software engineering techniques have unearthed numerous ways to manage complexity and ensure that modules (including modules with unknown internals) abide by strict behavioral protocols.

Open-source code: An extreme transparency measure is using open-source code. In such a scenario, the Prover does not have any security concerns about their code and therefore there is no concern about letting it be inspected.

Third party efforts to minimize the "algorithmic overhang": If open-source techniques are far behind closed-source techniques, then an "algorithmic overhang" may exist, where a Prover could undertake training that appears compliant but is not (see for example the password locked model example provided above). Minimizing the algorithmic overhang may therefore be desirable for a number of reasons. Third party states or other organizations could attempt to advance the state of open-source capabilities in an attempt to reduce this overhang and therefore provide less wiggle room for states to "sandbag" with their models. Of course, expanding the capabilities of open-source algorithms will also proliferate the ability to produce more capable models. One potential way to mitigate broad proliferation would be for state of the art algorithms to be shared among a more select group, such as a small group of experts or representatives of key states who contribute to the Verifier's knowledge of what the expectations should be for the performance of leading models. Of course, such a proposal would likely be strongly opposed by those who want to see broad-based development of AI.

Algorithm registry and layers of judgments: An algorithm registry could be as simple as a hash of the algorithm code along with an owner.[412] Judgments about algorithms would be rendered by authorities with the power to see algorithms in their full detail (such as domestic regulatory agencies). These judgments would be public. Therefore, if a large model is being proposed in a training plan (see Section 4.5.1.1.1), the algorithm it employs should already have been judged compliant by a domestic regulatory agency. If a regulatory and verification process at the international level calls into question the compliance of the algorithm, this could cast doubt on the ability or trustworthiness of the domestic regulatory agency that signed off on the algorithm in the first place. Third-party assessments of the algorithm could lend weight to this judgment. Hypothetically, the explicit endorsement of the algorithm by the domestic regulatory agency should make it more costly for the Prover in a scenario in which they are caught, thus slightly reducing the desirability of defection. Finally, if an agreement also requires that an algorithm be subjected to future evaluations that are developed, the Prover considering a regulatory circumvention would have to believe that they would eventually face a very serious risk of either being revealed to be out of compliance or having

[412] See also the related idea of a model registry (Appendix L.3).

to say “no” in response to a perfectly reasonable evaluation—which other actors might take as them being out of compliance just the same. See also Appendix F for more on this point.

Estimate the capability scaling of algorithm code: Even keeping the algorithm secret, small experiments could be done to test its ability—using known data sets. This could allow the scaling law for this algorithm to be estimated and thus allow sandbagging[413] attempts to be discovered (see more on this in Appendix G).

Require an open API version of every algorithm: Building on the idea above, another highly speculative technique for ensuring that the general capabilities of algorithms are well-known would be to require that an open API serve a small or medium-sized model trained using that algorithm and employing similar data to the real proposed model. Metadata about the small model could be available publicly, thus allowing some publicly verifiable tests of the learning capabilities of the model. The key reason why this approach may be entirely unworkable for sensitive applications is that even access to a small model over an API might reveal a lot about what the model is designed for, what it is capable of, and what its weak points are. All of these factors may in fact be deemed sensitive by the Prover for sensitive models. However, in the civilian realm, such a proposal is less outlandish provided that it can guard against the theft of trade secrets or training data.

Algorithmic escrow: Building on the simple algorithm registry discussed above, one can also build a political agreement that requires the open-sourcing of any algorithm that meets certain criteria—and the cryptographic commitments about the algorithm code mean that the revealed code will have to match the true code. For example, algorithms submitted to the registry more than a few years earlier might reasonably be revealed, presuming that the capabilities of open source models have caught up—or that all meaningful competitor models built by corporations or states are presumed to have already embodied these insights. Another potential trigger for open sourcing would be a desire to build a model with a large compute budget. It is important to realize that small specialized models differ from large multimodal models in compute budget by several orders of magnitude. If international governance is primarily worried about the loss of control or strategic effects of extremely large models, then that governance system could require that larger models have to reveal their algorithms more quickly—perhaps even before training is completed for those models that are at the upper end of the compute budget. A variant of this idea would include various levels of revelation described above, including limited assessor access, limited release to key personnel, open API access to a small model, and finally open-sourcing.

[413] Teun van der Weij et al., ‘AI Sandbagging: Language Models Can Strategically Underperform on Evaluations’ (arXiv, 6 February 2025), https://doi.org/10.48550/arXiv.2406.07358.

Appendix F: The "Who watches the watchers?" problem with privacy-preserving evaluations

Another variant of this problem is that the Verifier's secret evaluation content might be used to meaningfully explore the Prover's data in a security damaging way. For example, the Verifier might be able to test whether a Prover-provided AI model is capable of recognizing a particular camouflage pattern for military hardware—thus revealing information about a potential vulnerability of the Prover's military. Guarding against this kind of attack requires that Verifier-provided evaluation code and data also be subject to evaluation checks. Naturally, one might ask how the evaluation checks used in this domain are verified—thus creating a potentially infinite regress evocative of the "Who watches the watchers?" problem in politics—where empowered authorities must also be watched over by someone, thus raising the question of who is watching them in turn.

Robustly addressing this kind of problem is beyond the scope of this report, but a few avenues of work are worth mentioning:

1. Third-party states or institutions might be able to acquire somewhat more access to private code and data due to their political neutrality (see the process described in Section 3.5, but imagine that assessors are from neutral states). Their involvement in a verification process might allow an important check on the behaviors of the Prover and Verifier.

2. Public meta-evaluation code and data might be developed for testing private evaluations. While it would be possible for sophisticated efforts to circumvent these checks because they are publicly available, these circumventions do not come with zero cost. They would make it more difficult and costly for the circumvention to take place.

3. Similarly, private evaluation and meta-evaluation capabilities might be developed by many actors simultaneously, thus allowing direct comparison of outputs and checks among this set of codes. If a comparison effort yielded diverging results among different kinds of evaluations, it could lead to an escalating process of verification of those tools (see Section 3.5 for how humans might be brought in).

4. If declarations are made about the strict purpose and design of evaluations, then it is possible to imagine that the algorithm or the data of the algorithm could be subject to random changes—exchanging them with open-source evaluations that are collaboratively developed. If the behavior of the evaluation sharply diverges from the declaration when code or data is interchanged, it would be an indication that something is wrong with the declaration. To be feasible, this approach would need a very tight specification for evaluation code and data for the declaration to be built around.

5. Similarly, strictly structured declarations about the intent and structure of an evaluation would allow for claims about the private algorithm and data to be tested by various means such as simple codes. If the Prover does not get to see the code until after they have committed to the evaluation content (via a cryptographic commitment), then they run the risk of either being detected in their circumvention or having to say “no” in response to a clearly benign evaluation. In either way, the Prover would be embarrassed in front of its peers and potentially deemed out of compliance with the agreement. The challenge with this approach is that revelation of the code of meta-evaluations would be ongoing, thus reducing their usefulness as they can only be used in this way once (after use of such an evaluation, the Prover can prepare their own evaluations to face all known meta-evaluations). Overall, this approach does allow for the Prover to be “kept guessing” about the tests to which their committed evaluation will be subject to, thus making them potentially more leery of submitting evaluations that are in fact accomplishing different goals than their declared purpose. Furthermore, an agreement might require that an algorithm or evaluation is subject to any future tests as well—thus making any commitment by either side subject to a potentially infinite series of unseen tests, a prospect that would likely be politically daunting for any actor seeking to submit a resource that diverges from their declaration.

Appendix G: Model evaluation

Model evaluation encompasses a wide range of techniques that attempt to understand the capabilities and characteristics of AI models. This section will contextualize the problem and then provide a brief description of a few specific issue areas that relate to verification.

This section presumes that infrastructure for running code-based evaluations against AI models already exists and is privacy-preserving for both the Prover and the Verifier. For a description of how this might be done, see Section 3.5 Note also that in the context of verification, earlier parts of this report also discuss the regulation of model inputs (such as data, algorithms, and hyperparameters) and model behavior and attributes (Sections 4.5.1.2.2 and 4.5.1.2.3). Furthermore, another appendix examines the related challenge of algorithm verification (Appendix E). The question of how evaluation code and data can come to be trusted even if it is hidden is examined in Appendix F.

Presuming that privacy-preserving infrastructure exists, the remaining challenges may still be substantial. Sufficient information needs to be gleaned via tests using code and data, and they must be of reasonable compute and time requirements.[414]

Evaluations are an incredibly diverse category, including tools that differ on the basis of several dimensions, including:

- Goals, including benchmarking capabilities—such as dangerous capabilities like CBRN weapon construction knowledge; testing compliance with rules; and testing political dimensions such as apparent bias.[415]
- Technological implementations, including: request-response tests for a model, deep inspection of model activations and weights, and tests enabled by other AI tools—including everything ranging from simple classifiers to large multimodal models.
- Complexity of efforts to find hidden capabilities, including deep inspection as mentioned above and also capability elicitation using various kinds of tuning—including prompt tuning or fine-tuning of the entire model.

A much more complete survey of model evaluation is available in Reuel and Bucknall (2024), section 3.1.

G.1 Open-ended challenges

Even with full access to all the model inputs, model training snapshots, and the model itself, verifying some rules is far from trivial. Three major challenges are worth a brief discussion: sandbagging, aggregate models, and distillation. The issues in these spaces indicate that fur-

[414] For compute budget, see also Reuel and Bucknall (2024), Section 3.3.2.

[415] These are sometimes desired by the creators, regulators, or users.

ther work will be necessary to understand the potential and limits of model assessment even presuming that the assessment efforts have full access to model inputs, weights, and inference exchanges with a substantial compute budget.

G.1.1 Sandbagging

Under normal circumstances, model capabilities are often gradually discovered for months after the model's release as users attempt to accomplish different tasks with them. Evaluations of capabilities must always be seen therefore as a *lower bound* on model capability, not an upper bound.[416]

A more dangerous and concerning form of this problem is a covertly adversarial Prover who is attempting to hide non-compliant AI capabilities in a way that evaluation suites will not find them.[417] Most likely, a circumvention of this kind will require a non-public algorithmic insight that notably advances the capability frontier (see Appendix E). For example, a Prover might train a model with non-compliant (and perhaps dangerous) capabilities, but have those capabilities locked with a password. Without the password known only to the Prover, the dangerous capabilities will not be demonstrated. In theory, this allows the Prover to submit this model for evaluation fully expecting that no known evaluation technique will catch its dangerous capabilities—evaluation results might look reasonable, and capability elicitation may not work.[418]

It is not clear at this juncture whether such an attack would require algorithmic changes which could in theory be detected (see Appendix E) or whether it could be accomplished via data manipulations that are difficult to detect. It is also unclear whether even skilled humans with full information access could catch an attack of this kind unless they had the time and resources to extensively test the provided algorithm on other (known to be uncorrupted) data sets and thus check its scaling laws, since knowing the true capability curve of the algorithm might allow sandbagging to be detected (see Section 3.5 for more on human-in-the-loop full-text verification).

In terms of governance of this kind of attack, another avenue is to cap the maximum size of models somewhat below what is believed to be the smallest model that could possibly pose a significant threat, given public knowledge on the issue (see also 'Third party efforts to minimize the "algorithmic overhang"' in Appendix E).

[416] See Reuel and Bucknall (2024), Section 3.3.1.

[417] Teun van der Weij et al., 'AI Sandbagging: Language Models Can Strategically Underperform on Evaluations' (arXiv, 6 February 2025), https://doi.org/10.48550/arXiv.2406.07358.

[418] This is an open area of investigation. The reason why basic capability elicitation efforts might not work is that there may be no signal to boost. That is, the password locking of the model might be so effective that there are no hints of dangerous capabilities and thus weight updates based on those hints might not be possible. Direct manipulation of model weights would of course be possible, but that is an open research domain of its own. In general, without a clear understanding of how the model is actually thinking (another research domain known as interpretability), these efforts may fail.

G.1.2 Aggregate models

Another particularly thorny challenge is aggregate systems such as “mixture of experts” approaches. It might be unclear to the automated evaluations how the particular system being examined will relate to other systems. It might be possible to circumvent safeguards and build an enormous system that is composed of smaller models tied together. It is not clear precisely how a combination of evals could catch this type of circumvention. Perhaps one approach would be relatively rigid regulatory rules about the clear and obvious connection between 1) the types of training data and 2) the declared uses of the system—where mismatches between these expectations or obfuscation of either data or model uses would be an immediate cause for alarm. For example, known public algorithms could be trained on a random portion of the declared data to prove that the data actually is what it is supposed to be—or known public data could be used to train using the secret algorithm to prove that the algorithm is doing what it is supposed to be doing (see also Appendix E). Relatedly, strict declarations might be needed about data and control flow to prove that none of the declared AIs are the overseer of other powerful AIs.

G.1.3 Data-centric distillation

Distillation techniques may allow extremely capable AI models to be created from less data. This might be accomplished via an existing model being used to create training data that has the express purpose of being used to make a more powerful model. It is not known how powerful this technique could be.[419] Explicitly testing for the use of this technique might be very difficult, since it might employ otherwise standard algorithms and compute budgets. One hypothetical approach is data provenance—where the Prover has to demonstrate somehow where their data has come from to demonstrate that it is not distilled (e.g., by tracking all inference from all large models—which would be a gargantuan and probably infeasible task). It may be more practical to assume that techniques like this will be used, and both algorithmic progress indicators and the regulations that are based on them must also take these kinds of capability changes into account to ensure that regulations (and verification) are appropriately tuned to the challenge as it evolves.

[419] Toby Ord, ‘Inference Scaling Reshapes AI Governance’, 12 February 2025, https://www.tobyord.com/writing/inference-scaling-reshapes-ai-governance.

Appendix H: Systemically risky AI

One potential definition of *systemically risky AI* takes into account two pathways to systemic risk: a) a credible chance of great power war (such as via a major power shift) and b) the potential loss of human control and its consequent risk to all humans.[420]

Given what is known today about our limited ability to understand and control complex AI systems, one clear example of such a systemically risky AI would be an attempt to build an agentic artificial superintelligence. This category is largely unrelated to the intentions or character of the institution(s) building the AI, since there are hypothetical AIs powerful enough that no existing institution would reasonably be trusted with their creation and control.

For clarity, several other kinds of AI should be *excluded* from the definition of systemically risky AIs. These include:

- economically or socially valuable AIs which are clearly not capable enough to credibly cause either a great power war or loss of human control.
- many forms of "transformative AI", since widespread transformations might be accomplished via relatively small, narrow, and safe AIs.[421] While maintaining peace in such a scenario might be challenging, the strategic risks are very different from those posed by either the creation of a singular superintelligent system or by the race to create it—and therefore those risks should be dealt with in different ways.
- AIs that great powers are willing to allow to be proliferated. Even if a given AI or system of AIs is very capable, if leading states are not trying to control proliferation of such systems, that is a strong signal that such systems are not deemed a systemic risk. In fact, state efforts against proliferation should be expected for AIs that fall well short of being systemically risky, such as AIs with significant potential for misuse.

Systems of many AIs in combination should be considered as well, such as the "mixture of experts" approach, but this line of reasoning is not deeply explored in this report.

A new definition was desirable despite overlapping with several prior definitions because no prior definition explicitly captures systemic danger (an issue of clear international concern) without covering many kinds of AI which are not systemically dangerous.[422] Rather than a more technical definition centered on particular characteristics of an AI system, this new

[420] The EU AI Act has a narrower definition for models with *systemic risk*. They focus only on the potential for the models to directly harm humans, not their potential for triggering war. 'General-Purpose AI Models in the AI Act – Questions & Answers', European Commission, 20 November 2024, https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers.

[421] Karnofsky, 'Some Background on Our Views Regarding Advanced Artificial Intelligence'.

[422] For example, 'transformative AI' covers economically transformative AI without distinguishing between systemic risk and a more gradual and diffuse transformation. See Karnofsky.

definition foregrounds the potential political effects of an AI system and is therefore also workable in a world in which AI development paradigms sharply change.

Appendix I: Hardware configurations for "fixed set" policies

Verified hardware could be placed into configurations that reduce the usefulness of that hardware in non-compliant training runs. One example of such a configuration is the "fixed set" approach, where hardware controls prevent chips from being used in a large collaborating array by only allowing high-bandwidth communication with a specific, limited set of chips—thus making extremely large training runs more difficult and costly to complete.[423]

Prior work has described how *on-chip* hardware mechanisms could be implemented which enforce such a policy.[424] However, another approach is to implement communication limitations in other ways, such as via network hardware. For example, a pod of a few dozen GPUs could be configured to only be connected to the rest of the world through a single network interface with limited bandwidth in order to prevent it from being used as part of a larger cluster. Hardware controls and monitoring could then be used to verify that the cluster's physical configuration has not been changed. Such hardware configurations would be easily reconfigured to be non-compliant, thus necessitating comprehensive monitoring.

Depending on the politically desired slowdown in the training of the largest models, the inter-pod bandwidth in such a design would need to be surprisingly low (e.g., 1 Mb/s) to be robust against reasonable advances in the efficiency of distributed training.[425] If either kind of fixed set approach could be implemented and verified, it would allow a Verifier to be more confident that the maximum rate at which the Prover can train large systems is limited.

[423] Such runs typically require tens of thousands of chips running in tight synchrony for months.

[424] Gabriel Kulp et al., 'Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090' (RAND Corporation, 18 January 2024), https://www.rand.org/pubs/working_papers/WRA3056-1.html.

[425] Scher & Thiergart (2024) present calculations that help calibrate our expectations of how difficult this would be. They examine the challenge of making training approximately 100 times as difficult as it would be on ungoverned hardware—including expected advances in distributed training efficiency. What they find is that the inter-pod bandwidth limitations would need to be surprisingly low (e.g, 1 Mb/s), thus necessitating alternative methods for other crucial operations, such as loading models into memory. Aaron Scher and Lisa Thiergart, 'Mechanisms to Verify International Agreements About AI Development' (Machine Intelligence Research Institute, November 2024), page 121, https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development.

Appendix J: Chip density controls

If distributed hardware were deemed politically desirable, such controls could help verify such distributions. Equally however, the relative feasibility of distributed training might create a desire for increased chip concentration in a small number of governed data centers, as each controlled facility would require further marginal costs.[426]

Overall, chip concentration limits may be politically challenging given that commercial data center construction tends to concentrate data centers in specific regions such as Northern Virginia due to the ready availability of electric power, internet backbone connections, and other infrastructure. Furthermore, for military AI chips, only very coarse-grained location verification can be expected to be tolerable to militaries (see Section 1.5.1.1), thus requiring other approaches for the regulation of training scale (see Section 2.5.4.2).

Therefore, different chip location tracking regimes may be needed for the civilian and military spheres, since location accuracy that is fine-grained enough to mitigate concerns about a dense region such as Northern Virginia may be precise enough to be intolerable for militaries. Finally, the realistic potential for large improvements in the efficiency of distributed training may make it infeasible for a governance regime to make large models infeasible via location tracking alone—although other regulatory goals might be served well by such a system.[427]

[426] Scher & Thiergart (2024).

[427] Scher & Thiergart (2024).

Appendix K: Potential approaches for device-model mating with an encrypted model

As noted in Section 4.5.2.3.6.c, it could be useful to have a technique for embedding a model in a device in a way such that downstream users cannot feasibly a) exfiltrate the model or b) repurpose the hardware. This appendix outlines one speculative approach to the problem and another potentially interesting avenue of investigation.

Hardware functionality supporting model-hardware mating might be accomplished through the inventive use of a physical unclonable function (PUF).[428] This idea has roughly four parts: 1) A physical enclosure is placed around a chip, with the enclosure material creating a sensitive PUF which is then used as the private key of the device. 2) When a model is loaded onto the device, it is encrypted using the device's PUF-backed private key.[429] 3) When the model has been installed, a physical or digital switch is thrown, blowing select fuses inside the inner (enclosed) hardware, with the effect of disallowing another model installation thereafter. 4) Inference conducted by the model requires the encryption of inputs and the decryption of outputs using the PUF-backed private key. Thus, the model should be functional, but never reside at rest or even in memory in the hardware in a way that is decryptable without the private key. Exfiltrating the private key might be possible, but might be very difficult depending on how sensitive and reliable the PUF enclosure is. One of several major unknowns with this approach is whether a PUF enclosure can reasonably withstand the normal stresses of the environments in which these devices would be placed without deforming and thus rendering the device inoperable.

A very different approach to at least part of the problem would be to employ in-memory processing and a form of encrypted memory to protect the model. It should be noted at the outset that in-memory processing for AI inference might be impossible or very impractical. However, if it is possible, then a random bit mask could be generated at the hardware level which reliably flips half the bits that are stored in non-volatile memory. An attempt to extract the model from the memory units directly would also have to extract the entire bit mask. In-memory processing would be desirable because it would prevent the model from being loaded into random access memory in order for it to be used for inference. The combination of these mechanisms would mean that the model is never available in plaintext.

[428] For readers unfamiliar with physical unclonable functions, they are essentially a way to use physical objects as private keys. Crucially for the usage described here, if you change the object even slightly, the private key is lost. A similar scheme is described in Vincent Immler et al., 'Secure Physical Enclosures from Covers with Tamper-Resistance', IACR Transactions on Cryptographic Hardware and Embedded Systems, 2019, 51–96, https://doi.org/10.13154/tches.v2019.i1.51-96.

[429] Note that being encrypted with the PUF means that only this device can produce the decryption key, and that key would be destroyed if the enclosure is tampered with.

Appendix L: Other verification mechanisms

L.1 Model-specific training licensing

A model-specific training licensing system is a licensing mechanism that is intended to be used for permitting model training. Therefore it combines a training plan—all information needed to train a model, including the code, data, and hyperparameters (see Section 4.5.1.1.1)—with a licensing system that only allows hardware to be unlocked with an appropriately signed license. With such a scheme in place, the governed hardware can only be unlocked for model training purposes if the license issuing authority attaches a valid license to a training plan. Cryptographic techniques can be used to ensure that the license and training plan are only valid together, thus allowing the licenser (or anyone verifying all licenses) to see what models are being permitted to be built.

While basic licenses such as those described in the prior section appear to be a relatively well-scoped and workable hardware mechanism that could be added to future hardware,[430] it remains less clear at this point whether such a mechanism could be designed to verifiably enforce adherence to a training plan—making this a speculative mechanism.[431] In addition to being able to enforce offline licenses, this hardware mechanism would also have to be able to enforce training plans—or enable their enforcement. Therefore, if data or hyperparameters submitted by the Prover to the training process actually differ from the commitments made in the licensed training plan, the hardware must either refuse these non-compliant operations or take some other kind of enforcement action.[432]

How could this be used for governance? Consider that a licensing authority (see Section 3.6.2) could require that a training plan be put through privacy-preserving evaluation (see Section 3.5) before it is allowed to begin. This would mean that the data, algorithm, and hyperparameters would be verified before training could begin. Given that the hardware is locked—and this could be confirmed via other channels as well (see Section 2.5.2.2)—this would then be a credible mechanism for demonstrating that the governed hardware is not conducting training without the Verifier seeing and verifying the training plans via the pre-licensing process.

[430] Aarne, Fist, and Withers, 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'; Kulp et al., 'Hardware-Enabled Governance Mechanisms'.

[431] Confidential computing can accomplish similar governance goals, but given its general-purpose nature, it might be more difficult to secure than a purpose-built mechanism.

[432] Such action could be designed to enable later inspection of the chip (via remote attestation or physically) to detect that non-compliant training information was used. An important caveat for this sort of thing is that digital errors do occur, so a reasonable rate of errors should be expected and planned for when rolling out a scheme like this to many tens of thousands of chips.

L.2 Model-specific inference licensing

Using the same approach as that described for licensing of training plans above, this mechanism would only unlock the full power of hardware for inference purposes if an appropriate license is provided along with the licensed inference plan (see Section 4.5.2.2.1). The same technical and political limitations apply to this mechanism as its cousin described above, with two changes. First, single-chip inference licensing is somewhat more likely than single chip training licenses, since a minimal inference system for a relatively small model might indeed be using a single chip. Still, licensing at the level of a pod would probably be more common and useful. Second, automatically enforcing an inference plan might be meaningfully different from enforcing a training plan. One way to make these different processes depend on the same hardware mechanism would be to have the "plan" in either case simply be a cryptographic commitment for the complete package of code and settings which contains all of the information needed to either train a model or perform inference on it. This kind of simplification might allow both inference plan enforcement and training plan enforcement to be enabled via the same mechanism.[433]

L.3 Model registry

A model registry is a database tracking crucial governance information for each model.[434] For example, the data pertaining to a single model could include its fingerprint alongside other metadata such as who created the model, what evaluations were done on the model, and the results of those evaluations, and the total compute budget of the training process including all prior models used as inputs.[435] A model registry can be useful for governance processes such as recognizing models later in time and also tracking how they have been tested. Depending on the sensitivity of the information contained in a model registry—or how crucial it is for successful verification—detailed questions might arise about its technical composition and location. This report does not investigate this particular question further, but engages in a similar discussion in Section 3.6.

L.4 Model fingerprint attestation

The goal of model fingerprint attestation is to credibly demonstrate that a given model is actually loaded into memory on specified hardware (e.g., a chip or pod). Simple versions of this are available via remote attestation and confidential computing (see Sections 2.2.4.2 and 2.2.4.4), but more specialized hardware-enabled versions are also theoretically possible and may have some advantages.

[433] The similarity between this mechanism and remote attestation is not accidental. Both mechanisms rely on the hardware root of trust to provide credibility to the attestations. Neither mechanism can fully guard against circumventions post-attestation without either a more complicated protocol or other supporting mechanisms.

[434] Elliot McKernon et al., 'AI Model Registries: A Foundational Tool for AI Governance' (arXiv, 12 October 2024), https://doi.org/10.48550/arXiv.2410.09645.

[435] The idea of a "compute graph" or similar concept mapping all compute embodied in all models is also raised in Petrie et al. (2024).

Ideally, such a mechanism could robustly demonstrate that a precise, identifiable chip (or pod) has precisely a specific model in memory. In practice, it might actually be advisable for this commitment mechanism to lean less on the hardware private key and more on computations that can only be done by a chip of at least the expected capability with the model *already in memory*. This means that even if the hardware root of trust is violated on all the chips, thus allowing the Prover to run other models on their chips—the Prover would still have to keep an equal-or greater performance chip geared up and prepared to respond to cryptographic challenges about the model that it is claiming to be running. Since loading the model into memory would take at least a number of seconds (if not much longer), it would be infeasible for hardware running a completely different model to respond to the challenge rapidly enough if the time horizon for the challenge is kept short enough. This approach would be similar in concept to a proof of work,[436] but conceptually it would be a proof of work that could only be done by a chip that is not lying about the work that it is doing. Designing a cryptographic challenge that fits the design specifications of this mechanism is an area for future work.[437]

Presuming that a neutral cluster is available, these proofs could be verified in batches there, where a copy of the model can be used. Doing these proofs *after* the challenges have taken place provides an additional level of certainty. For example, even if the Verifier's communications could all be read by the Prover, if the random challenges are generated right before they are sent, there is no way for the Prover to prepare their answers ahead of time—and even the Verifier does not know the right answer when the challenge is sent.

An expanded version of this mechanism challenges all inference chips in the Prover's inventory simultaneously, essentially amounting to a proof that the right models are loaded for all the declared chips simultaneously (thus making it infeasible for the prover to use declared chips to cover for each other). This kind of broad proof of work would cause a small system-wide slowdown for the Prover, so the workload would have to be designed to provide a credible response relatively quickly without costing much marginal computational power or time (presuming that the Prover actually has the models loaded that they say they have loaded).

[436] A proof of work allows a Prover to demonstrate that they completed a computationally costly operation. In examining a proof of work, the Verifier would have to spend extremely little computing power. The asymmetry between these two processes can allow a Prover to credibly demonstrate what they did with a portion of their total computing power, thus allowing other verification processes to focus on the remaining portion of their computing power. On this latter point, see Scher and Theirgart (2024).

[437] Roughly, what is needed is a problem which fits criteria something like the following. Provide a problem which can only feasibly be responded to by hardware of the appropriate type (or more powerful than the declared type) using an entire model (or pre-allocated shard) in memory, and which takes a maximum of about 50 ms to complete—but which does not create a major interruption for ongoing inference calculations. The overhead must be minimal. There can't be obvious ways to truncate the difficulty of the problem to allow attacks. This idea should all be developed in public and subjected to incentivized attacks (bug bounties) to find issues. If it is infeasible for current hardware to do challenges this fast, then the problem can be reframed to only use a random portion of the in-memory weights. Since this portion is randomly based on the content of the challenge, it would not be feasible for the Prover to prepare the chip memory before the challenge arrives except to actually have their model or shard in memory as expected. As the total computational burden of the challenge is brought down, the timing requirements can be tightened, with them culminating in a very tight check that costs very little, and can thus be run relatively often.

L.5 Model tenancy ledger attestation

This mechanism allows hardware to attest to the identity of all models that have been loaded into its memory. A simple version of this would have the hardware write to a nonvolatile memory store at least the model fingerprints and timestamps for all operations that load models into memory. A remote attestation call or local inspection could reveal this list of fingerprints and timestamps, thus allowing verification. Eventually this list could be quite long, so in theory it could be truncated assuming that timestamps are included. It is unclear whether this scheme can be robustly secured other than relying on the hardware private key as is done for remote attestation and confidential computing. One conceptually similar idea with the potential to be more difficult to circumvent would be fingerprint attestations as described above (Appendix L.4) that are done extremely often by a neutral system—thus demonstrating that it is infeasible that any other models are being run at scale.

Bibliography

Aarne, Onni, Tim Fist, and Caleb Withers. 'Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing'. Center for a New American Security, 2024.

Allentuck, Jack. 'Challenge Inspections in Arms Control Treaties: Any Lessons for Strengthening NPT Verification?' Brookhaven National Lab., Upton, NY (United States), 1992. https://www.osti.gov/biblio/10174104.

Anderljung, Markus, and Julian Hazell. 'Protecting Society from AI Misuse: When Are Restrictions on Capabilities Warranted?' arXiv, 29 March 2023. https://doi.org/10.48550/arXiv.2303.09377.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, et al. 'Frontier AI Regulation: Managing Emerging Risks to Public Safety'. arXiv, 11 July 2023. http://arxiv.org/abs/2307.03718.

Anthropic. 'Claude Can Now Use Tools', 30 May 2024. https://www.anthropic.com/news/tool-use-ga.

Apsey, Emily, Phil Rogers, Michael O'Connor, and Rob Nertney. 'Confidential Computing on NVIDIA H100 GPUs for Secure and Trustworthy AI'. NVIDIA Technical Blog, 3 August 2023. https://developer.nvidia.com/blog/confidential-computing-on-h100-gpus-for-secure-and-trustworthy-ai/.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. 'Racing to the Precipice: A Model of Artificial Intelligence Development'. AI & Society 31, no. 2 (2016): 201–6.

Askell, Amanda, Miles Brundage, and Gillian Hadfield. 'The Role of Cooperation in Responsible AI Development'. arXiv, 10 July 2019. http://arxiv.org/abs/1907.04534.

Aumann, Yonatan, and Yehuda Lindell. 'Security Against Covert Adversaries: Efficient Protocols for Realistic Adversaries'. In Theory of Cryptography, edited by Salil P. Vadhan, 137–56. Berlin, Heidelberg: Springer, 2007. https://doi.org/10.1007/978-3-540-70936-7_8.

Australia Group. 'Common Control Lists'. Accessed 1 March 2025. https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/common-control-lists.html.

Axelrod, Robert. The Evolution of Cooperation. Basic Books, New York, 1984.

Aytbaev, Bulat, Dmitry Grigoriev, Vladislav Lavrenchuk, and Noah C Mayhew. 'Don't Let Nuclear Accidents Scare You Away from Nuclear Power'. Bulletin of the Atomic Scientists, 31 August 2020. https://thebulletin.org/2020/08/dont-let-nuclear-accidents-scare-you-away-from-nuclear-power/.

Beimel, Amos. 'Secret-Sharing Schemes: A Survey'. In Coding and Cryptology, edited by Yeow Meng Chee, Zhenbo Guo, San Ling, Fengjing Shao, Yuansheng Tang, Huaxiong Wang, and Chaoping Xing, 6639:11–46. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. https://doi.org/10.1007/978-3-642-20901-7_2.

Baker, Mauricio. 'Nuclear Arms Control Verification and Lessons for AI Treaties'. arXiv, 8 April 2023. https://doi.org/10.48550/arXiv.2304.04123.

Baker, Mauricio, Gabriel Kulp, Oliver Marks, Miles Brundage, Lennart Heim. 'Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment'. Forthcoming.

Barez, Fazl, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, et al. 'Open Problems in Machine Unlearning for AI Safety'. arXiv, 9 January 2025. https://doi.org/10.48550/arXiv.2501.04952.

Bengio, Yoshua, Andrew Yao, Geoffrey Hinton, Zhang Ya-Qin, Stuart Russell, Gillian Hadfield, Mary Robinson, Xue Lan, et al. 'IDAIS-Venice'. International Dialogues on AI Safety, 2024. https://idais.ai/dialogue/idais-venice/.

Bengio, Yoshua, et al. 'International AI Safety Report', 29 January 2025.

———. 'AI and Catastrophic Risk'. Journal of Democracy, October 2023. https://www.journalofdemocracy.org/articles/ai-and-catastrophic-risk/.

———. 'International Scientific Report on the Safety of Advanced AI - Interim Report'. DSIT UK, 17 May 2024.

Bernauer, Thomas. The Projected Chemical Weapons Convention: A Guide to the Negotiations in the Conference on Disarmament. New York: United Nations Institute for Disarmament Research, 1990.

Bertelsen, Olga. 'Secrecy and the Disinformation Campaign Surrounding Chernobyl'. International Journal of Intelligence and CounterIntelligence 35, no. 2 (3 April 2022): 292–317. https://doi.org/10.1080/08850607.2021.2018262.

Black, James, Mattias Eken, Ryan Bain, Jacob Parakilas, Stuart Dee, Kiran Suman-Chauhan, Maria Chiara Aquilino, et al. 'Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence'. RAND, 2024.

Boulanin, Vincent, and Maaike Verbruggen. 'Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies'. Stockholm International Peace Research Institute, 2017.

Brass, Asher, and Onni Aarne. 'Location Verification for AI Chips'. Institute for AI Policy and Strategy, April 2024. https://www.iaps.ai/research/location-verification-for-ai-chips.

Brown, Robert L., and Jeffrey M. Kaplow. 'Talking Peace, Making Weapons: IAEA Technical Cooperation and Nuclear Proliferation'. Journal of Conflict Resolution 58, no. 3 (1 April 2014): 402–28. https://doi.org/10.1177/0022002713509052.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, et al. 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims'. arXiv, 20 Apr 2020. https://arxiv.org/abs/2004.07213.

Bucknall, Ben, Robert F. Trager, and Michael A. Osborne. 'Position: Ensuring Mutual Privacy Is Necessary for Effective External Evaluation of Proprietary AI Systems'. arXiv, 3 March 2025. https://doi.org/10.48550/arXiv.2503.01470.

Calderaro, Andrea, and Anthony J. S. Craig. 'Transnational Governance of Cybersecurity: Policy Challenges and Global Inequalities in Cyber Capacity Building'. Third World Quarterly 41, no. 6 (2 June 2020): 917–38. https://doi.org/10.1080/01436597.2020.1729729.

Carnegie, Allison, and Austin Carson. 'The Disclosure Dilemma: Nuclear Intelligence and International Organizations'. American Journal of Political Science 63, no. 2 (April 2019): 269–85. https://doi.org/10.1111/ajps.12426.

Casper, Stephen, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, et al. 'Black-Box Access Is Insufficient for Rigorous AI Audits'. arXiv, 25 January 2024. https://doi.org/10.48550/arXiv.2401.14446.

Center for AI Safety. 'Statement on AI Risk', 30 May 2023. https://www.safe.ai/statement-on-ai-risk.

Center for Security and Emerging Technology, Saif Khan, and Alexander Mann. 'AI Chips: What They Are and Why They Matter'. Center for Security and Emerging Technology, April 2020. https://doi.org/10.51593/20190014.

Chadefaux, Thomas. 'Bargaining over Power: When Do Shifts in Power Lead to War?' International Theory 3, no. 2 (2011): 228–53.

Che, Wenjie, Fareena Saqib, and Jim Plusquellic. 'PUF-Based Authentication'. In Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 337–44. ICCAD '15. Austin, TX, USA: IEEE Press, 2015.

Cheng, Deric. 'Evaluating An AI Chip Registration Policy'. Convergence Analysis, April 2024.

Chevrier, Marie Isabelle. 'Verifying the Unverifiable: Lessons from the Biological Weapons Convention'. Politics and the Life Sciences 9, no. 1 (1990): 93–105. https://doi.org/10.1017/S073093840001025X.

Choi, Dami, Yonadav Shavit, and David Duvenaud. 'Tools for Verifying Neural Models' Training Data'. arXiv, 2 July 2023. https://doi.org/10.48550/arXiv.2307.00682.

Christian, Brian. The Alignment Problem: Machine Learning and Human Values. WW Norton & Company, 2020.

CloudFlare. 'The DNSSEC Root Signing Ceremony'. Accessed 2 October 2024. https://www.cloudflare.com/dns/dnssec/root-signing-ceremony/.

Coe, Andrew J., and Jane Vaynman. 'Why Arms Control Is so Rare'. American Political Science Review 114, no. 2 (2020): 342–55.

Coe, Andrew. 'Costly Peace: A New Rationalist Explanation for War', 2011.

Cohen, Michael K., Noam Kolt, Yoshua Bengio, Gillian K. Hadfield, and Stuart Russell. 'Regulating Advanced Artificial Agents'. Science, 5 April 2024. https://doi.org/10.1126/science.adl0625.

Coming to Terms with Security: A Handbook on Verification and Compliance. Geneva: United Nations Institute for Disarmament Research, 2003.

Dalrymple, David 'davidad'. 'Safeguarded AI: Constructing Guaranteed Safety'. Advanced Research and Invention Agency, UK Government, 2024. https://www.aria.org.uk/media/3nhijno4/aria-safeguarded-ai-programme-thesis-v1.pdf.

Dalrymple, David 'davidad', Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, et al. 'Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems'. arXiv, 8 July 2024. https://doi.org/10.48550/arXiv.2405.06624.

Daniels, Mario. 'Controlling Knowledge, Controlling People: Travel Restrictions of US Scientists and National Security'. Diplomatic History 43, no. 1 (2019): 57–82.

Deng, Jiangyi, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. 'SOPHON: Non-Fine-Tunable Learning to Restrain Task Transferability For Pre-Trained Models'. arXiv.org, 19 April 2024. https://arxiv.org/abs/2404.12699v1.

Dennis, Claire, Sam Manning, Stephen Clare, Ben Garfinkel, Boxi Wu, Jake Okechukwu Effoduh, Chinasa T Okolo, Lennart Heim, and Katya Klinova. 'Options and Motivations for International AI Benefit Sharing'. Centre for the Governance of AI, 2025.

Dennis, Claire, Stephen Clare, Rebecca Hawkins, Morgan Simpson, Eva Behrens, Gillian Diebold, Zaheed Kara, et al. 'What Should Be Internationalised in AI Governance?' Oxford Martin AI Governance Initiative, 2024.

Ding, Jeffrey. ‘Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies’. European Journal of International Relations, 27 April 2024, 13540661241246622. https://doi.org/10.1177/13540661241246622.

‘Economic Impact Assessment of the Global Minimum Tax’. Organisation for Economic Co-operation and Development, January 2024.

‘Equipment, Software And Technology Annex’. Missile Technology Control Regime, 14 March 2024.

European Commission. ‘General-Purpose AI Models in the AI Act – Questions & Answers’, 20 November 2024. https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers.

‘Evals’, 19 May 2023. https://github.com/openai/evals.

Financial Action Task Force. ‘Mutual Evaluations’. Accessed 13 July 2023. https://www.fatf-gafi.org/en/topics/mutual-evaluations.html.

Garg, Sanjam, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Guru-Vamsi Policharla, and Mingyuan Wang. ‘Experimenting with Zero-Knowledge Proofs of Training’. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, 1880–94. Copenhagen Denmark: ACM, 2023. https://doi.org/10.1145/3576915.3623202.

Georgiev, Petko, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, et al. ‘Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context’. arXiv, 8 August 2024. http://arxiv.org/abs/2403.05530.

Gibbons, Rebecca Davis. ‘Supply to Deny: The Benefits of Nuclear Assistance for Nuclear Nonproliferation’. Journal of Global Security Studies 5, no. 2 (1 April 2020): 282–98. https://doi.org/10.1093/jogss/ogz059.

Gottemoeller, Rose. ‘Looking Back: The Intermediate-Range Nuclear Forces Treaty’. Arms Control Today, 2007. https://www.armscontrol.org/act/2007-06/looking-back-intermediate-range-nuclear-forces-treaty.

GOV.UK. ‘The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023’. Accessed 2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

Gupta, Ritwik, Leah Walker, Rodolfo Corona, Stephanie Fu, Suzanne Petryk, Janet Napolitano, Trevor Darrell, and Andrew W. Reddie. ‘Data-Centric AI Governance: Addressing the Limitations of Model-Focused Policies’. arXiv, 25 September 2024. https://doi.org/10.48550/arXiv.2409.17216.

Grunewald, Erich. ‘Are Consumer GPUs a Problem for US Export Controls?’ Institute for AI Policy and Strategy, May 2024.

‘Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System (Annex III)’. United Nations Office of Disarmament Affairs: Convention on Certain Conventional Weapons (CCW) - Group of Governmental Experts on Lethal Autonomous Weapons Systems, 2019.

Hales, Steven D. ‘Thinking Tools: You Can Prove a Negative’. Think 4, no. 10 (2005): 109–12.

Halstead, John. ‘Managing Risks from AI-Enabled Biological Tools’. Centre for the Governance of AI (blog), 5 August 2024. https://www.governance.ai/analysis/managing-risks-from-ai-enabled-biological-tools.

Harahan, Joseph P. On-Site Inspections Under The INF Treaty, A History of the On-Site Inspection Agency and Treaty Implementation, 1988-1991. On-Site Inspection Agency, United States Department of Defense, 1993.

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A Osborne, and Noa Zilberman. 'Governing Through The Cloud: The Intermediary Role Of Compute Providers In AI Regulation'. Oxford Martin AI Governance Initiative, March 2024.

Heim, Lennart. 'The Rise of DeepSeek: What the Headlines Miss'. RAND Corporation, 28 January 2025. https://www.rand.org/pubs/commentary/2025/01/the-rise-of-deepseek-what-the-headlines-miss.html.

———. 'AI Benefit Sharing Options'. Lennart Heim (blog), 28 September 2024. https://blog.heim.xyz/ai-benefit-sharing-options/.

———. 'The Case for Pre-Emptive Authorizations for AI Training'. Lennart Heim (blog), 10 June 2023. https://blog.heim.xyz/the-case-for-pre-emptive-authorizations/.

———. 'Limitations of Satellite Imagery Analysis for AI-Specific Data Centers'. Lennart Heim (blog), 13 September 2024. https://blog.heim.xyz/limitations-of-satellite-imagery/.

Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. 'Algorithmic Progress in Language Models'. arXiv, 9 March 2024. https://doi.org/10.48550/arXiv.2403.05812.

Ho, Lewis, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, et al. 'International Institutions for Advanced AI'. arXiv, 11 July 2023. http://arxiv.org/abs/2307.04699.

Hooker, Sara. 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv, 29 July 2024. https://doi.org/10.48550/arXiv.2407.05694.

'Huawei Cyber Security Evaluation Centre (HCSEC) Oversight Board Annual Report 2021'. Huawei Cyber Security Evaluation Centre Oversight Board, 20 July 2021. https://assets.publishing.service.gov.uk/media/60f6b6be8fa8f50c7a1b9ffd/2021_HCSEC_OB_REPORT_FINAL__1_.pdf.

'IAEA Safety Standards: Functions and Processes of the Regulatory Body for Safety'. International Atomic Energy Agency, 2018.

Immler, Vincent, Johannes Obermaier, Kuan Kuan Ng, Fei Xiang Ke, JinYu Lee, Yak Peng Lim, Wei Koon Oh, Keng Hoong Wee, and Georg Sigl. 'Secure Physical Enclosures from Covers with Tamper-Resistance'. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2019, 51–96. https://doi.org/10.13154/tches.v2019.i1.51-96.

Intel. 'Intel On Demand'. Accessed 13 March 2025. https://www.intel.com/content/www/us/en/products/docs/ondemand/overview.html.

Islam, Md Nazmul, and Sandip Kundu. 'Enabling IC Traceability via Blockchain Pegged to Embedded PUF'. ACM Transactions on Design Automation of Electronic Systems 24, no. 3 (31 May 2019): 1–23. https://doi.org/10.1145/3315669.

Jaghouar, Sami, Jack Min Ong, Manveer Basra, Fares Obeid, Jannik Straube, Michael Keiblinger, Elie Bakouch, et al. 'INTELLECT-1 Technical Report'. arXiv, 2 December 2024. https://doi.org/10.48550/arXiv.2412.01152.

Jauernig, Patrick, Ahmad-Reza Sadeghi, and Emmanuel Stapf. 'Trusted Execution Environments: Properties, Applications, and Challenges'. IEEE Security & Privacy 18, no. 2 (March 2020): 56–60. https://doi.org/10.1109/MSEC.2019.2947124.

Jennings, Brian, Sean Alcorn, Bonnie Canion, Joshua Cunningham, Monica Maceira, and Michael J. Willis. 'Advanced Portal Monitoring for Arms Control Treaty Verification'. Oak Ridge National Laboratory, 2024. https://www.osti.gov/servlets/purl/2472697.

Johnson, James. 'AI-Security Dilemma: Insecurity, Mistrust, and Misperception under the Nuclear Shadow'. In AI and the Bomb: Nuclear Strategy and Risk in the Digital Age. Oxford University Press, 2023. https://doi.org/10.1093/oso/9780192858184.003.0005.

Karnofsky, Holden. 'Some Background on Our Views Regarding Advanced Artificial Intelligence'. Open Philanthropy Project (Blog), Open Philanthropy Project, 2016. https://www.openphilanthropy.org/research/some-background-on-our-views-regarding-advanced-artificial-intelligence/.

Kissinger, Henry, Eric Schmidt, and Craig Mundie. Genesis: Artificial Intelligence, Hope, and the Human Spirit. Little Brown and Company, 2024.

Koblentz, Gregory D. 'Saddam versus the Inspectors: The Impact of Regime Security on the Verification of Iraq's WMD Disarmament'. Journal of Strategic Studies 41, no. 3 (16 April 2018): 372–409. https://doi.org/10.1080/01402390.2016.1224764.

Kulp, Gabriel, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman. 'Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090'. RAND Corporation, 18 January 2024. https://www.rand.org/pubs/working_papers/WRA3056-1.html.

Lagerros, Jacob. Presentation at the Paris AI Security Forum, 9 Feb 2025.

Lencucha, Raphael, and Shashika Bandara. 'Trust, Risk, and the Challenge of Information Sharing during a Health Emergency'. Globalization and Health 17, no. 1 (18 February 2021): 21. https://doi.org/10.1186/s12992-021-00673-9.

Liang, Jiacheng, Ren Pang, Changjiang Li, and Ting Wang. 'Model Extraction Attacks Revisited'. arXiv, 8 December 2023. https://doi.org/10.48550/arXiv.2312.05386.

Lindell, Yehuda. 'Secure Multiparty Computation'. Communications of the ACM 64, no. 1 (January 2021): 86–96. https://doi.org/10.1145/3387108.

Maas, Matthijs M., and José Jaime Villalobos. 'International AI Institutions: A Literature Review of Models, Examples, and Proposals'. SSRN Scholarly Paper. Rochester, NY, 22 September 2023. https://doi.org/10.2139/ssrn.4579773.

Manheim, David, Sammy Martin, Mark Bailey, Mikhail Samin, and Ross Greutzmacher. 'The Necessity of AI Audit Standards Boards'. arXiv, 11 April 2024. https://doi.org/10.48550/arXiv.2404.13060.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, et al. 'The AI Index 2025 Annual Report'. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2025.

McKernon, Elliot, Gwyn Glasser, Deric Cheng, and Gillian Hadfield. 'AI Model Registries: A Foundational Tool for AI Governance'. arXiv, 12 October 2024. https://doi.org/10.48550/arXiv.2410.09645.

'Memorandum of Agreement Regarding the Implementation of the Verification Provisions of the Treaty Between the United States of America and the Union of Soviet Socialist Republics on the Elimination of Their Intermediate-Range and Shorter-Range Missiles', 21 December 1989. https://nuke.fas.org/control/inf/text/inf-mouanx.htm.

Meserole, Chris. 'Artificial Intelligence and the Security Dilemma'. Lawfare, 4 November 2018. https://www.lawfaremedia.org/article/artificial-intelligence-and-security-dilemma.

'METR: Model Evaluation and Threat Research'. Accessed 30 September 2024. https://metr.org/.

Microsoft. ‘Transparency Centers’, 29 October 2024. https://learn.microsoft.com/en-us/security/engineering/contenttransparencycenters.

Miller, Chris. Chip War: The Fight for the World’s Most Critical Technology. Simon and Schuster, 2022.

Miller, Nicholas L. ‘Why Nuclear Energy Programs Rarely Lead to Proliferation’. International Security 42, no. 2 (1 November 2017): 40–77. https://doi.org/10.1162/ISEC_a_00293.

Miotti, Andrea, and Akash Wasil. ‘Taking Control: Policies to Address Extinction Risks from Advanced AI’. arXiv, 31 October 2023. http://arxiv.org/abs/2310.20563.

Mithril Security. ‘AICert’. Accessed 2 October 2024. https://www.mithrilsecurity.io/aicert.

Muszyński-Sulima, Wawrzyniec. ‘Cold War in Space: Reconnaissance Satellites and US-Soviet Security Competition’. European Journal of American Studies 18, no. 2 (30 June 2023). https://doi.org/10.4000/ejas.20427.

Nasu, Hitoshi. ‘State Secrets Law and National Security’. International & Comparative Law Quarterly 64, no. 2 (2015): 365–404.

Nertney, Rob. ‘Confidential Compute on NVIDIA Hopper H100’. NVIDIA, 25 July 2023. https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/HCC-Whitepaper-v1.0.pdf.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. ‘Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models’. RAND Corporation, 30 May 2024. https://www.rand.org/pubs/research_reports/RRA2849-1.html.

Nuclear Suppliers Group. ‘Guidelines’. Accessed 1 March 2025. https://www.nuclearsuppliersgroup.org/index.php/en/guidelines/nsg-guidelines.

NVIDIA. ‘NVIDIA Blackwell Architecture’. Accessed 13 March 2025. https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/.

NVIDIA. ‘NVIDIA GB200 NVL72 GPU – Optimized for AI and Data Centers’. Accessed 15 March 2025. https://www.nvidia.com/en-gb/data-center/gb200-nvl72/.

OpenMined Blog. ‘When Data Sharing Is a Problem, PySyft 0.9 Is the Solution’, 6 August 2024. https://blog.openmined.org/announcing-pysyft-09/.

Ord, Toby. ‘Inference Scaling Reshapes AI Governance’, 12 February 2025. https://www.tobyord.com/writing/inference-scaling-reshapes-ai-governance.

Peng, Bowen, Jeffrey Quesnelle, and Diederik P. Kingma. ‘DeMo: Decoupled Momentum Optimization’. arXiv, 29 November 2024. https://doi.org/10.48550/arXiv.2411.19870.

Petrie, James. ‘Near-Term Enforcement of AI Chip Export Controls Using A Firmware-Based Design for Offline Licensing’. arXiv, 28 May 2024. https://doi.org/10.48550/arXiv.2404.18308.

Petrie, James, Onni Aarne, Nora Ammann, and David Dalrymple. ‘Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees’, 23 August 2024.

Philippe, Sébastien, Alexander Glaser, and Edward W. Felten. ‘A Cryptographic Escrow for Treaty Declarations and Step-by-Step Verification’. Science & Global Security 27, no. 1 (2 January 2019): 3–14. https://doi.org/10.1080/08929882.2019.1573483.

Philippe, Sébastien, Robert J. Goldston, Alexander Glaser, and Francesco d’Errico. ‘A Physical Zero-Knowledge Object-Comparison System for Nuclear Warhead Verification’. Nature Communications 7, no. 1 (20 September 2016): 12890. https://doi.org/10.1038/ncomms12890.

Pilz, Konstantin, Lennart Heim, and Nicholas Brown. ‘Increased Compute Efficiency and the Diffusion of AI Capabilities’. arXiv, 13 February 2024. https://doi.org/10.48550/arXiv.2311.15377.

Poast, Paul. ‘Issue Linkage and International Cooperation: An Empirical Investigation’. Conflict Management and Peace Science 30, no. 3 (2013): 286–303.

Pouget, Hadrien, Claire Dennis, Jon Bateman, Robert F. Trager, Renan Araujo, Belinda Cleeland, Malou Estier, et al. ‘The Future of International Scientific Assessments of AI’s Risks’. Oxford Martin AI Governance Initiative, August 2024.

Powell, Robert. ‘Guns, Butter, and Anarchy’. American Political Science Review 87, no. 1 (1993): 115–32.

Reuel, Anka, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, et al. ‘Open Problems in Technical AI Governance’. arXiv, 20 July 2024. https://arxiv.org/abs/2407.14981.

Roberts, Guy B. Arms Control without Arms Control: The Failure of the Biological Weapons Convention Protocol and a New Paradigm for Fighting the Threat of Biological Weapons. USAF Institute for National Security Studies, 2003.

Rockwood, Laura. ‘IAEA Safeguards: Correctness and Completeness of States’ Safeguards Declarations’. In Nuclear Law: The Global Debate, 205–22. The Hague: T.M.C. Asser Press, 2022. https://doi.org/10.1007/978-94-6265-495-2_10.

Rueckert, George. On-Site Inspection in Theory and Practice: A Primer on Modern Arms Control Regimes. Westport, Conn.: Praeger, 1998.

Safire, William. ‘On Language; Weapons Of Mass Destruction’. The New York Times Magazine, 19 April 1998. https://www.nytimes.com/1998/04/19/magazine/on-language-weapons-of-mass-destruction.html.

Sandbrink, Jonas B. ‘Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools’. arXiv, 23 December 2023. https://doi.org/10.48550/arXiv.2306.13952.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O’Keefe, et al. ‘Computing Power and the Governance of Artificial Intelligence’. arXiv, 13 February 2024. http://arxiv.org/abs/2402.08797.

Schelling, Thomas C. The Strategy of Conflict. Harvard University Press, 1960.

Scharre, Paul, and Megan Lamberth. ‘Artificial Intelligence and Arms Control’. The Center for a New American Security, 12 October 2022. https://www.cnas.org/publications/reports/artificial-intelligence-and-arms-control.

Scher, Aaron, and Lisa Thiergart. ‘Mechanisms to Verify International Agreements About AI Development’. Machine Intelligence Research Institute, November 2024. https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development.

‘Secure Enclaves for AI Evaluation’. OpenMined, 12 December 2024. https://openmined.org/blog/secure-enclaves-for-ai-evaluation/.

Shah, Agam. ‘Nvidia Shipped 3.76 Million Data-Center GPUs in 2023, According to Study’. HPCwire, 10 June 2024. https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/.

Shamsujjoha, Md, Qinghua Lu, Dehai Zhao, and Liming Zhu. ‘Swiss Cheese Model for AI Safety: A Taxonomy and Reference Architecture for Multi-Layered Guardrails of Foundation Model Based Agents’. In 2025 IEEE 22nd International Conference on Software Architecture (ICSA), 37–48. IEEE, 2025. https://ieeexplore.ieee.org/abstract/document/10978931/.

Shavit, Yonadav. ‘What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring’. arXiv, 20 March 2023. https://doi.org/10.48550/arXiv.2303.11341.

Shirriff, Ken. 'Standard Cells: Looking at Individual Gates in the Pentium Processor', July 2024. http://www.righto.com/2024/07/pentium-standard-cells.html.

Sommerhalder, Maria. 'Hardware Security Module'. In Trends in Data Protection and Encryption Technologies, edited by Valentin Mulder, Alain Mermoud, Vincent Lenders, and Bernhard Tellenbach, 83–87. Cham: Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-33386-6_16.

South, Tobin, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, and Alex 'Sandy' Pentland. 'Verifiable Evaluations of Machine Learning Models Using zkSNARKs'. arXiv, 22 May 2024. https://doi.org/10.48550/arXiv.2402.02675.

Stafford, Eoghan, Robert F Trager, and Allan Dafoe. 'Safety Not Guaranteed: International Races for Risky Technologies', November 2022. https://cdn.governance.ai/International_Races_for_Risky_Technologies_DRAFT_NOV_2022.pdf.

'Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet'. Paris AI Action Summit, 11 February 2025. https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet.

Sun, Haochen, Jason Li, and Hongyang Zhang. 'zkLLM: Zero Knowledge Proofs for Large Language Models'. arXiv, 24 April 2024. https://doi.org/10.48550/arXiv.2404.16109.

Tamirisa, Rishub, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, et al. 'Tamper-Resistant Safeguards for Open-Weight LLMs'. arXiv, 8 August 2024. https://doi.org/10.48550/arXiv.2408.00761.

Thadani, Akhil, and Gregory C. Allen. 'Mapping the Semiconductor Supply Chain: The Critical Role of the Indo-Pacific Region'. Center for Strategic and International Studies, 30 May 2023. https://www.csis.org/analysis/mapping-semiconductor-supply-chain-critical-role-indo-pacific-region.

Toivanen, Henrietta N. 'The Significance of Strategic Foresight in Verification Technologies: A Case Study of the INF Treaty'. Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2017. https://www.osti.gov/biblio/1502006.

Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, and Seán Ó hÉigeartaigh. 'International Governance of Civilian AI: A Jurisdictional Certification Approach'. Oxford Martin AI Governance Initiative, 2023.

Trager, Robert F., Paolo Bova, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe. 'Safety-Performance Tradeoff Model: Exploring Safety Insights in AI Competition'. Modeling Cooperation, December 2022. https://spt.modelingcooperation.com/.

Trask, Andrew, Emma Bluemke, Teddy Collins, Eric Drexler, Claudia Ghezzou Cuervas-Mons, Iason Gabriel, Allan Dafoe, William Isaac, and Ben Garfinkel. 'Beyond Privacy Trade-Offs with Structured Transparency'. arXiv, 2020. https://doi.org/10.48550/arXiv.2012.08347.

Trask, Andrew, and Irina Bejan. 'Privacy, Security, and Innovation – Friends Not Foes'. Center for Security and Emerging Technology. Accessed 24 January 2025. https://cset.georgetown.edu/event/privacy-security-and-innovation-friends-not-foes/.

Tuyls, Pim, and Boris Škorić. 'Strong Authentication with Physical Unclonable Functions'. In Security, Privacy, and Trust in Modern Data Management, edited by Milan Petković and Willem Jonker, 133–48. Berlin, Heidelberg: Springer, 2007. https://doi.org/10.1007/978-3-540-69861-6_10.

United States Office of the Assistant Secretary of Defense. 'Strategic Arms Reduction Talks (START) Treaty', 20 November 1991. https://www.acq.osd.mil/asda/ssipm/sdc/tc/start1/start1-aaa/START1lett-sub.html.

Vaynman, Jane, and Tristan A. Volpe. 'Dual Use Deception: How Technology Shapes Cooperation in International Relations'. International Organization 77, no. 3 (March 2023): 599–632. https://doi.org/10.1017/S0020818323000140.

'Verifying LAWS Regulation - Opportunities and Challenges'. International Panel on the Regulation of Autonomous Weapons, 2019. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-77413-1.

Waiwitlikhit, Suppakit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, and Daniel Kang. 'Trustless Audits without Revealing Data or Models'. arXiv, 6 April 2024. https://doi.org/10.48550/arXiv.2404.04500.

Walter, Barbara F. Committing to Peace: The Successful Settlement of Civil Wars. Princeton University Press, 2002.

Wasil, Akash, Lukas Berglund, Tom Reed, Miro Plueckebaum, and Everett Smith. 'Understanding Frontier AI Capabilities and Risks through Semi-Structured Interviews'. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 1 July 2024. https://doi.org/10.2139/ssrn.4881729.

Wasil, Akash R., Tom Reed, Jack William Miller, and Peter Barnett. 'Verification Methods for International AI Agreements'. arXiv.org, 28 August 2024. https://arxiv.org/abs/2408.16074v1.

Weij, Teun van der, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. 'AI Sandbagging: Language Models Can Strategically Underperform on Evaluations'. arXiv, 6 February 2025. https://doi.org/10.48550/arXiv.2406.07358.

Zaidi, Waqar H. ' "Aviation Will Either Destroy or Save Our Civilization": Proposals for the International Control of Aviation, 1920—45'. Journal of Contemporary History 46, no. 1 (1 January 2011): 150–78. https://doi.org/10.1177/0022009410375257.

Zhu, Sally, Ahmed Ahmed, Rohith Kuditipudi, and Percy Liang. 'Independence Tests for Language Models'. arXiv, 17 February 2025. https://doi.org/10.48550/arXiv.2502.12292.

———. Technological Internationalism and World Order. Cambridge University Press, 2021.

Ziosi, Marta, Claire Dennis, Robert Trager, Ben Bucknall, Simeon Campos, Charles Martinet, Adam L Smith, and Merlin Stein. 'AISIs' Roles in Domestic and International Governance', 2024.