Verification methods for international AI agreements

Akash R. Wasil[1,2] **Tom Reed**[2] **Jack William Miller**[2] **Peter Barnett**[3]
[1]Georgetown University [2]University of Cambridge, ERA AI Fellowship [3]Independent
aw1404@georgetown.edu

Abstract

What techniques can be used to verify compliance with international agreements about advanced AI development? In this paper, we examine 10 verification methods that could detect two types of potential violations: unauthorized AI training (e.g., training runs above a certain FLOP threshold) and unauthorized data centers. We divide the verification methods into three categories: (a) national technical means (methods requiring minimal or no access from suspected non-compliant nations), (b) access-dependent methods (methods that require approval from the nation suspected of unauthorized activities), and (c) hardware-dependent methods (methods that require rules around advanced hardware). For each verification method, we provide a description, historical precedents, and possible evasion techniques. We conclude by offering recommendations for future work related to the verification and enforcement of international AI governance agreements.

1 Introduction

The development of advanced artificial intelligence poses major global security risks. Significant threats include the potential for pervasive surveillance, the development of autonomous weapons, and misuse by malicious actors. Some of the most concerning risks stem from loss of control and misalignment (Bengio et al., 2024; Bostrom, 2014; Ngo et al., 2022). A sufficiently powerful misaligned AI system could autonomously act against human interests following an objective function which does not capture human values (Pan et al., 2022). There is a great amount of uncertainty around what kinds of safeguards will be necessary to prevent misalignment (Gabriel, 2020). Many experts believe that safeguards may require many years or decades of concerted research effort.

AI risks are exacerbated by race dynamics — companies are rapidly progressing in the hope of being the first to develop artificial superintelligence (Armstrong et al., 2016; Hogarth, 2023). In the context of an AI race, nations may not have sufficient time to carefully and cautiously develop or evaluate such safeguards.

International agreements could help avoid or mitigate a race between nations. Even though governments are at early stages of understanding AI risks, key figures in the United States and China have already acknowledged concerns about AI global security risks and expressed interest in global governance approaches Wasil and Durgin (2024). As governments become more aware of AI risks, they may become interested in global governance strategies that curb these race dynamics. Alternatively, nations might agree to cede the development of advanced AI to a joint international project. The international institution would be responsible for carrying out certain forms of advanced AI development, which would be illegal outside the context of this secure joint project (Hogarth, 2023).

International agreements require verification. Nations might be much more likely to form international agreements around rules that they can reliably verify (Fearon, 1995; Baker, 2023). By "verify", we mean that nations would be able to detect non-compliance with agreements. Ideally, verification methods (methods used to detect non-compliance) would provide **early and reliable warning signs**. "Early", in that non-compliance could be detected relatively quickly (before a nation achieved

2nd Workshop on Regulatable ML at NeurIPS 2024.

any meaningful unauthorized advantage in advanced AI development), and "reliable" in that non-compliance would be very likely to be detected.[1]

In this paper, we provide an overview of verification methods for international AI agreements. We begin by outlining the potential targets such an international agreement. We then outline 10 verification methods. For each verification method, we provide a description, some precedents for how the verification method has been used in the past, and an example evasion technique to illustrate how an adversary could attempt to circumvent the method. Finally, we discuss limitations of the verification methods and directions for future work.[2]

2 What to verify: Unauthorized AI development and unauthorized data centers

An international agreement on AI could take many forms, depending on how the technology and its associated risks evolve. In scenarios where continued AI development leads to substantial acknowledged global security risks, we anticipate that verification methods would need to be capable of detecting two primary types of potential violations.

Unauthorized data centers. International governance of AI could plausibly set restrictions on the form, size, quantity, and location of large-scale computing facilities. Verification methods would therefore be needed to detect the construction or operation of data centers that violate these agreed-upon standards. **Unauthorized training runs**. An effective international system for governing AI would likely include restrictions on the scale and characteristics of AI development. Beyond detecting unauthorised data centers, methods to verify that *known* data centers are compliant with agreed-upon standards would also be necessary. For example, an agreement might stipulate that AI training runs should not exceed a certain FLOP threshold (Heim, 2024), use specific types of training data, or employ certain training algorithms. Verification methods would be needed to detect whether AI development activities occurring within facilities violate such standards.

3 Methodology

Our process for compiling verification methods involved a few steps: (a) reviewing relevant literature on AI verification and international AI governance, (b) reviewing relevant literature on verification methods for agreements in other fields (e.g., nuclear security, biosecurity, arms control), and (c) conducting informal interviews with experts in technical AI governance. Through this process, we identified 10 verification methods. For each verification method, we examined its application in other fields to inform our description of how the method could be used in the context of AI disagreements and to inform our section about the method's precedent in other fields. We also grouped the methods into categories based on the circumstances in which they could be implemented: universally (national technical means), only in cases where a nation provides access (access-dependent), or only in cases in which nations have agreed to rules around the design of advanced hardware (hardware-dependent). Our process is *not* intended to be systematic and our work should not be considered a comprehensive overview of verification methods. Rather, it is meant to serve as an initial step toward better understanding a set of specific verification methods and their limitations.

Defining "verification method". In this piece, a verification method is **a method that could directly be used to detect defection or non-compliance from an agreement.** That is, we assume an adversarial setup in which one party is explicitly attempting to "hide" unauthorized data centers or unauthorized AI training.

Categorizing verification methods. Some verification methods can be implemented without any buy-in from nations suspected of non-compliance, some verification methods require cooperation or authorization from the suspected nation, and some verification methods require cooperation from hardware manufacturers. These distinctions are useful for determining which verification methods

[1] Early and reliable warning signs have also been discussed in the context of international agreements for nuclear security (see Barnard and Acheson (1946), specifically pages 15 to 34.

[2] Readers should also note that there is a forthcoming report by authors at the RAND Corporation that aims to provide a detailed examination of verification methods, analyze trade-offs with various methods, and discuss their technical implementation.

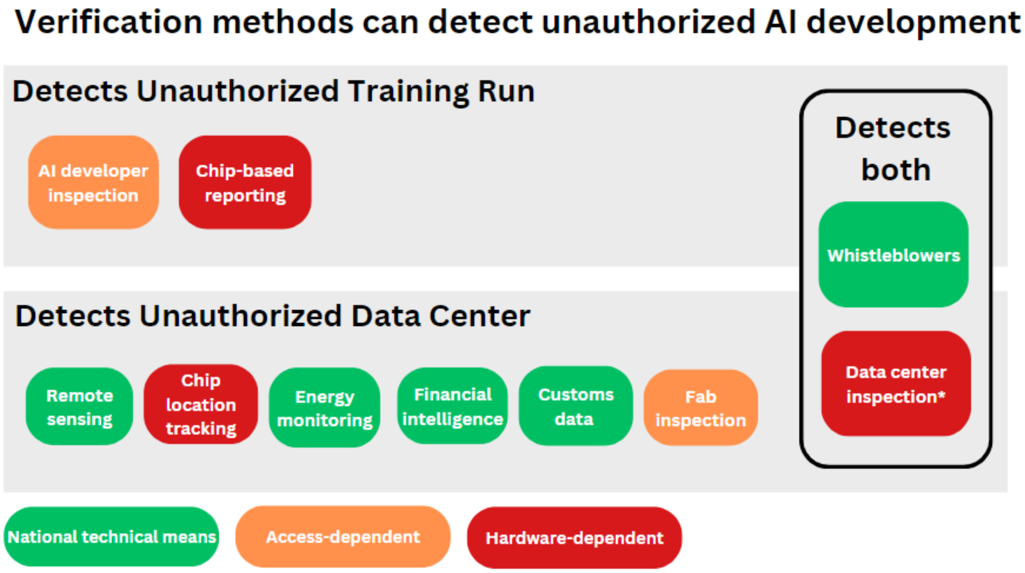


Figure 1: Verification methods can help detect unauthorized training runs and unauthorized data centers. Some methods require authorization from the country suspected of engaging in unauthorized activities, and some methods require hardware measures.

might be feasible under various circumstances. Thus, we divide verification methods into three categories: (a) **national technical means** (methods that can be implemented without the approval of individual nations; we borrow this term from the security literature, see U.S. Department of State (2001)), (b) **access-dependent verification methods** (methods that require international agreements that include the suspected nation), and (c) **hardware-dependent verification methods** (methods that require international agreements that include major hardware manufacturers). See Figure 1 and **??** for a visual summary of the verification methods.

3.1 National Technical Means

National technical means offer a valuable starting point for verifying compliance with AI governance agreements. Nations already have extensive experience using these methods to verify compliance with other kinds of international agreements. These methods can plausibly be used to detect large-scale AI infrastructure and unusual patterns in energy consumption, hardware imports, and financial transactions. However, these methods have important limitations. In particular, adversaries could attempt to disguise data centers as other high-energy facilities like power plants, or when compute is distributed across multiple smaller sites.

3.1.1 Remote sensing

Remote sensing techniques, including satellite imagery and other forms of aerial observation, can detect potential undeclared data centers using visual, infrared and other electromagnetic signatures. Advanced commercial satellites, which can achieve sub-meter resolutions (Statista, 2022), could identify specialized cooling units and the movement of computing equipment.

Infrared imaging is particularly promising for detecting concealed data centers, as GPUs and other computing hardware generate significant heat signatures (Yuan et al., 2023) that are difficult to mask. This could reveal large-scale computing facilities even when visually concealed.

Recent advancements in machine learning have further enhanced remote sensing capabilities for verification. Drawing from nuclear non-proliferation efforts (Rutkowski and Niemeyer, 2020), AI-driven approaches such as supervised and unsupervised classification techniques can be applied to remotely sensed data. These methods could significantly improve the identification and monitoring of potential AI development facilities without requiring on-site access, bolstering national technical means for AI governance verification.

While remote sensing can be used without formal agreements, international commitments similar to START could enhance its effectiveness by ensuring non-interference and facilitating data exchange (U.S. Department of State, 2023).

Precedent. The International Atomic Energy Agency (IAEA) routinely employs satellite imagery to evaluate state-provided information about nuclear activities and to plan inspections (Baker, 2023).

Non-state actors have also demonstrated the power of commercial imagery; for example, the Open Nuclear Network used it to reveal maintenance and possible expansion at China's Lop Nur nuclear testing site, including new tunneling and drilling activities (Open Nuclear Network, 2024).

3.1.2 Whistleblowers

Insiders with knowledge of undeclared facilities or operations could provide valuable information not detectable through external means. Potential whistleblowers include employees, contractors, or local residents aware of suspicious activities. Governments could incentivize whistleblowing by (1) establishing robust protection frameworks specifically for AI and technology sector, (2) offering financial incentives for verified information, (3) creating secure, anonymous reporting channels, (4) providing legal support and job protection, and (5) developing international cooperation for cross-border whistleblower protection (Loyens and Vandekerckhove, 2018).

It is important to note that incentivization alone may not be sufficient to ensure the effectiveness of whistleblower schemes, given that determined adversaries might attempt to physically or digitally block employees from contacting a verifying authority. One possible solution to this limitation is to implement regular in-person communication with employees, such as through semi-structured interviews (Wasil et al., 2024a).

Precedent. The SEC Whistleblower Program, established under the Dodd-Frank Act in 2010, created a system for reporting securities violations (U.S. Securities and Exchange Commission, 2017). It includes strong protections and incentives like monetary awards for whistleblowers, who are entitled to anywhere between 10-30% of the sanctions resulting from their information (Reuters, 2018). For example, in 2016, three whistleblowers revealed Merrill Lynch's misuse of up to $58 billion daily in customer funds, leading to a $415 million settlement and $83 million in whistleblower awards (Securities and Exchange Commission (SEC), 2016; Reuters, 2018).

3.1.3 Energy monitoring

Unauthorized data centers or the use of data centers for unauthorized training runs could be detected by monitoring energy consumption, either passively (through grid data obtained via espionage), or actively (using devices to measure grid activity).

If the total amount of energy reaching a data center can be measured with reasonable accuracy, it should be possible to convert the energy estimate into a reasonable approximation of the number of FLOPs completed by that facility (Desislavov et al., 2021). Using the FLOP/s, we could ascertain whether the facility is at an unauthorized size. However, these coarse-grained measurements may only be capable of detecting large-scale violations, and further research is needed to understand how such measurements can be more accurately translated into relevant units like FLOPs. More precise methods of energy monitoring may be necessary for detecting smaller-scale violations or unauthorized training runs.

Precedent. Economists use energy monitoring to verify economic data. For example, economists have used discrepancies between reported GDP growth and energy consumption to suggest exaggerated growth figures in China (Owyang and Shell, 2017). This principle could be applied to detect unauthorized data centers, given the direct relationship between energy consumption and FLOPS.

3.1.4 Customs data analysis

Governments can use customs data to track the movement of key components for large-scale AI computing facilities. Import and export records could be analyzed to identify unusual or unexplained patterns in the movement of critical hardware, equipment or raw materials. A sudden surge in imports of high-performance GPUs or other critical components to a specific region of concern, far exceeding the known requirements of declared facilities in that region, would indicate non-compliance.

Precedent. The U.S. government's End-Use Monitoring (EUM) programs, particularly the Blue Lantern program for direct commercial sales, provide a robust precedent for tracking and verifying the use of sensitive technologies (U.S. Department of State, 2021). Under the Blue Lantern program, the Department of State conducts pre-license, post-license/pre-shipment, and post-shipment checks to verify the legitimacy of proposed transactions and ensure compliance with use, transfer, and security requirements. This program has been successful in promoting understanding of U.S. defense trade controls, building mutual confidence among stakeholders, mitigating risks of diversion and unauthorized use, and uncovering violations of the Arms Export Control Act. A similar approach could be adapted for monitoring the movement and use of critical AI hardware components in countries at different stages of the chip supply chain.

3.1.5 Financial intelligence

Governments could track suspicious financial transactions relating to the purchase of important components of AI development. Financial institutions could be required to flag large or unusual purchases of specialized AI hardware, monitor transactions to known AI chip manufacturers, and cross-reference financial data with customs information.

Precedent. In the US, the Financial Crimes Enforcement Network (FinCEN) uses the Suspicious Activity Report (SAR) system and FinCEN Exchange, a public-private partnership, to combat money laundering, terrorism financing, and organised crime (Financial Crimes Enforcement Network, 2024).

In the early 2010s, an SAR filed by a bank led to the discovery of a complex international bribery scheme. The case resulted in multiple arrests and the seizure of over $100 million in criminal proceeds (Financial Crimes Enforcement Network, 2011). This demonstrates how financial intelligence can uncover sophisticated international financial crimes, potentially adaptable to detecting undeclared AI development activities.

3.2 Access-dependent verification methods

Access-dependent methods can allow for in-depth inspections of key facilities such as AI development facilities, hardware manufacturing facilities, and data centers. If international inspectors have sufficient access to these facilities, this provides a great deal of robustness to a verification regime. However, such methods may be perceived as invasive, and they may rely on the permission of nations that are suspected of unauthorized activity. Access-dependent methods can also be somewhat flexible depending on the amount of political will and the level of access that nations are willing to provide. To preserve privacy or trade secrets, inspectors may receive limited access — enough access to verify that an unauthorized training run is not being conducted but not enough access to see exactly what kind of tasks are being performed.

3.2.1 On-site inspections of data centers

On-site inspections involve physical visits to declared data centers to verify compliance with agreements on computing power. These inspections would focus on several aspects. (1) Chip identifiers — AI-capable chips could be required to have unique identifiers (Aarne et al., 2024). Inspectors could catalog these identifiers to ensure they match declared inventories. (2) Chip activity logs — Require chips to have activity logs that inspectors can analyze to verify that chips are being used in accordance with their declared purposes and within agreed-upon limits, and only licensed code is being executed on the chips (Shavit, 2023). (3) FLOP/s limit compliance — ensuring the data center's total computing power is below agreed thresholds. (4) Certified chip usage — verifying that only approved chip models are in use. (5) Training run evidence — examining records and transcripts of large-scale AI training activities. (6) Hardware integrity — inspecting for any evidence of chip tampering (Aarne et al., 2024).

In addition to requiring periodic inspections, an agreement could also require *continuous monitoring* of certain facilities. In a continuous monitoring setup, inspectors are present at facilities at all times to catch any violations of agreements (such as tampering with hardware). Continuous monitoring could also be used to monitor training logs to verify that a given data center is not being used to conduct an unauthorized training run. A final idea could be to conduct challenge inspections, similar to those conducted by the Organization for the Prohibition of Chemical Weapons (OPCW), where inspections can be called for on short notice based on suspicions of non-compliance (OPCW, 1997).

Precedent. The New START treaty signed by the USA and Russia provides for 18 annual on-site inspections for the American and Russian inspections. Inspections allow for verification activities such as confirming the number of reentry vehicles on deployed missiles, counting nuclear weapons on bombers, and verifying the elimination of weapon systems. The treaty's approach of allowing a limited number of highly structured inspections, focused on counting and verifying specific hardware, is a suggestive precedent for inspections of data centers. Notably, an earlier treaty (START I) also provided for continuous monitoring of specific facilities (Arms Control Association, 2022).

The most significant precedent for the detailed inspection of hardware is the IAEA's mandated use of bespoke tamper-evident containment seals for nuclear materials (International Atomic Energy Agency, 2011). These seals – each of which bears a unique identifier – are designed to provide clear evidence of any tampering or unauthorized access. IAEA inspectors examine these seals during on-site visits, allowing them to detect any undeclared movement or use of nuclear materials.

3.2.2 On-site inspections of semiconductor manufacturing facilities

Inspections of semiconductor manufacturing facilities could be used to determine the quantity and nature of chips produced. The manufacturing of advanced chips is a highly specialized activity, and only a few entities have this capacity (Sastry et al., 2024). For example, it is well known that ASML produces EUV lithography systems which are needed to manufacture the latest generation of advanced chips (Khan et al., 2021). If inspectors identified the existence of such machines, it would be relatively easy to know what kind of chips are possible to construct at the manufacturing site. Inspectors may also be able to use basic metrics like the square-meterage of a facility or the number of lithography machines to bound the number of chips that are possible to produce in such facilities.

These inspections could also verify that facilities are producing chips in accordance with any hardware-related agreements. For example, if nations agreed to only build chips with certain on-chip hardware governance mechanisms, inspections of semiconductor manufacturing facilities could identify non-compliance. Inspectors could look at a sample of chips, potentially midway through production, to ensure they have the correct mechanisms.

Precedent. The use of on-site inspections for monitoring compliance with international agreements has been well-established in other domains, particularly in controlling extreme risks.

Organization for the Prohibition of Chemical Weapons (OPCW). The OPCW conducts inspections at facilities that produce toxic chemicals and their precursors. These inspections involve an initial tour, followed by a detailed inspection plan, physical inspections, and a review of the facility's records to verify compliance. The intensity and duration of inspections vary depending on the perceived risk, with chemicals categorized into three schedules based on their threat level (OPCW, 2024).

Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO). The CTBTO, although not fully operational due to the Comprehensive Nuclear-Test-Ban Treaty's pending entry into force, has established protocols for on-site inspections (OSI). These inspections are intended to verify compliance with the treaty, particularly in detecting and investigating potential nuclear explosions. If the treaty enters into force, an OSI could be initiated upon the request of a State Party. The inspection area could cover up to 1000 km^2 (The Comprehensive Nuclear-Test-Ban Treaty Organisation, 2024).

3.2.3 On-site inspections of AI developers

An international inspection team could visit an AI development facility to ensure that developers are running authorized code, ensure that developers are properly implementing model evaluations and safeguards, assess safety culture, and assess security concerns. Inspections could involve various components, such as reviewing code (Casper et al., 2024), assessing compliance with commitments from safety cases, and conducting semi-structured interviews with key personnel to solicit security-relevant concerns (see Wasil et al. (2024a)). Inspections could uncover the usage of unauthorized or unlicensed AI algorithms. A number of privacy-preserving technologies in development could facilitate such inspections to make inspections less intrusive.

Precedent. The closest precedent is the IAEA's use of on-site inspections, as discussed above. The IAEA's approach demonstrates the feasibility of conducting thorough on-site inspections in sensitive, high-tech environments, which could be adapted for AI development facilities. One key difference is that AI inspections would focus more on software and computational resources rather than physical materials, requiring inspectors with specialized expertise in AI technologies and development practices.

3.3 Hardware-dependent verification methods

Hardware-dependent verification methods may offer robust and privacy-preserving tools for detecting non-compliance. However, these methods require nations with advanced hardware manufacturing capabilities to agree to rules around hardware manufacturing. Another challenge is that advanced chips are already in circulation (without hardware-enabled mechanisms built-in). A verification regime relying on hardware-dependent measures may need to address this "legacy hardware", potentially through retrofitting techniques or gradual phase-out strategies.

If successfully implemented, these methods could dramatically enhance the effectiveness of other verification approaches, particularly on-site inspections. However, they also raise important concerns about privacy, national sovereignty, and potential misuse that must be carefully addressed. Overall, hardware-dependent methods represent a promising but long-term goal, requiring sustained international cooperation and technological innovation to realize their full potential in AI governance.

3.3.1 Chip location tracking

Chip location tracking involves implementing a system to monitor the movement and use of AI-capable chips (Brass, 2024). This method requires international agreement on chip manufacturing standards and the implementation of tracking mechanisms directly into the hardware. Each chip above a certain computational threshold would be assigned a unique identifier and equipped with secure tracking capabilities.

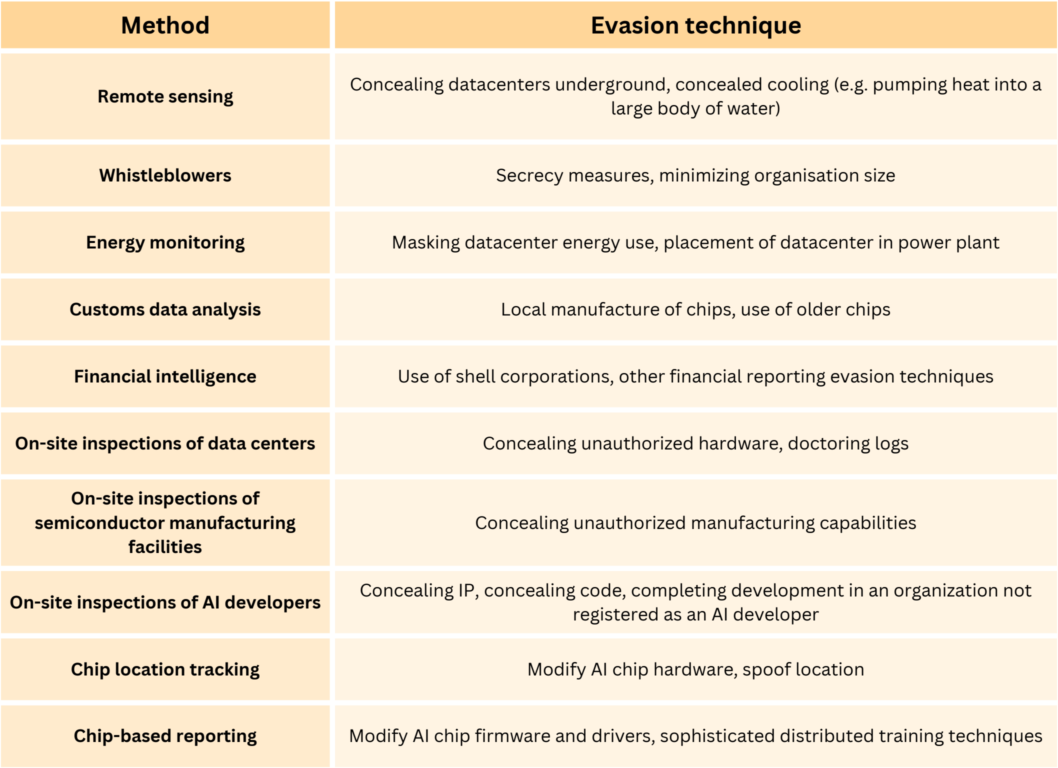
Precedent. The concept of tracking and monitoring critical technology has several precedents across different industries, particularly where security, compliance, and international regulation are concerned.

Nuclear Material Tracking. The IAEA monitors and tracks nuclear materials globally using systems like the Integrated Nuclear Fuel Cycle Information System (International Atomic Energy Agency, 2024b). The IAEA also maintains a databse of incidents involving trafficking or other unauthorized uses of nuclear or radioactive materials (International Atomic Energy Agency, 2024a).

Pharmaceutical Supply Chain Tracking. The Drug Supply Chain Security Act (DSCSA) in the United States currently outlines steps to achieve, "an interoperable and electronic way to identify and trace certain prescription drugs at the package level as they move through the supply chain" (U.S. Food and Drug Administration, 2024).

3.3.2 Chip-based reporting

Chip-based reporting involves implementing mechanisms within AI-capable chips and closely associated hardware (e.g., networking cards) to automatically detect and signal when they are being used in ways that violate agreed-upon constraints. These constraints might include thresholds on the number of chips connected together, or specific operations the chip is not authorized to perform. By embedding these reporting mechanisms at the lowest levels of the software stack — within the firmware and drivers of the AI-capable chips or associated networking devices — it may become more challenging for developers to bypass these safeguards. As one moves up the software stack, toward components that operate at higher levels of abstraction, it becomes easier for developers to replace authorized programs with their own software, potentially circumventing the constraints. Therefore, focusing on the lower levels of the stack, such as firmware, which is the (often read-only) software residing on the device (NASA, 2004), and the drivers, which allow the operating system to communicate with the device (Microsoft, 2023), is crucial for effective enforcement of constraints. These components are typically developed by the chip maker, further limiting the number of developers who could foreseeably edit reporting mechanisms.



Method	Evasion technique
Remote sensing	Concealing datacenters underground, concealed cooling (e.g. pumping heat into a large body of water)
Whistleblowers	Secrecy measures, minimizing organisation size
Energy monitoring	Masking datacenter energy use, placement of datacenter in power plant
Customs data analysis	Local manufacture of chips, use of older chips
Financial intelligence	Use of shell corporations, other financial reporting evasion techniques
On-site inspections of data centers	Concealing unauthorized hardware, doctoring logs
On-site inspections of semiconductor manufacturing facilities	Concealing unauthorized manufacturing capabilities
On-site inspections of AI developers	Concealing IP, concealing code, completing development in an organization not registered as an AI developer
Chip location tracking	Modify AI chip hardware, spoof location
Chip-based reporting	Modify AI chip firmware and drivers, sophisticated distributed training techniques

Figure 2: Example evasion techniques for all verification methods considered.

Precedent. The closest precedent for this type of firmware-based reporting is the Light Hash Rate (LHR) GPUs developed by NVIDIA. These GPUs can detect, via mechanisms implemented in their firmware and drivers, whether they are being used for Ethereum mining (Nvidia, 2021). Similar strategies could foreseeably be developed to report unauthorized AI training.

4 Discussion

This paper examined verification methods that could help nations detect non-compliance with international agreements prohibiting unauthorized AI development and unauthorized data centers. Verification methods vary in their feasibility, intrusiveness and effectiveness. National technical means offer a valuable starting point, capable of detecting large-scale AI infrastructure and unusual patterns in energy consumption, hardware imports, and financial transactions. However, national technical means are limited in their ability to identify software-level violations or concerted attempts at concealment. Access-dependent methods, such as on-site inspections, provide more robust reassurance but require nations to agree to international inspections. Hardware-dependent approaches offer additional robustness (potentially even guarantees) but face some implementation challenges, including the need to address existing legacy hardware.

Table 1 summarizes gaps in individual verification methods, as well ways each method can be complemented by other methods. Figure 3 lists the verification methods based on the amount of additional research required to implement each method.

4.1 Future research directions

Our work provides a starting point for discussions about verification methods, but there are many open questions that can be addressed by future work. We outline some of these directions include below.

Red-teaming exercises for international verification. In a "red-team" step, the authors could brainstorm how an adversary might try to hide an unauthorized training run or unauthorized data center. Then, in a "blue team" step, the authors could identify how one or more verification methods could catch the adversary. Then, in a subsequent "red team" step, the authors could brainstorm if there are feasible ways for the adversary to avoid or undermine the verification method(s). This process could be used to determine likely ways that adversaries may try to evade verification methods and highlight ways of strengthening international verification regimes (e.g., by identifying suites of methods with complementary failure modes).

Design of international AI governance institutions. Compliance with international agreements is often verified by international institutions. Some early work has proposed international organizations that could set and verify compliance with safety standards (Ho et al., 2023; Cass-Beggs et al., 2024), certify national licensing agencies (Trager et al., 2023), verify compliance with a variety of potential agreements (see Maas and Villalobos (2023)), and participate in joint AI safety research (Cass-Beggs et al., 2024). One avenue for future research is to provide more details about how an international verification agency could be structured, how decision-making power is distributed between nations, how the agency handles disputes over non-compliance, and what powers ought to be granted to the agency. Such work could draw from lessons learned from the design and implementation of other international institutions (such as the IAEA and the OPCW) and bilateral or multilateral agreements (such as the Strategic Arms Reduction Treaties and the Wassenaar Agreement).

Enforcement of international agreements. Our paper focused on *verification*– detecting whether or not nations are complying with an agreement. A separate important question is *enforcement*– how nations should react in the event that non-compliance is identified. Such work could examine what kinds of responses would be proportionate to the violation. For example, evidence of small-scale chip smuggling would warrant a less strong response than evidence of an illegal or unauthorized training run.

Research on hardware-enabled mechanisms to enhance verification and/or enforcement. Hardware-enabled mechanisms can unlock new verification methods and make existing verification methods more robust. Some hardware-enabled mechanisms are ready to be implemented swiftly, while others may take several years of research to further develop. Additionally, there are open questions relating to how to make hardware-enabled mechanisms more tamper-proof and privacy-preserving (see Kulp et al. (2024)).

Detecting unauthorized AI deployment or inference. Our paper focuses on detecting unauthorized AI *development*. Nations may also wish to have agreements in which they agree not to *deploy* advanced AI systems in certain ways (for example, nations might prohibit AI from being deployed in the context of nuclear systems, military R&D research, or AI R&D research that could trigger uncontrolled AI development.) Future work could examine verification methods that could detect the unauthorized deployment of AI systems, potentially through hardware-enabled licenses that detect the presence of unauthorized code used for inference.

Detecting compliance with agreements around model evaluations. International agreements may require the usage of model evaluations to detect potential safety or security issues (see Shevlane et al. (2023)). Reliable risk evaluations and risk mitigation strategies could become a minimum safety bar imposed by international agreements. Future work could examine verification methods that allow international authorities to ensure that parties are implementing a set of internationally-required model evaluations, as well as any specific model evaluations that a developer proposed as part of a safety case or licensing application (see Clymer et al. (2024); Wasil et al. (2024b)).

5 Conclusion

Our work provides an initial step toward a better understanding of how compliance with international AI agreements could be verified. Efforts to improve our understanding of verification methods will be especially important if global security risks from advanced AI become concerning enough to motivate coordinated national and international action. We believe some AI governance work should aim to prepare in advance for such scenarios. Such "future-oriented" AI governance work could address questions that would inform policymaking efforts in scenarios where concerns about global security risks became significantly stronger. Our hope is that our work on verification methods illustrates an example of promising work in this category.

References

Aarne, O., Fist, T., and Withers, C. (2024). Secure, governable chips, january 2024. *URL https://www.cnas.org/publications/reports/secure-governable-chips. Accessed*, pages 1–28.

Arms Control Association (2022). START I at a glance.

Armstrong, S., Bostrom, N., and Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & society*, 31:201–206.

Baker, M. (2023). Nuclear arms control verification and lessons for AI treaties.

Barnard, C. I. and Acheson, D. (1946). A report on the international control of Atomic Energy. Technical report, US Department of State.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., et al. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition.

Brass, A. (2024). Location verification for AI chips. Technical report, Center for AI Governance. Accessed: 2024-08-09.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., et al. (2024). Black-box access is insufficient for rigorous AI audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272.

Cass-Beggs, D., Clare, S., Dimowo, D., and Kara, Z. (2024). Framework convention on global AI challenges.

Clymer, J., Gabrieli, N., Krueger, D., and Larsen, T. (2024). Safety cases: Justifying the safety of advanced AI systems. *arXiv preprint arXiv:2403.10462*.

Desislavov, R., Martínez-Plumed, F., and Hernández-Orallo, J. (2021). Compute and energy consumption trends in deep learning inference. *arXiv preprint arXiv:2109.05472*.

Fearon, J. D. (1995). Rationalist explanations for war. *International organization*, 49(3):379–414.

Financial Crimes Enforcement Network (2011). SAR leads to recovery of funds derived from foreign corruption. Accessed: 2024-08-09.

Financial Crimes Enforcement Network (2024). FinCEN advisory to financial institutions to counter the financing of Iran-backed terrorist organizations. Accessed: 2024-08-09.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

Heim, L. (2024). Training compute thresholds: Features and functions in AI governance. *arXiv preprint arXiv:2405.10799*.

Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., et al. (2023). International institutions for advanced AI. *arXiv preprint arXiv:2307.04699*.

Hogarth, I. (2023). We must slow down the race to God-like AI. *Financial Times*. `https://archive.is/jFfBQ`. Accessed 2024-08-09.

International Atomic Energy Agency (2011). *Safeguards Techniques and Equipment: 2011 Edition*. Number 1 (Rev. 2) in International Nuclear Verification Series. International Atomic Energy Agency. `https://www.iaea.org/publications/10416/safeguards-techniques-and-equipment-2011-edition`.

International Atomic Energy Agency (2024a). IAEA Incident and Trafficking Database (ITDB). Accessed: 2024-08-09.

International Atomic Energy Agency (2024b). IAEA Nuclear Fuel Cycle Information System (NFCIS). Accessed: 2024-08-09.

Khan, S. M., Mann, A., and Peterson, D. (2021). The semiconductor supply chain: Assessing national competitiveness. *Center for Security and Emerging Technology*, 8(8):1–98.

Kulp, G., Gonzales, D., Smith, E., Heim, L., Puri, P., Vermeer, M. J. D., and Winkelman, Z. (2024). Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090.

Loyens, K. and Vandekerckhove, W. (2018). Whistleblowing from an international perspective: A comparative analysis of institutional arrangements. *Administrative Sciences*, 8(3):30.

Maas, M. M. and Villalobos, J. J. (2023). International AI institutions: A literature review of models, examples, and proposals. *AI Foundations Report*, 1.

Microsoft (2023). What is a driver? `https://learn.microsoft.com/en-us/windows-hardware/drivers/gettingstarted/what-is-a-driver-` Accessed 2024-08-09.

NASA (2004). Software safety standard. `https://klabs.org/ce_watch/sw_documents/871913b.pdf`. Accessed 2024-08-09.

Ngo, R., Chan, L., and Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Nvidia (2021). A further step to getting geforce cards into the hands of gamers. `https://blogs.nvidia.com/blog/lhr/`. Accessed 2024-08-09.

OPCW (1997). Convention on the prohibition of the development, production, stockpiling and use of chemical weapons and on their destruction.

OPCW (2024). Industry Inspections: What to Expect. `https://www.opcw.org/industry-inspections-what-to-expect`. Accessed 2024-08-09.

Open Nuclear Network (2024). Strengthening Nuclear Test Ban Monitoring and Verification: the Role of Commercial Satellite Imagery. `https://oneearthfuture.org/sites/default/files/2024-06/TheRoleOfCommercialSatelliteImagery-.pdf`. Accessed 2024-08-09.

Owyang, M. T. and Shell, H. (2017). China's economic data: an accurate reflection, or just smoke and mirrors? *The Regional Economist*, 25(2).

Pan, A., Bhatia, K., and Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*.

Reuters (2018). U.S. SEC awards Merrill Lynch whistleblowers a record $83 million. `https://www.reuters.com/article/business/u-s-sec-awards-merrill-lynch-whistleblowers-a-record-83-million-idUSKBN1GV2UX/`. Accessed 2024-08-09.

Rutkowski, J. and Niemeyer, I. (2020). Remote sensing data processing and analysis techniques for nuclear non-proliferation. *Nuclear Non-proliferation and Arms Control Verification: Innovative Systems Concepts*, pages 339–350.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., et al. (2024). Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*.

Securities and Exchange Commission (SEC) (2016). Merrill Lynch to Pay $415 Million for Misusing Customer Cash and Putting Customer Securities at Risk. `https://www.sec.gov/newsroom/press-releases/2016-128`.

Shavit, Y. (2023). What does it take to catch a Chinchilla? Verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Statista (2022). Commercially available satellite imagery worldwide in 2022, by spatial resolution. https://www.statista.com/statistics/1293723/commercial-satellite-imagery-resolution-worldwide/ Accessed 2024-08-09.

The Comprehensive Nuclear-Test-Ban Treaty Organisation (2024). On-site inspection. https://www.ctbto.org/our-work/on-site-inspection. Accessed 2024-08-09.

Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., hÉigeartaigh, S. Ó., et al. (2023). International governance of civilian AI: A jurisdictional certification approach. *arXiv preprint arXiv:2308.15514*.

U.S. Department of State (2001). Interim Agreement Between The United States of America and The Union of Soviet Socialist Republics on Certain Measures With Respect to the Limitation of Strategic Offensive Arms (SALT I).

U.S. Department of State (2021). End-use monitoring of U.S.-origin defense articles. https://www.state.gov/end-use-monitoring-of-u-s-origin-defense-articles/. Accessed 2024-08-09.

U.S. Department of State (2023). New START Treaty. https://www.state.gov/new-start/. Accessed 2024-08-09.

U.S. Food and Drug Administration (2024). Drug Supply Chain Security Act (DSCSA). Accessed: 2024-08-09.

U.S. Securities and Exchange Commission (2017). SEC 2017 Annual Report: Whistleblower Program. Technical report, U.S. Securities and Exchange Commission.

Wasil, A., Berglund, L., Reed, T., Plueckebuam, M., and Smith, E. (2024a). Understanding frontier AI capabilities and risks through semi-structured interviews. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4881729.

Wasil, A., Clymer, J., Krueger, D., Dardaman, E., Campos, S., and Murphy, E. (2024b). Affirmative safety: An approach to risk management for high-risk AI. https://arxiv.org/pdf/2406.15371.

Wasil, A. and Durgin, T. (2024). US-China perspectives on extreme AI risks and global governance. https://arxiv.org/abs/2407.16903.

Yuan, X., Liang, Y., Hu, X., Xu, Y., Chen, Y., and Kosonen, R. (2023). Waste heat recoveries in data centers: A review. *Renewable and Sustainable Energy Reviews*, 188:113777.

A Limitations of methods and possible solutions using complementary methods

Table 1: Limitations of methods and possible solutions using complementary methods.

Verification method	Primary limitations	Complementary methods
Satellite imagery	Data centers could be concealed underground or camouflaged	*National intelligence services* can provide human intelligence and signals intelligence to identify hidden facilities that satellite imagery might miss. They can gather information on construction activities, personnel movements, and communications that could indicate the presence of a concealed data center. *Energy monitoring* complements satellite imagery by detecting unusual power consumption patterns that might indicate a hidden data center. Even if a facility is visually concealed, its energy requirements are difficult to hide, especially for large-scale AI operations. *Chip location tracking* can determine the approximate location of data centers and discourage concealment.
Whistleblowers	Reliability issues: Whistleblowers may provide incomplete, biased, or false information. Limited access: Not all potential whistleblowers have access to critical information. Fear of retaliation: Potential whistleblowers may be deterred by fears of personal or professional consequences.	*National intelligence services* can corroborate or refute whistleblower claims through other intelligence gathering methods. *On-site inspections* can be triggered by whistleblower reports, allowing for direct verification of claims. Inspectors can look for specific evidence pointed out by whistleblowers, increasing the effectiveness of the inspection. Financial intelligence can be used to verify claims about resource allocation or unusual transactions mentioned by whistleblowers.

Verification method	Primary limitations	Complementary methods
National intelligence services	Using national intelligence can be unnecessarily invasive and infringe on national sovereignty. The source of intelligence is often classified, which makes it a poor foundation for transparent discussion in international forums.	*Satellite imagery* can provide visual confirmation of intelligence reports about suspected facilities, offering a less intrusive method of verification. *Financial intelligence* can corroborate intelligence about resource allocation and unusual transactions, providing a paper trail for activities identified through other intelligence means.
Energy monitoring	Though plausible in theory, this method is unproven in practice. Energy consumption may be disguised as other high-energy activities. Obtaining detailed energy consumption data is also likely to be challenging.	Customs data analysis can corroborate energy monitoring data by tracking the import of high-performance computing equipment to areas with suspicious energy consumption patterns.
Customs data	Countries with advanced domestic manufacturing capabilities may be able to produce key components internally, reducing the effectiveness of customs data analysis. Distinguishing between components intended for authorised and unauthorised is likely to be challenging.	*On-site inspections of semiconductor manufacturing* facilities can verify whether domestic production capabilities match declared capacities, helping to identify discrepancies that might indicate undeclared production bypassing customs. *Chip location tracking*, if implemented, can help verify the end destination and use of key components that have passed through customs, ensuring they are being used as declared.
Financial intelligence	Many AI-related purchases may have legitimate alternate uses, making it difficult to distinguish between authorized and unauthorized activities. This method may be disproportionately invasive. Financial intelligence is also limited by banking secrecy laws and a potential lack of international cooperation.	*Customs data analysis* can corroborate financial intelligence by providing physical evidence of hardware purchases and movements that correspond to suspicious financial transactions. *Whistleblowers* can provide insider information about financial practices, helping to interpret complex transactions or reveal hidden financial structures used to fund unauthorized AI development.
Data center inspections	Inspections can only be carried out with the agreement of the host nation, potentially allowing time for concealment of violations. Thorough inspections are also both invasive and very resource-intensive, requiring significant time, expertise and resources.	*Chip location tracking*, if implemented, can be verified during inspections to ensure that the physical location of AI-capable chips matches their reported locations. *Whistleblower* information can guide inspectors to look for specific evidence of non-compliance that might otherwise be overlooked.

Verification method	Primary limitations	Complementary methods
Fab inspection	Like all inspections, these inspections are also resource-intensive, invasive and pose threats to intellectual property. The technological complexity of chip manufacturing may also make it challenging for inspectors to detect potential violations without highly specialized expertise.	*Chip location tracking*, if implemented, can be initiated during the inspection process, ensuring that newly manufactured AI-capable chips are properly registered and tracked from the point of production.
AI developer inspection	Unlike hardware, software can be quickly modified or hidden, making violations difficult to detect. Such inspections also require highly specialized knowledge, and may pose a disproportionate risk to proprietary algorithms and research.	*Whistleblowers* can provide insider information about development practices, guiding inspectors to specific areas or systems of concern. *Financial intelligence* can be cross-referenced to ensure declared AI projects match financial records.
Chip location tracking	Requires agreement on chip manufacturing standards. Sophisticated actors may find ways to disable tracking mechanisms. The effectiveness of this intervention would be limited to the production of new chips.	*On-site inspections*of manufacturing sites can ensure that chips are being made with the required location tracking mechanisms. *Satellite imagery* can provide additional, more precise location tracking.
Fixed set reporting	Requires agreement on chip manufacturing standards. False positives/negatives: Balancing sensitivity to catch violations without triggering false alarms is difficult. The effectiveness of this intervention would be limited to the production of new chips.	*On-site inspections* of data centers can be triggered by automatic signals of non-compliance, allowing for rapid verification of potential violations.
Firmware-based reporting	Requires agreement on chip manufacturing and implementation standards; difficult to implement; may come with an economic or computational cost.	*On-site inspections* could be triggered by automatic signals of non-compliance, allowing for rapid verification of potential violations.

B Research and development estimates for verification methods

	Reasoning	Research Needed
Whistleblowers	Has been used in the past	🟩
Customs data	Existing US monitoring of semiconductor exports	🟩
Financial intelligence	Has been used in the past	🟩
Remote sensing	Satellite imagery has been used to monitor actor activities	🟩
Fab inspections	Expertise exists but need to know what standards to require	🟧
Data center inspections	Expertise exists but need to know what standards to require	🟧
Energy Monitoring	Already exists but little work on connection to data centers and training	🟧
AI developer inspection	Expertise exists but need to know what standards to require	🟧
Chip location tracking	Further research needed to develop the hardware	🟥
Chip-based reporting	Further research needed to develop the hardware, firmware and drivers.	🟥

Figure 3: Estimated research and development needed for verification methods investigated. Note that green indicates little additional research needed, orange indicates some additional research and red indicates significant additional research.