Faster AI Diffusion Through Hardware-Based Verification

How to use privacy-preserving verification in the AI hardware stack to build trust and limit misuse

Nora Ammann and David ‘davidad’ Dalrymple

Faster AI Diffusion Through Hardware-Based Verification

How to use privacy-preserving verification in the AI hardware stack to build trust and limit misuse | Nora Ammann and David 'davidad' Dalrymple

Written in the authors' personal capacity

This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

Summary

There's often thought to be an irreconcilable tension between AI diffusion and security. Widespread access to powerful AI systems increases the risk of deliberate misuse by malicious actors or loss of control, while tight restrictions on access can stifle innovation and entrench power asymmetries. Developing hardware-enabled verification capabilities would allow us to escape that tradeoff. These capabilities would make it possible to verify complex claims about AI development and usage without exposing sensitive information. For example, AI developers could prove that their model passed certain safety evaluations, or that it was trained with specific architectural or data safeguards in place. Data center providers could prove that the workloads they host have not been altered or sabotaged. The technology could also make it feasible to export AI chips and capabilities broadly, while protecting against IP theft and maintaining fine-grained guarantees about how these chips can be used. As a result, these hardware-enabled verification capabilities would reduce trust- and security-based barriers to widespread AI adoption, simplify compliance, defend intellectual property, and significantly reduce the security risks of AI diffusion — while still preserving privacy and avoiding intrusive surveillance. To enact this vision, we

recommend a dedicated, publicly-backed, open R&D initiative — like a DARPA-style program or a Focused Research Organization — to provide policymakers with a powerful new instrument for fostering responsible AI progress.

Motivation

The missing tools to achieve secure and widespread AI adoption

There are countless gains to be had in allowing AI to be deployed broadly: acceleration of scientific progress, supercharging of economic growth, and hardening of critical infrastructure, to name a few. But AI diffusion, whether at a national or global scale, also increases the likelihood that AI systems will be stolen, sabotaged, or misused. Diffusion without security risks irreversibly proliferating dangerous capabilities. The negative consequences might be numerous: enabling authoritarian, criminal, or adversarial actors to execute cyberattacks or develop novel bioweapons, for instance.

To facilitate widespread, secure adoption, policymakers and end users alike must be able to trust AI's capabilities, security, and safety. This points to the need for reliable mechanisms to verify that AI systems are secure, do what they claim, and are not being misused — but without resorting to intrusive government oversight and surveillance that would clash with fundamental values of freedom and privacy. Unfortunately, the current AI technology stack does not provide trustworthy, privacy-preserving verification mechanisms. This forces today’s policy options into an uneasy tradeoff: either forgo oversight and risk misuse, or implement intrusive oversight and sacrifice privacy. Hardware-enabled verification would break this deadlock by giving end users the means to prove no misuse, without revealing other sensitive data.

The availability of such mechanisms would not just help secure against future risks, but would also offer immediate benefit to actors across the burgeoning AI economy. The demand for verification exists across the ecosystem: AI developers want to be able to make credible claims about their models’ capabilities or safety

architectures, while keeping their proprietary methods and sensitive data protected; data center providers want to attest to the confidentiality and integrity of the workloads they host; customers want to be able to trust in the systems they use, and have confidence that they are getting the models the providers are claiming, with the properties they claim. But today, the technological means to do this remain meager.

To use a comparison from the history of technology, the AI market is currently like the market for used cars before the introduction of the odometer (a device that measures the total distance traveled by a vehicle). Before the odometer, used car sellers had asymmetrically more information than buyers. Buyers were concerned about the cars being in much worse shape than expected, making them reluctant to buy. After the odometer was introduced, sellers could prove how used their car was, which enabled this market to thrive. In a similar way, enhancing developers' ability to make trustworthy claims about their models would significantly enable the widespread use of these models in critical or high-stakes deployment situations.

As it stands, no one can make verifiable claims about how they train AI models or what safeguards they have implemented, without full access to the models themselves. Conducting an AI evaluation — part of today's standard toolkit for assessing the performance and reliability of artificial intelligence systems — currently requires a cumbersome, expensive, and largely trust-based process. Evaluations can also impose significant information disclosure requirements on AI developers, by requiring them to submit their most valuable intellectual property – the AI models themselves – to a third-party auditor. At the same time, without access to that information, these evaluations risk being ultimately ineffective and untrustworthy, as they aren't able to offer high confidence that a given evaluation result was run on the model in question, or that no further changes have been made to the model after the evaluation has been passed.

Updating the hardware stack to build privacy and trust into the foundations of AI systems

New mechanisms, built into the hardware of AI chips themselves, would make it possible to verify — that is, produce highly reliable evidence about — critical

claims concerning the development and use of AI. These same mechanisms could also provide guarantees against misuse, by enforcing specific guardrails that preempt violations, all while preserving privacy.

Theoretically, verifying some of these claims would be achievable without new hardware elements, by using software-only solutions such as zero-knowledge proofs. But to date, such cryptography-based methods remain egregiously expensive in the context of AI (despite continual progress motivated by blockchain applications, which have vastly smaller quantities of data to verify). While these capabilities should no doubt be pursued, it is hard to achieve sufficient levels of security without grounding the security properties all the way into the hardware layer. In the absence of this, numerous low-cost methods exist to tamper with and undermine security and monitoring methods.

The new hardware mechanisms would be added directly to AI chips, combining a tamper-proof enclosure with an auxiliary guarantee processor capable of verifying, and optionally enforcing, claims about how the chips are used. This would enable:

1. **Privacy-preserving AI evaluation:** These hardware mechanisms could attest, directly and locally, that an AI model has successfully passed an evaluation, without needing to reveal any information beyond that fact.
2. **Compute thresholds:** The mechanisms could assess the amount of training compute used to train an AI model — which is an important, albeit not perfect, proxy for the model's capabilities — and provide auditors greater insight into the state of the AI frontier. They could also be used to enforce compute thresholds, above which a model cannot be trained, or only with a valid license.
3. **Location verification:** Similarly, the mechanisms could verify the geographical location of a chip or compute cluster, and make it possible to adjust usage policies based on location.
4. **Model safeguards:** More ambitiously, the technology could even enable direct hardware attestation that an AI model implements specific architectural features, such as safety fine-tuning, or has not been trained on certain types of data, such as data relevant to the development of biological weapons.
5. **Protection against weight theft:** The same technological primitives would also improve our ability to protect model weights — often the most commercially

and strategically sensitive software component of modern AI systems — from being stolen, by enabling cryptographic capabilities locally at the level of the AI hardware itself.

6. **Prevention of model sabotage:** As AI becomes increasingly integrated into critical processes, the stakes of undetected sabotage of AI workloads rise. Hardware-enabled verification could help detect or prevent sabotage by ensuring the AI workload being trained precisely matches developer specifications and has not been tampered with, protecting integrity from development to deployment.
7. **Export:** By exporting AI chips with integrated verification and anti-tamper mechanisms alongside AI models that only run on such guaranteeable hardware, it becomes possible to ensure that embedded safety and security features remain uncompromised, regardless of where or how the model is used. This would make it feasible to export AI chips broadly, while protecting against IP theft and maintaining fine-grained guarantees about how these chips can be used (like the ones listed above).

While the existence of these capabilities would benefit the AI ecosystem as a whole, market participants are unlikely to develop them on their own, and even less likely to do so in a timely manner. None of the usual industry players possess the unique combination of skills, resources, and incentives required to build this technology. More broadly, critical public goods like public safety and cybersecurity are often undersupplied by the private sector (as evidenced, for example, by the persistent costs of ransomware). This technology, specifically, would not be developed at the pace required to incorporate it in AI data center build-outs over the critical next few years, which will constitute most of the future's AI hardware.

Solution

Bringing the vision of hardware-enabled verification and guarantee technologies to fruition demands a focused R&D initiative. The goal isn't necessarily for a publicly funded program to achieve full production readiness. Rather than direct industrial implementation, the goal is to thoroughly demonstrate its technical feasibility,

driving agreement on industry standards, and ultimately empowering industry to independently implement these capabilities to open specifications.

Program design

A concerted R&D program must integrate three key ingredients: funding, talent, and speed. Done correctly, we believe a 3-year time-bound program, funded at around $30 million could achieve sufficient progress to turn the stack over to industry to pursue.

This effort could be pursued as a DARPA-style R&D program, which has a proven track record of success. Another promising option is to establish a Focused Research Organization (FRO), funded publicly, privately, or both. FROs are typically not-for-profit entities designed to tackle specific, high-impact scientific or technical problems. Unlike traditional academic labs or for-profit ventures, FROs pursue well-defined, time-bound technical milestones to create public goods that address research bottlenecks, especially in areas not immediately of high-priority interest to private investors.

While a DARPA-style program offers the benefit of scale which enables the exploration of multiple parallel hypotheses through separate, competing teams, an FRO can leverage its highly integrated operational model and startup-like agility for enhanced speed and focus.

This will require the government to depart from a typical model of scientific funding whereby an R&D goal is attacked through a disconnected collection of academic grants. Such an approach risks resulting in incremental, slow, and scattered efforts that lack sufficient relevance to frontier hardware. Similarly, it's vital for this work to occur in the open. As decades of experience in cybersecurity have demonstrated, “sunshine is the best disinfectant:” building securely means building openly. Furthermore, openness allows a broader community of practitioners to understand, contribute to, test, and — most importantly — come to trust what's being built.

In an initial R&D sprint, a team comprising experts from across the hardware and software stack would develop these hardware-based verification capabilities into a mature prototype, targeting a Technology Readiness Level (TRL) of 5–6 by

demonstrating its functionality in a relevant environment. Throughout this process, active engagement with relevant stakeholders (including hyperscalers, frontier labs, chip designers, and manufacturers) is paramount in order to ensure that the technology is capable of integrating with the existing stack. Once TRL 5–6 is achieved, the collaboration with key industry stakeholders would intensify in order to develop the design to full maturity, accounting for practical constraints ranging from energy costs and performance metrics to maintenance procedures, and ultimately to drive adoption across the industry.

Technical objectives

To realize the full potential of secure AI diffusion, these hardware-enabled mechanisms must incorporate the following key high-level design features:

- **Verifiability:** The hardware design needs to provide the ability to cryptographically prove claims about AI development and usage. Specifically, it must provide the technical foundation to prove, for instance, how much compute an AI model was trained with, that it passed a certain safety test, or that its architectural features match its specifications, all through secure, hardware-enabled cryptographic attestation.
- **Privacy-preservation:** The design needs to ensure that sensitive or proprietary information (e.g., model weights, training data) remains confidential while verification occurs. This can be achieved by leveraging local verification, on-chip de/encryption capabilities, and remote attestation from a physically unclonable private key. The crux is to eliminate the need to send sensitive data to third parties entirely, protecting intellectual property and privacy by design.
- **Security:** The design must adhere to the highest security standards, incorporating a robust root of trust and resilience against sophisticated side-channel attacks. Ideally, it should feature active security measures such as tamper-responsive mechanisms that detect and react to unauthorized physical or electrical interference, making it extremely difficult to compromise.
- **Auditability:** To build trust, the design of this hardware should be developed openly and transparently. It should allow independent security researchers and experts to thoroughly verify its integrity and confirm the absence of hidden

vulnerabilities or backdoors. This commitment to openness is vital for widespread adoption and trust.

- **Flexibility and generality:** The verification capabilities shouldn't be limited to a narrow set of applications. Instead, they need to be broadly applicable, capable of supporting a wide range of verification regimes across diverse AI applications. This flexibility is essential to enable the hardware, once deployed, to adapt seamlessly to evolving policy needs, scientific advancements, and unforeseen future AI use cases.
- **Updatability**: Building on the above, the design must permit secure, post-deployment updates to its firmware and security features. This capability is critical for addressing newly discovered threats, patching vulnerabilities, or adapting to changes in policy requirements as the AI landscape continues to shift rapidly.

This can be achieved by integrating the following key components:

1. **A guarantee processor** able to encrypt and authenticate data coming from the AI chip and from the guarantee processor itself, and thus enable privacy-preserving, sophisticated, and programmable verification schemes.
2. **An anti-tamper enclosure** to protect the system from snooping or interference, including a tamper-detection system which, if triggered, would activate mechanisms that wipe any secret information, and render the chip non-functional to prevent the potentially-compromised chip from being misused.
3. **A secure updating mechanism** for the guarantee processor's firmware in order to keep all devices up-to-date on security updates and possible rule-set changes
4. **An interlock** between the guarantee processor and the AI chip's data path to enable robust verification, without needing to trust the rest of the hardware stack.

Further detail is available in the author's three-part report on Flexible, Hardware-Enabled Guarantees.

Further Resources

About Flexible Hardware-Enabled Guarantees

- flexHEG, flexHEG.com, n.d.
- flexHEG Report Series, (1) Overview, (2) Technical Options, (3) International Security Applications, 2025.
- Davidad, Principles of flexHEG, FlexHEG Builder Workshop, Berkeley, 2025

Additional background on AI compute and hardware-enabled mechanisms

- Center for the Governance of AI, Computing Power and the Governance of AI, 2024.
- Center for Security & Emerging Technologies, The AI Triad and What It Means for National Security Strategy, 2020.
- RAND, Hardware-Enabled Governance Mechanisms, 2024.
- Center for a New American Security, Secure, Governable Chips, 2024.