Verifying Restrictions on Frontier AI Research

Aaron Scher [1]

Abstract

The premature development of artificial superintelligence poses major risks to humanity, so researchers have proposed international agreements halting such development until it can be done safely. AI progress depends primarily on compute, algorithms, and data; a durable halt would address all three so that advances in one input do not counteract restrictions on another. Improvements to AI algorithms are driven largely through research activities, so this research may need to be restricted during a halt. Given low international trust, signatories will want to verify compliance. This paper analyzes how such restrictions on AI research could be verified, while remaining agnostic about what specific research would be prohibited. It first explores key considerations that affect the verifiability of research restrictions, such as the computational infrastructure necessary for experiments. It then catalogs 28 candidate verification mechanisms. These mechanisms include whistleblowers, search warrants, reviews of AI training code, standard intelligence gathering tools, and more. Some of these mechanisms are not yet implementation-ready, and some might be undesirable upon further inspection. By examining the space of potential options, this work provides a foundation for future research to develop the most promising mechanisms into deployable tools.

1. Introduction

Many experts in the field of artificial intelligence (AI) worry that increasingly intelligent AI systems will pose catastrophic risks to humanity, including human extinction (Center for AI Safety, 2023; Yudkowsky & Soares, 2025). To address this threat, some have proposed delaying the development of artificial superintelligence (ASI) until development can be done safely (Future of Life Institute, 2025; Barnett & Scher, 2025; Scher et al., 2025; Ramiah et al., 2025; Martin et al., 2024).

To this end, Scher et al. (2025) provide a draft international agreement aimed at preventing the premature creation of ASI. For simplicity, we use the agreement from Scher et al. (2025) as a scaffold for this work, asking how research restrictions could be verified under that agreement; however, these mechanisms would also be useful in other policy proposals that involve research restrictions. The agreement has two main pillars focused on stopping the development of better AIs: *chip controls* and *research restrictions*. Both of these are needed; neither would be sufficient on their own. The primary drivers of AI progress are compute, data, and algorithms (Buchanan, 2020). If only compute is controlled, progress would still continue via advances in data and algorithms (Ho et al., 2024; Hooker, 2024; Heim & Koessler, 2024). The proposed agreement includes research restrictions to fill this gap, focused primarily on research that is likely to advance the state of frontier AI capabilities (e.g., frontier AI research, general-purpose AI capabilities research, potentially AI chip design research). Certain research would also endanger the verification system established in the agreement, so it may be desirable to prohibit such research as well (e.g., decentralized training methods). The question of which specific research should be restricted during a halt is out of scope for this paper.

We focus instead on the verification question: **How could governments verify compliance with restrictions on AI research, especially within signatory countries?** We answer this question in two ways:

- **First, we analyze key factors influencing the verifiability of research restrictions.** These include the need for compute in testing research ideas, the present rarity of human experts who can push the frontier, and the likely automation of AI research activities.
- **Second, we map the space of potential verification mechanisms, cataloging 28 high-level mechanisms** that could be used to verify compliance with research restrictions. Examples of these mechanisms include whistleblowers, automated code reviews, and sting operations.

[1]Technical Governance Team, Machine Intelligence Research Institute, Berkeley, USA. Correspondence to: Aaron Scher <aaron.scher@intelligence.org>.

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

To simplify, we use specific terminology throughout this paper. *Restricted research* refers to research that is prohibited by an international agreement, for instance because it contributes to AI capabilities or threatens the verification system; this paper remains agnostic about where the boundaries of that category should be drawn. Scher et al. (2025) use restricted research as an umbrella term covering two forms: *controlled* research, which is subject to monitoring, and *banned* research, which is disallowed. In this paper, we use restricted research to refer to the latter. *High-risk organizations and researchers* are organizations and individuals that are relatively likely to carry out impactful restricted research, for instance because their legitimate work is similar to restricted research (i.e., it falls in the controlled category) or because they have relevant expertise and resources. For example, current AI developers such as OpenAI and Anthropic would qualify as high-risk organizations due to their expertise and resources. After an agreement, they would not be allowed to continue conducting frontier capabilities research, but they could, for instance, serve existing models.

The *verification authorities* are a group of people responsible for verification, for instance a technical body staffed by representatives from certain signatory countries, as in the agreement from Scher et al. (2025) (though we aim to be agnostic to the design of such a body as this is primarily a political rather than technical question). The choice of verification authorities affects the political viability and security risk of mechanisms, but these considerations will be balanced against countries' need to obtain justified confidence in compliance, not trusting each other to self-enforce an agreement. *Inspectors* are employees of the verification authority who carry out specific monitoring or inspection tasks. Finally, we use the term *verification* to refer to a broad set of methods and actions that could make one party justifiably more confident that another is adhering to agreed-upon rules. This is somewhat broader than the standard use as it includes preventative mechanisms and useful components as well as core information collection—we note when our use is more broad than typical.

To ground this work, our investigation is motivated by three main threat actors who may attempt to violate research restrictions. First, covert nation-state actors who are highly competent at security, may have secret data centers, and whose existence may not be known to verification authorities (e.g., a secret Manhattan Project for ASI). Second, established organizations that are known to verification authorities, are highly competent at AI development, and are generally law-abiding (e.g., OpenAI, DeepSeek, Prime Intellect). Third, rogue academics and less-resourced researchers who may be ideologically motivated and are likely to make progress either slowly (and with contributions from many individuals) or through new paradigm advances, if at all. In some cases, the potential violator is known, so verification would involve countries agreeing to monitoring and inspections to verify compliance. In other cases the main verification difficulty is to detect the presence of such a project.

2. Key considerations for the verifiability of research restrictions

Compute is needed to test many research ideas within the machine learning paradigm. Barnett (2025) investigates how much compute was used to develop the algorithmic innovations used in Llama 3 and DeepSeek-V3 and finds that around half of the innovations were developed using less than 8 H100 equivalents of computing capacity (or around 1×10^{21} FLOP when measuring total computations). This quantity is larger than what average consumers own, but small enough to be effectively undetectable: we are not aware of any detection mechanism that could reliably find all such clusters in the world, though such an analysis is outside the scope of this work.

Validating research ideas at scale requires substantial compute. Industry practitioners report that many ideas which perform well in small experiments nonetheless scale poorly, and to identify promising directions requires larger-scale experiments (Dean & Shazeer, 2025, 00:45:43). Recent work by Gundlach et al. (2025) finds that several of the most consequential algorithmic innovations are compute-dependent: their benefit differs when applied at small vs. large training scales. Algorithm researchers seeking to validate such ideas and push the capability frontier would therefore require more compute, rendering their activity more detectable.

The extent to which frontier AI research is bottlenecked by compute is not publicly established. Prior work examining the implications of automated AI research has diverged on the severity of this bottleneck (Anthropic, 2026, p. 44); see also Whitfill & Wu (2025) and Davidson (2025).

The population of researchers relevant to frontier AI development is small—plausibly on the order of thousands or tens of thousands. Scher et al. (2025) provide rough estimates for the number of these researchers, ranging from thousands (i.e., technical staff at frontier AI companies) to tens of millions (i.e., all software engineers and AI chip company employees), with tens of thousands being a reasonable overall estimate. The skills required to advance the frontier are rare. In a world where humans, rather than AI systems, remain the primary drivers of AI research progress, the set of individuals whose activities warrant monitoring is correspondingly bounded.

AI systems may come to perform most AI research in the future. At the time of writing, frontier AI companies are working to automate the labor of their human engineers

and researchers (Ecoffet et al., 2026; Reedi et al., 2026), though humans remain important in the process. This balance is likely to shift substantially within the next several years (Favaro & Clark, 2026; Altman & Pachocki, 2026). If and when an international agreement enters into force, AI systems will likely conduct a substantial share of AI research and development—perhaps all of it. This shift cuts in both directions for verification. On one hand, AI systems are trivially copyable and considerably harder to detect than human researchers, especially smaller models capable of running on consumer hardware. On the other hand, imposing intensive monitoring regimes on AI systems raises fewer normative and legal objections (Mark & Scher, 2025).

Criminalization would substantially reduce willingness to engage in affected research. Under the agreement proposed by Scher et al. (2025), specified categories of AI research would become illegal and would likely acquire significant social stigma. The agreement does not specify enforcement mechanisms in detail, requiring only that each party "prohibits and prevents" restricted research; in practice, penalties could plausibly include prison sentences. Individuals and corporations typically comply with legal prohibitions for a range of reasons, including the expected cost of detection and sanction (Nagin, 2013). Investment in the affected areas would likewise contract in response to heightened legal and financial exposure. As an informal estimate, we expect aggregate effort directed toward frontier AI research to fall to approximately 1–10% of its pre-agreement level, though this figure is speculative and sensitive to the credibility of enforcement.

3. Mechanisms that could be used to verify compliance with research restrictions

We catalog 28 of the most promising high-level mechanisms from among a larger set considered. The full set draws on precedent, related fields, and an understanding of the AI R&D process. We group mechanisms by their target—identifying covert projects, monitoring declared high-risk organizations, verifying use of declared chips, or broad non-proliferation. We then sort each group in descending order of priority—an informal assessment of the costs (e.g., privacy concerns, amount of R&D needed, difficulty of implementation) and benefits (e.g., effectiveness against the threat actors). The least promising approaches are omitted entirely.

3.1. Covert projects

During an international agreement on AI, actors may seek to continue prohibited development in secret. The mechanisms below can help detect, identify, or prevent such covert projects.

Table 1. Catalog of 28 candidate verification mechanisms for restrictions on AI research, grouped by their target.

Covert projects
Whistleblowers
Intelligence gathering on known high-risk targets
Intelligence gathering to identify unknown high-risk targets
Interviews with researchers and relevant officials
International search warrants and inspections
Leadership emphasis on compliance†
Polygraphs and lie detection
Sting operations for covert projects
Declared high-risk organizations
AI-assisted code review for compliance
Auditors in high-risk organizations
Employee-monitoring software overseen by inspectors
Monitoring the use of chip design tools
Monitoring chip fabrication
Multi-party compliance attestations
Pre-registration and review of controlled research plans
Pre-publication review of controlled research
Researcher education and outreach†
Recordkeeping and disclosure on request
Evaluating AI models to detect unexplained capability gains
Controlled work environments to prevent evasion
Chip-use verification for declared chips
Review AI training for novel methods of relevance
Monitoring AI inference content for restricted research
Review AI training data for prohibited data improvements
Broad chip-use monitoring for restricted experiments
Restricting chips to only run approved models
Non-proliferation
Preventing the development of AI-R&D-capable AIs†
Preventing the proliferation of AI-R&D-capable AIs†
Non-proliferation of AI algorithms and data†

† Not a verification mechanism in the traditional use of the term, but still useful for obtaining confidence in compliance.

Whistleblowers Establish secure channels for whistleblowing about the existence of secret AI projects, including anonymous channels. If possible, establish strong incentives around whistleblowing. The SEC's Dodd-Frank whistleblower program, which offers 10–30% of collected sanctions and has paid over $2 billion since 2010, provides a direct model for financially incentivized reporting with anonymity and anti-retaliation protections. Asylum guarantees that extend to a whistleblower's family may also incentivize reporting.

Intelligence gathering on known high-risk targets Use traditional intelligence gathering methods to monitor high-risk organizations and individuals. States might employ such methods as signals and communications intelligence to understand whether restricted research projects are taking place. They might also use financial intelligence to detect AI chip rentals or hiring, and human intelligence, including interviews with researchers and those close to them.

Intelligence gathering to identify unknown high-risk targets Intelligence agencies search for covert projects by first identifying individuals and organizations that are likely to engage in restricted research. The methods used might parallel those used in the past to detect covert WMD projects, including human intelligence and signals intelligence (Committee of Privy Counsellors, 2004). Compared to historical analogues, AI research appears more difficult to detect: data centers and chips are the only input with a notable physical footprint, and AI research requires relatively few chips. Arms control treaties from SALT I to New START explicitly authorize national technical means for verification and prohibit parties from interfering with each other's monitoring capabilities. In some cases, actors who continue AI research despite global restrictions might leave public clues. They might publish their research findings online (e.g., for ideological reasons), organize events, or otherwise signal non-compliance. Therefore, open source intelligence and monitoring public AI research might be effective against non-covert rulebreakers. The IAEA systematically collects and analyzes open source scientific literature as part of its safeguards verification activities, and this practice can cleanly adapt to monitoring public AI research.

Interviews with researchers and relevant officials Verification authorities interview high-risk researchers, aiming to uncover whether the researchers performed restricted research after the restrictions took effect. To guard against misconduct, a researcher's country of origin might oversee the interview. Interviews are a precedented measure; the Chemical Weapons Convention (CWC) Verification Annex explicitly grants inspectors the right to interview facility personnel.

International search warrants and inspections Carry out search warrants for properties, files, and computers, a standard method for gathering evidence about suspected illegal activity. Information collected from these sources can identify whether the targets are engaged in restricted research. Research is difficult, and researchers benefit from recording their ideas and experiments, often creating a paper trail. Warrants and inspection bodies would likely need to be international to avoid lax self-policing or overreach by foreign authorities. Possible design features include approval from an international judge appointed with the consent of the host country. To guard against collusion with or tipoffs from domestic sources, warrants could follow the precedent set by the CWC, which established a right to short-notice challenge inspections at any facility in any state party, with no right of refusal, although this right has not yet been exercised.

Leadership emphasis on compliance Major world leaders publicly declare that participation in restricted research or covert frontier AI development is a serious offense, even if directed by a superior. Leaders encourage whistleblowing, discourage restricted AI research, and emphasize the seriousness of this position at all levels of command. CIA Director Schlesinger's 1973 directive (Schlesinger, 1973) ordering all employees to report activities outside the Agency's charter is a precedent for senior leaders using explicit directives to signal that compliance is a top priority. This action could make covert projects less likely and would be visible to rivals, but it is not a verification measure in the traditional use of the term.

Polygraphs and lie detection Sponsor research into lie detection technology and apply it as an added layer of screening in conversations with high-risk researchers and with public officials likely to know about covert projects. Improvements in lie detection would render interviews, meetings, and public statements more credible. Polygraph examinations are required for all CIA and NSA personnel, but the scientific community has not established that they are reliable screening tools (National Research Council, 2003). Even weak or unreliable signals could be useful to prompt further investigation.

Sting operations for covert projects Run sting operations or false-flag recruitment to catch researchers willing to unlawfully participate in a covert project, chip owners willing to rent or sell to such projects, or other nascent threats. Sting operations have been used by law enforcement in domains ranging from drug enforcement to terrorism and WMD procurement.

3.2. Declared high-risk organizations

To aid verification, an agreement would likely require that organizations and individuals that pose a significant risk of conducting restricted research declare themselves to the verification authorities and accept monitoring protocols such as those below.

AI-assisted code review for compliance Require that all production and research code at high-risk organizations be screened by an AI to flag possible restricted research. Automated screening might extend to codebases, experiment logs, and research documents, and might output only compliance-relevant findings (e.g., "this code implements a novel training algorithm, thus constituting restricted research") without revealing proprietary details to human inspectors. One difficulty is in third-party inspectors verifying that the right code review happens every time it needs to, and another is ensuring the AIs doing code review will do a good job (i.e., they are not jailbreakable, are not attempting to sabotage verification, and are good at the relevant skills). Automated tools such as linters and SAST are common in coding, and AI tools such as Claude Code Review and Cursor Bugbot also see growing adoption. While these tools traditionally focus on bugs and security vulnerabilities, some of them could be adapted to detect violations of an agreement.

Auditors in high-risk organizations Maintain full-time human auditors, provided by the verification authorities, within select high-risk organizations. Auditors aim to ensure that the organization is not engaging in restricted research. The Nuclear Regulatory Commission has stationed full-time resident inspectors at every U.S. nuclear power plant since 1978 to observe daily operations and verify compliance (U.S. Nuclear Regulatory Commission, 2025).

Employee-monitoring software overseen by inspectors Require employee-monitoring software on the work devices of high-risk organizations and researchers. Potential implementations vary in the data they collect and how it is analyzed. Inspectors would need sufficient access to verify compliance, but could analyze data using automated methods to preserve privacy. Similar software is widespread in corporate and classified government settings, including data loss prevention software, endpoint detection and response, and application-specific user activity logging, though having external inspectors oversee such monitoring on researchers' devices is a novel addition to existing practice.

Monitoring the use of chip design tools Collaborate with Electronic Design Automation (EDA) companies (Synopsys, Cadence, Siemens EDA) to ensure their tools are not being used in ways that violate the agreement. Semiconductor foundries (e.g., TSMC, Intel) provide designers with a Process Design Kit (PDK) specific to their fab and necessary for fabrication. These companies could work with verification authorities to limit tooling use to legitimate organizations performing only permitted research. This mechanism is relevant if research restrictions include research into better AI chips.

Monitoring chip fabrication Collaborate with chip fabrication companies to ensure new chips are not the product of restricted research. Cutting-edge chips are fabricated by a small handful of companies (e.g., TSMC, Intel, SMIC, Samsung) who have significant visibility into chip design. Various performance characteristics of chips could be inferred from either chip designs or fabricated chips (e.g., via independent teardown analysis), so the world's chip production could be monitored to ensure that chips are not advancing along the relevant axes—for agreements in which research on improving frontier AI chips is restricted.

Multi-party compliance attestations Require high-risk organizations to periodically submit attestations from multiple independent employees (e.g., 5+ signatories from different divisions) asserting that no restricted research is active or planned. This is cheap to implement but would be ineffective against a highly competent adversary. The Sarbanes-Oxley Act (United States Congress, 2002, Sections 302 and 906) provides a model for executive-level compliance attestations, requiring CEOs and CFOs to personally certify financial reports each quarter and specifying criminal penalties for false statements.

Pre-registration and review of controlled research plans Require high-risk organizations and researchers to pre-register their research plans with verification authorities. Pre-registration provides early screening of research proposals likely to trigger restrictions, and could be automated to minimize added friction. This approach primarily prevents restricted research from happening by accident, and may also reduce plausible deniability for bad actors: violations either result from a classification error by the verification authorities, or from bad actors circumventing the registration process, itself a flagrant violation. One difficulty in implementing this mechanism is that predicting downstream effects of research is inherently difficult. Pre-registration and review of research exists in the U.S. for human subjects research (Institutional Review Boards) and clinical trials (U.S. Food and Drug Administration), though these processes are much slower than what could be achieved with automation.

Pre-publication review of controlled research Require high-risk organizations and researchers to submit their research results to verification authorities before publication. Verification authorities would check that the work does not constitute restricted research, potentially using automated classifiers to speed up the process. The Atomic Energy Act classifies nuclear weapons design information by default (United States Congress, 1954), and the Invention Secrecy Act of 1951 authorizes the government to prohibit publication of patent applications with national security implications (United States Congress, 1952). In contrast, computer science research is sometimes conducted semi-publicly, for instance using public code repositories, and for work in high-risk categories, this practice would need to change.

Researcher education and outreach Interface with high-risk researchers to ensure they understand the research restriction rules, the consequences of non-compliance, and available whistleblower pathways. This would be relatively cheap to implement and would strengthen both whistleblowers and multi-party compliance attestations, though it is not a verification mechanism in its own right. Training and debriefings are commonplace in fields that use sensitive information; for example, under the U.S. National Industrial Security Program, cleared contractor employees must complete insider-threat training.

Recordkeeping and disclosure on request Require high-risk organizations to maintain an auditable record of important activities and submit select records to inspectors upon request. This is easier and likely more palatable to implement than embedded auditors, but offers weaker evidence about compliance. Required documentation might include meeting notes, code changes, compute workloads, and other important records. Verifiable outputs from the organization (such as product releases) may serve as an additional point of comparison for validating records. Alternatively, a regime might require a cryptographically-signed ledger of high-risk organizations' software code, recorded for later audit (Philippe et al., 2019). In a direct precedent for mandatory recordkeeping and disclosure, the CWC requires state parties to declare relevant activities to the OPCW and provide access to facility records during inspections.

Evaluating AI models to detect unexplained capability gains Perform third-party evaluations of AI models. The agreement from Scher et al. (2025) would constrain compute and research in order to prevent AI capabilities from improving. Therefore, capability increases would indicate that progress is being made in at least one of these areas. This mechanism is useful if there is low visibility into AI development for verification, for instance due to low political will. Verification authorities would need access to any model they intend to evaluate, meaning this mechanism cannot function when development is entirely secret, though it could function if authorities are granted early access to models before deployment. A further challenge is that the field of AI presently lacks a strong basis for predicting how inputs to AI development will affect downstream model capabilities (Ganguli et al., 2022), and also lacks evaluation techniques that are robust to increasingly test-aware AI models (Anthropic, 2026; MSL Preparedness & Red Teaming & Alignment Team & MSL AI Security Team, 2026). Thus, measured model performance cannot provide definitive evidence, only indicate a need for further investigation.

Controlled work environments to prevent evasion Augment employee-monitoring software with methods that hinder evasion. For instance, require that high-risk researchers and high-risk organizations work in Sensitive Compartmented Information Facilities or other secured environments without access to personal devices, and require work devices to run monitoring software. Draw from existing companies' efforts to ensure their employees are not secretly working a second job: monitoring network activity, badge/key entry logs, and more. This mechanism is useful if states are concerned that monitoring alone will not suffice to prevent high-risk researchers from covertly performing restricted research.

3.3. Chip-use verification for declared chips

The agreement proposed by Scher et al. (2025) involves locating and monitoring a large fraction of the world's AI chips, aiming for all non-trivial clusters of chips to be located. We refer readers to previous work for mechanisms that could help locate AI chips (Scher et al., 2025; Baker et al., 2025; Wasil et al., 2024; Scher & Thiergart, 2024; Brass & Aarne, 2024; Fist & Grunewald, 2023; Petrie, 2024). The agreement then applies chip-use verification protocols to ensure that chips are only doing permitted activities such as inference on existing models, but not prohibited activities such as large-scale training. A wide array of previous work aims to identify or build these chip-use verification methods (Baker et al., 2025; Scher & Thiergart, 2024; Harack et al., 2025; Shavit, 2023; Heim et al., 2024). In this section, we assume the vast majority of AI chips have been located and are being monitored, and we ask what specific claims about the chips would be relevant to ensuring restricted research is not taking place. The effectiveness of these mechanisms depends on how much coverage the chip monitoring system has over the total stock of relevant chips—a question we leave to other work. Similar to the mechanisms for declared high-risk organizations, these mechanisms involve a known and legitimate prover (i.e., the chip owner) cooperating with verification authorities.

Review AI training for novel methods of relevance Require that training on monitored compute clusters (this could be all training or all training above some compute scale, such as the *monitored threshold* in Scher et al. (2025)'s agreement) be submitted to inspectors, including all code used, and review this training to ensure it does not use novel AI methods of relevance. Under an agreement, *large* compute clusters are very likely to be known and monitored, whereas covert violators might have *small* secret data centers that are unmonitored. These violators would prefer to use monitored clusters to help them with their AI development because these are major sources of compute. For example, they might conduct restricted research using their secret clusters, and then do larger-scale AI training using the larger (monitored) clusters because training is compute-intensive. Because these larger clusters are monitored, it would be possible to detect them doing so, specifically by reviewing the AI training on monitored clusters to determine if it

uses novel methods of relevance (the results of restricted research). This check would ensure that all the AI algorithms used were either previously known or irrelevant to frontier AI, and the code and algorithms could be reviewed in an automated fashion to preserve privacy and security while ensuring compliance. Given that the mechanism can only be applied to monitored chips, it cannot preclude algorithmic progress being conducted in secret data centers and applied to training in secret data centers. Even with that limitation, this mechanism adds value by preventing new algorithms from being applied to monitored chips, a category which, under most plausible restriction regimes, includes all of the largest data centers.

Monitoring AI inference content for restricted research Apply classifiers to the content of all inference workloads (user prompts, tool use, model outputs) to detect restricted research. If chips are running inference on an AI model that is capable of restricted research, it is necessary to ensure that the model is not being used for this purpose. One difficulty is that AI models or users might obfuscate the workload, tricking the classifier into thinking the workload is permitted when it is not, similar to jailbreaking strategies that bypass external classifiers (Davies et al., 2026). Another difficulty is that this approach can only be applied to monitored chips; because inference requires relatively few chips (DeepSeek AI, 2025), this mechanism would only obtain high confidence if virtually all of the relevant chip clusters were monitored and had this classification applied. While many AI companies already deploy classifiers on their own inference workloads to detect prohibited uses (Sharma et al., 2025; OpenAI, 2025), there would be a small difficulty in adding AI R&D to the set of prohibited uses and a large difficulty in proving comprehensive monitoring to external inspectors.

Review AI training data for prohibited data improvements Log and review AI training data to verify the data complies with research restrictions. Improvements to the data used to train AIs can contribute to dangerous capabilities, and thus might be restricted by an agreement, but such improvements could be difficult to detect. Choi et al. (2023) propose methods of verifying that certain data was used to train a particular model, but further research is needed to make these methods implementation-ready. Given such methods, however, inspectors could then attempt to confirm whether the verified training data complies with restrictions. For example, if there are rules limiting the use of synthetic data, inspectors could apply AI-text-detection methods to the training data. An agreement may find it challenging to define and classify data improvements; potential solutions may include restricting the amount of AI-generated data, restricting which models can be used for synthetic data generation, or restricting what domains the data covers. Current AI development involves classifying and filtering training data, for instance to remove CSAM, but this filtering is typically not verified by an external party.

Broad chip-use monitoring for restricted experiments Classify the activities of AI chips to detect restricted research. Many algorithmic advances require experiments to test them. This experimentation would be prohibited by an agreement, and chip-use verification could be applied to ensure chips are not running such experiments. For instance, automated classifiers could analyze characteristics of the workloads being run—power draw, memory bandwidth utilization, user organization—to determine if these are likely part of a restricted research project (Tang et al., 2022; Rahman & Tajdari, 2026). If restrictions prohibit new AI training, classification could be aimed at differentiating training from inference. However, most realistic definitions of restricted research would not be so simple to enforce via classification of chip activities—ease of classification depends on how restrictions are defined. Classification can only cover monitored chips, and due to the relatively low number of chips required for research, this monitoring should be widely applied to make this approach work well. In a structurally similar practice, cloud providers offer tools for detecting unauthorized cryptocurrency mining (Google, 2026).

Restricting chips to only run approved models Ensure that AI chips do not run models capable of restricted research. This could involve an allowlist (approve only certain models), a blocklist (prohibit only certain models), or some other method of differentiating models (e.g., allow models below a certain size). Because AI capabilities are difficult to verify, it is hard to rule out that a model might conduct restricted research, and this state of affairs suggests an allowlist. Identifying models for an allowlist might involve data filtering or "unlearning" to limit AI skill at restricted research, training models to refuse it, or evaluating models to confirm they cannot perform it. At present, all of these methods face significant challenges in robustness, but they may nonetheless add value. Further challenges emerge in requiring chips to attest to the identity of models being run, given the proprietary nature of current AI software and hardware stacks, and like many alternatives, this mechanism can only be applied to monitored chips. The closest precedent is Digital Rights Management (DRM), which uses cryptographic checks to prevent unapproved uses of digital media or software, but DRM has a poor track record (Biddle et al., 2003; Patat et al., 2022).

3.4. Non-proliferation

Some mechanisms broadly aim at preventing dangerous resources or knowledge from reaching bad actors. Prevention is typically not labeled as a form of verification, but we discuss these non-proliferation options because they would

correctly cause a verifying party to have greater assurance that rules are being followed. In the context of AI governance, non-proliferation is typically applied to AI chips, for instance via export controls. Chip-level non-proliferation would be important for verifying an international agreement, but here we defer to prior work on the topic (Fist & Grunewald, 2023; Allen, 2022; Scher et al., 2025), and instead focus on mechanisms more uniquely relevant to verifying research restrictions.

Preventing the development of AI-R&D-capable AIs Prevent, as a first line of defense, the development of AIs advanced enough to automate restricted research. Prior to an international agreement, domestic regulation could delay or partially limit such development. With an agreement in place, AI development can be restricted more effectively. An agreement might require that AI training have no more than 5% of its training data be coding tokens and that the final model score below some threshold on AI R&D benchmarks. Data use could be verified by a combination of training-data identification and classification as code or non-code. AI R&D potential could be verified by evaluations, which are likely to include fine-tuning and other elicitation strategies; note, however, that upper bounding an AI's capabilities is not currently feasible (Barnett & Thiergart, 2024). This mechanism complements restrictions on the models that AI chips may run, as it seeks to prevent dangerous models from being developed at all.

Preventing the proliferation of AI-R&D-capable AIs Prevent, as a second line of defense, the spread of AI systems that could automate restricted research. Under most plausible regimes, automating restricted research would require both unmonitored chips and the weights of a sufficiently capable AI. Access to the latter might be limited by heightened security in AI companies, regulations and international norms around releasing model weights, and explicit corporate non-proliferation policies. Note, however, that corporate security is unlikely to stand up to attacks by well-resourced state actors. This mechanism complements restrictions on the models that AI chips may run, as it seeks to prevent the proliferation of dangerous models. The Wassenaar Arrangement's 2013 addition of "intrusion software" to its dual-use control list is the closest precedent for multilateral export controls on a category of software (Wassenaar Arrangement, 2013).

Non-proliferation of AI algorithms and data Secure the algorithms and data available to AI research organizations today. To the extent that groups capable of frontier AI research do not publish or leak their methods, knowledge proliferation will be limited to what can be gleaned from the open-source community, perhaps several months behind the frontier. Non-proliferation could apply to both algorithmic secrets and to data that advances AI capabilities (e.g., high-quality RL environments). Even if state actors have extracted some secrets, these secrets might lag the frontier, and these state actors could implement their own security against further proliferation. This mechanism would initially delay restricted research projects, but it would not provide ongoing assurance that such projects have not advanced. One difficulty is the fact that algorithmic secrets are often small pieces of knowledge that are hard to secure. The Atomic Energy Act's "born secret" doctrine provides precedent for non-proliferation of research ideas, as it renders information pertaining to the design of nuclear weapons classified by default (United States Congress, 1954).

4. Discussion and Conclusion

We have analyzed both key considerations affecting the verifiability of research restrictions and specific high-level mechanisms that could be used to implement those restrictions. This analysis leads to three main takeaways.

First, the most difficult challenge is the risk of secret data centers that are not known to verification authorities. These data centers could be used for experimentation (testing research ideas as part of the illicit research process) and for unmonitored inference (AIs thinking through research ideas and writing code for experiments). This consideration is especially important if AI models become sufficiently capable at AI R&D tasks before an agreement. Before that point, both human experts and compute are necessary components for most AI R&D, but after that point, only compute is needed.

Second, a few mechanisms stand out as especially useful. Whistleblowers and intelligence gathering are well established, and would likely be the most effective methods for detecting covert projects. Meanwhile, the best methods for monitoring declared organizations would be automated methods, such as automated code reviews and inference monitoring. These methods would aim to identify illicit experiments or AIs thinking about and writing code for restricted research. They could be designed to uphold privacy and security by default, only escalating to human review when needed. Also especially valuable is reviewing AI training for novel methods of relevance, which may prevent covert projects from using monitored chips to scale any restricted research they conduct.

Third, available mechanisms vary widely in ease of implementation. Whistleblowers, interviews, traditional intelligence gathering, and multi-party compliance attestations would be straightforward to implement, while some mechanisms first require significant R&D, especially chip-use verification.

One major risk related to the mechanisms in this report is that, if implemented without adequate safeguards, they

could be misused by bad actors for purposes such as political repression, suppression of legitimate research, surveillance overreach, and security leakage. Some mechanisms, such as whistleblowers, are minimally invasive and pose little risk of co-option, while others would require extremely careful implementation to avoid misuse. Additionally, well-meaning actors could implement restrictions poorly, causing collateral damage to unrestricted research or other harms. We hope future work builds implementation-ready versions of some of the mechanisms in this report while carefully considering security, privacy, and efficacy.

Overall, we remain uncertain about how easy it will be for countries to verify restrictions on frontier AI research in an international agreement; there are points favoring both optimism and pessimism. Insofar as it is necessary to have research restrictions and verify compliance with them, this work advances the state of thinking on the topic by highlighting important considerations and potential mechanisms—some of which require urgent R&D if they are to be ready soon.

Impact statement

This work aims to advance the field of technical AI governance by exploring verification methods. This topic is important to the viability of international agreements, and providing more detailed proposals for verification can make conversations about such agreements more concrete. There are risks from this work, including that verification mechanisms could be misappropriated toward harmful ends, and that research on verification could serve as a guide for evaders. We believe our counterfactual impact on both of these is sufficiently small that it is overwhelmed by the benefit of creating more detailed AI governance proposals.

LLM usage statement

LLMs were primarily used to convert this paper into LaTeX and to review drafts for typos. They were also used occasionally in brainstorming, locating citations, or to assist in rewriting particular sentences. The vast majority of the ideas in this paper and final text were human-contributed.

References

Allen, G. C. Choking off China's Access to the Future of AI. Technical report, October 2022. URL https://www.csis.org/analysis/choking-chinas-access-future-ai.

Altman, S. and Pachocki, J. Built to benefit everyone: our plan, June 2026. URL https://openai.com/index/built-to-benefit-everyone-our-plan/.

Anthropic. System Card: Claude Mythos Preview, April 2026. URL https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf.

Baker, M., Kulp, G., Marks, O., Brundage, M., and Heim, L. Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment, July 2025. URL https://arxiv.org/abs/2507.15916v2.

Barnett, P. Compute Requirements for Algorithmic Innovation in Frontier AI Models, July 2025. URL https://arxiv.org/abs/2507.10618v1.

Barnett, P. and Scher, A. AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions, May 2025. URL https://arxiv.org/abs/2505.04592v1.

Barnett, P. and Thiergart, L. What AI evaluations for preventing catastrophic risks can and cannot do, November 2024. URL http://arxiv.org/abs/2412.08653. arXiv:2412.08653 [cs].

Biddle, P., England, P., Peinado, M., and Willman, B. The Darknet and the Future of Content Distribution. In *ACM Workshop on digital rights management*, volume 6, pp. 54, October 2003. URL https://www.semanticscholar.org/paper/The-Darknet-and-the-Future-of-Content-Distribution-Biddle-England/d09ea4f3bb7a0b8ffc596eaa92af8ae367485358.

Brass, A. and Aarne, O. Location Verification for AI Chips, April 2024. URL https://www.iaps.ai/research/location-verification-for-ai-chips.

Buchanan, B. The AI Triad and What It Means for National Security Strategy, August 2020. URL https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy/.

Center for AI Safety. Statement on AI Risk | CAIS, 2023. URL https://aistatement.com.

Choi, D., Shavit, Y., and Duvenaud, D. Tools for Verifying Neural Models' Training Data, July 2023. URL http://arxiv.org/abs/2307.00682. arXiv:2307.00682.

Committee of Privy Counsellors. Review of Intelligence on Weapons of Mass Destruction. Technical report, The Stationery Office, London, 2004.

Davidson, T. Will compute bottlenecks prevent a software intelligence explosion? — AI Alignment Forum. April 2025. URL https://www.alignmentforum.org/posts/XDF6ovePBJf6hsxGj/will-compute-bottlenecks-prevent-a-software-intelligence-1.

Davies, X., Giglemiani, G., Lau, E., Winsor, E., Irving, G., and Gal, Y. Boundary Point Jailbreaking of Black-Box LLMs, February 2026. URL http://arxiv.org/abs/2602.15001. arXiv:2602.15001 [cs].

Dean, J. and Shazeer, N. Jeff Dean & Noam Shazeer — 25 years at Google: from PageRank to AGI, February 2025. URL https://www.dwarkesh.com/p/jeff-dean-and-noam-shazeer.

DeepSeek AI. Day 6: One More Thing, DeepSeek-V3/R1 Inference System Overview, February 2025. URL https://github.com/deepseek-ai/open-infra-index/blob/main/202502OpenSourceWeek/day_6_one_more_thing_deepseekV3R1_inference_system_overview.md.

Ecoffet, A., Altman, S., Achiam, J., and Nicholson, C. Event Replay: Sam Altman on Building the Future of AI - Video, April 2026. URL https://forum.openai.com/public/videos/event-replay-sam-altman-on-building-the-future-of-ai-2026-04-06.

Favaro, M. and Clark, J. When AI builds itself, June 2026. URL https://www.anthropic.com/institute/recursive-self-improvement.

Fist, T. and Grunewald, E. Preventing AI Chip Smuggling to China. Technical report, Center for New American Security, October 2023. URL https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china.

Future of Life Institute. Statement on Superintelligence, September 2025. URL https://superintelligence-statement.org.

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., and Elhage, N. Predictability and surprise in large generative models. pp. 1747–1764, 2022.

Google. Cryptomining detection best practices | Security Command Center, April 2026. URL https://docs.cloud.google.com/security-command-center/docs/cryptomining-detection-best-practices.

Gundlach, H., Fogelson, A., Lynch, J., Trisovic, A., Rosenfeld, J., Sandhu, A., and Thompson, N. On the Origin of Algorithmic Progress in AI, November 2025. URL https://arxiv.org/abs/2511.21622v1.

Harack, B., Trager, R. F., Reuel, A., Manheim, D., Brundage, M., Aarne, O., Scher, A., Pan, Y., Xiao, J., and Loke, K. Verification for international AI governance. *AI Governance Initiative, Oxford Martin School, University of Oxford*, 2025.

Heim, L. and Koessler, L. Training Compute Thresholds: Features and Functions in AI Regulation, August 2024. URL http://arxiv.org/abs/2405.10799. arXiv:2405.10799 [cs].

Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., Osborne, M. A., and Zilberman, N. Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation, March 2024. URL http://arxiv.org/abs/2403.08501. arXiv:2403.08501.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. Algorithmic progress in language models, March 2024. URL http://arxiv.org/abs/2403.05812. arXiv:2403.05812 [cs].

Hooker, S. On the Limitations of Compute Thresholds as a Governance Strategy, July 2024. URL http://arxiv.org/abs/2407.05694. arXiv:2407.05694 [cs].

Mark, A. and Scher, A. Constitutional Law and AI Governance: Constraints on Model Licensing and Research Classification, September 2025. URL https://arxiv.org/abs/2509.05361v2.

Martin, S., Bullock, J., and Katzke, C. Analysis of Global AI Governance Strategies | Convergence Analysis, December 2024. URL https://www.convergenceanalysis.org/research/analysis-of-global-ai-governance-strategies.

MSL Preparedness & Red Teaming & Alignment Team and MSL AI Security Team. Muse Spark Safety & Preparedness Report. Technical report, April 2026. URL https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report/.

Nagin, D. S. Deterrence: A Review of the Evidence by a Criminologist for Economists. *Annual Review of Economics*, 5(Volume 5, 2013):83–105, August 2013. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-072412-131310. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-072412-131310.

National Research Council. *The Polygraph and Lie Detection*. National Academies Press, Washington, D.C, 2003. ISBN 978-0-309-08436-9.

OpenAI. Preparing for future AI capabilities in biology, 2025. URL https://openai.com/index/preparing-for-future-ai-capabilities-in-biology/.

Patat, G., Sabt, M., and Fouque, P.-A. Exploring Widevine for Fun and Profit. In *2022 IEEE Security and Privacy Workshops (SPW)*, pp. 277–288, May 2022. doi: 10.1109/SPW54247.2022.9833867. URL https://ieeexplore.ieee.org/abstract/document/9833867. ISSN: 2770-8411.

Petrie, J. Near-Term Enforcement of AI Chip Export Controls Using A Firmware-Based Design for Offline Licensing, May 2024. URL http://arxiv.org/abs/2404.18308. arXiv:2404.18308.

Philippe, S., Glaser, A., and Felten, E. W. A Cryptographic Escrow for Treaty Declarations and Step-by-Step Verification. *Science & Global Security*, 27(1):3–14, January 2019. ISSN 0892-9882, 1547-7800. doi: 10.1080/08929882.2019.1573483. URL https://www.tandfonline.com/doi/full/10.1080/08929882.2019.1573483.

Rahman, R. and Tajdari, S. Detecting Hidden ML Training With Zero-Overhead Telemetry. In *Technical AI Governance Research Workshop at the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2606.19262.

Ramiah, A. A., Koopmanschap, R., Thorsteinson, J., Khan, S., Zhou, J., Noh, S., Meindertsma, J., and Shafiq, F. Toward a Global Regime for Compute Governance: Building the Pause Button, June 2025. URL https://arxiv.org/abs/2506.20530v1.

Reedi, A. J., Bonamy, H., Di Cosmo, Y., von Werra, L., and Tunstall, L. ml-intern: an agent that autonomously researches, writes, and ships good quality ML related code using the Hugging Face ecosystem, 2026. URL https://github.com/huggingface/ml-intern.

Scher, A. and Thiergart, L. Mechanisms to Verify International Agreements About AI Development, November 2024. URL https://arxiv.org/abs/2506.15867.

Scher, A., Abecassis, D., Barnett, P., and Abeyta, B. An International Agreement to Prevent the Premature Creation of Artificial Superintelligence, November 2025. URL https://arxiv.org/abs/2511.10783v2.

Schlesinger, J. Memorandum From Director of Central Intelligence Schlesinger to All Central Intelligence Agency Employees, May 1973. URL https://www.cia.gov/readingroom/docs/memo%20for%20all%20cia%20employee%5B15132481%5D.pdf.

Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., Ong, E., Peng, A., Agarwal, R., Anil, C., Askell, A., Bailey, N., Benton, J., Bluemke, E., Bowman, S. R., Christiansen, E., Cunningham, H., Dau, A., Gopal, A., Gilson, R., Graham, L., Howard, L., Kalra, N., Lee, T., Lin, K., Lofgren, P., Mosconi, F., O'Hara, C., Olsson, C., Petrini, L., Rajani, S., Saxena, N., Silverstein, A., Singh, T., Sumers, T., Tang, L., Troy, K. K., Weisser, C., Zhong, R., Zhou, G., Leike, J., Kaplan, J., and Perez, E. Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming, January 2025. URL http://arxiv.org/abs/2501.18837. arXiv:2501.18837 [cs].

Shavit, Y. What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring, March 2023. URL https://arxiv.org/abs/2303.11341v2.

Tang, B. J., Chen, Q., Weiss, M. L., Frey, N. C., McDonald, J., Bestor, D., Yee, C., Arcand, W., Bergeron, W., and Byun, C. The MIT Supercloud Workload Classification Challenge. pp. 708–714. IEEE, 2022. ISBN 1-6654-9747-5.

United States Congress. Invention Secrecy Act of 1951, February 1952. URL https://www.govinfo.gov/link/statute/66/3. Published: Pub. L. No. 82-256, ch. 4, 66 Stat. 3.

United States Congress. Atomic Energy Act of 1954, 1954. Published: 42 U.S.C. § 2014(y).

United States Congress. Sarbanes-Oxley Act of 2002, July 2002.

U.S. Nuclear Regulatory Commission. Resident Inspector Program Overview, 2025. URL https://www.nrc.gov/reactors/operating/oversight/rop-description/resident-insp-program.

Wasil, A. R., Reed, T., Miller, J. W., and Barnett, P. Verification methods for international AI agreements, August 2024. URL https://arxiv.org/abs/2408.16074v2.

Wassenaar Arrangement. List of Dual-Use Goods and Technologies and Munitions List, December 2013. URL https://www.wassenaar.org/app/uploads/2019/consolidated/WA-LIST%20%2813%29%201.pdf. Published: WA-LIST (13) 1.

Whitfill, P. and Wu, C. Will Compute Bottlenecks Prevent an Intelligence Explosion?, July 2025. URL https://arxiv.org/abs/2507.23181v2.

Yudkowsky, E. and Soares, N. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company, New York, 2025. ISBN 978-0-316-59564-3.