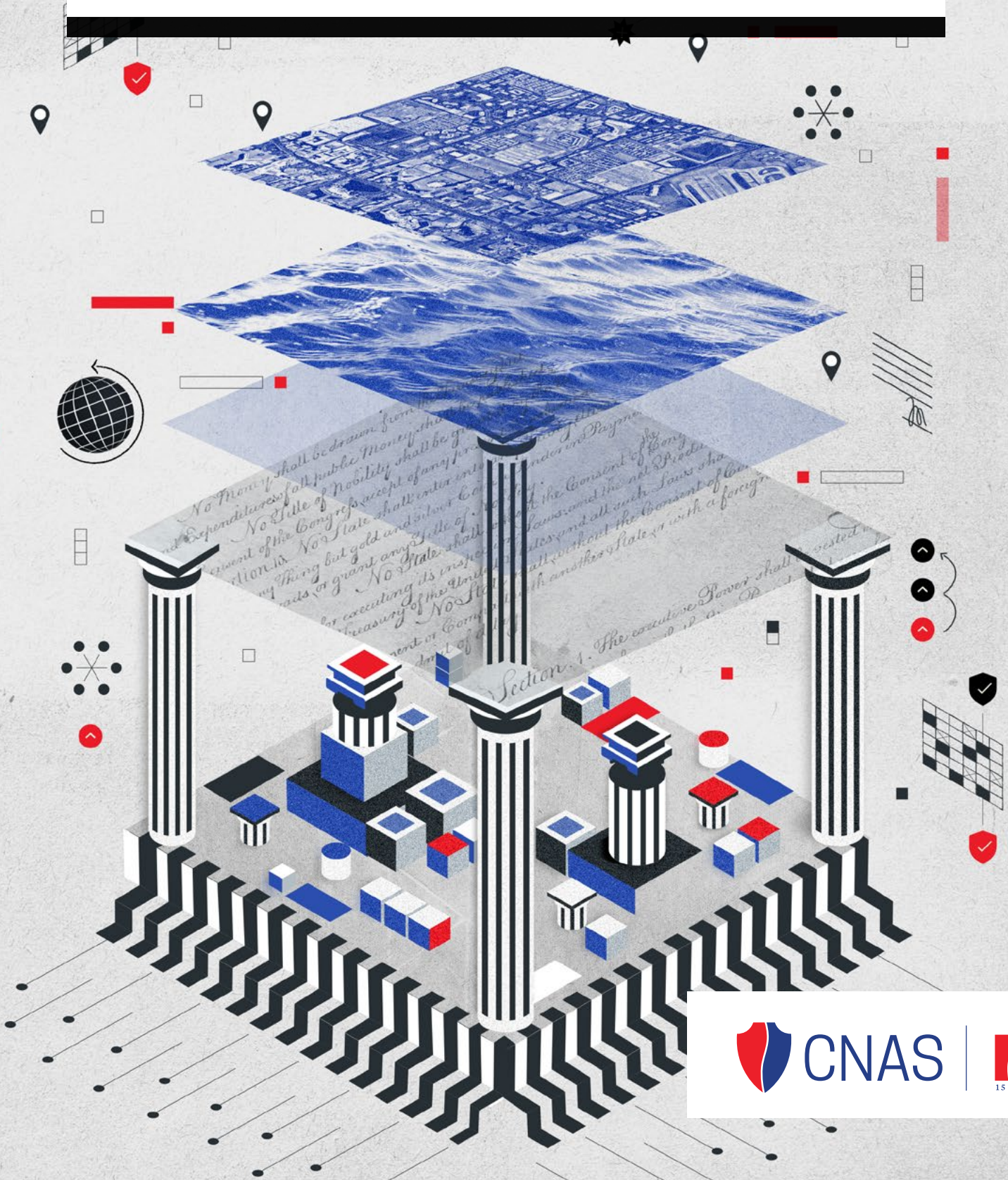


JANUARY 2024

# Secure, Governable Chips

Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing

Onni Aarne, Tim Fist, and Caleb Withers



CNAS



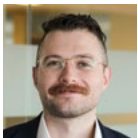
## About the Authors



**Onni Aarne** is a consultant with the compute governance team at the Institute for AI Policy and Strategy. He previously conducted compute governance research at Rethink Priorities, another research nonprofit organization. He has a BSc in computer science and an MSc in data science from the University of Helsinki.



**Tim Fist** is a Fellow with the Technology and National Security Program at the Center for a New American Security (CNAS). He has an engineering background and previously worked as the Head of Strategy & Governance at Fathom Radiant, an AI hardware company. Prior to that, he worked as a machine learning engineer, building and deploying AI systems in commercial settings. He holds a B.A. (Honors) in aerospace engineering and a B.A. in political science from Monash University.



**Caleb Withers** is a Research Assistant for the Technology and National Security Program at CNAS. Before CNAS, he worked as a policy analyst for a variety of New Zealand government departments. He has an M.A. in security studies from Georgetown University, concentrating in technology and security, and a Bachelor of Commerce from Victoria University of Wellington, majoring in economics and in information systems.

## About the Technology and National Security Program

The CNAS Technology and National Security program explores the policy challenges associated with emerging technologies. A key focus of the program is bringing together the technology and policy communities to better understand these challenges and together develop solutions.

## About the Artificial Intelligence Safety & Stability Project

The CNAS AI Safety & Stability Project is a multiyear, multiprogram effort that addresses the established and emerging risks associated with artificial intelligence. The work is focused on anticipating and mitigating catastrophic AI failures, improving the U.S. Department of Defense's processes for AI testing and evaluation, understanding and shaping opportunities for compute governance, understanding Chinese decision-making on AI and stability, and understanding Russian decision-making on AI and stability.

## Acknowledgments

The authors would like to acknowledge the CNAS Publications Teams for their support, design, and editing. The authors also would like to thank Paul Scharre, Executive Vice President and Director of Studies, for reviews of various iterations of this work. This report was produced in collaboration with the Institute for AI Policy and Strategy. The authors would also like to thank the large number of reviewers and experts consulted as part of this project, especially Samuel Hammond, Brady Helwig, and Gabriel Kulp. This project is made possible with the generous support of Open Philanthropy.

As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues, and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its [website](#) annually all donors who contribute.

# TABLE OF CONTENTS

<b>01</b>	<b>Executive Summary</b>
<b>05</b>	<b>Introduction</b>
<b>09</b>	<b>What Would Effective On-Chip Governance Look Like?</b>
<b>10</b>	<b>Policies that On-Chip Governance Mechanisms Could Enable</b>
<b>13</b>	<b>Technical Underpinnings</b>
<b>17</b>	<b>Challenges for Implementation</b>
<b>21</b>	<b>Implementation Timelines</b>
<b>23</b>	<b>Recommendations</b>
<b>25</b>	<b>Limitations and Conclusion</b>
<b>26</b>	<b>Appendix A: Glossary for AI Compute</b>
<b>28</b>	<b>Appendix B: Additional Security Considerations</b>

**On-chip governance mechanisms can safeguard the development and deployment of broadly capable AI and supercomputing systems in a way that is complementary to American technology leadership.**

## **Executive Summary**

**B**roadly capable AI systems, built and deployed using specialized chips, are becoming an engine of economic growth and scientific progress. At the same time, these systems also could be used by irresponsible actors to enable mass surveillance, conduct cyberattacks, and design novel biological weapons. This makes securing and governing the supply chain for AI chips important for mitigating risks to U.S. national security. But today's semiconductor export controls are lackluster as a stand-alone solution. To be effective, they need to be far-reaching, which harms the competitiveness of U.S. firms, risks the "de-Americanization" of chip supply chains, and risks alienating commercial AI developers and partner nations. Far-reaching controls are also hard to enforce: AI chip smuggling is already happening today and could significantly grow in volume over the coming years.<sup>1</sup>

The unique challenges of AI governance and the opportunities afforded by modern security technologies suggest alternative approaches are both necessary and possible. What if policies concerning AI chips could be implemented directly on the chips themselves? What if updates to export regulations could be deployed through a simple software update, backed by secure hardware? This report introduces the concept of "on-chip governance mechanisms": secure physical mechanisms built directly into chips or associated hardware that could provide a platform for *adaptive governance*. Its key findings are as follows.

**On-chip governance mechanisms could help safeguard the development and deployment of broadly capable AI and supercomputing systems in a way that is complementary to American technology leadership.**

One especially promising near-term application is export control enforcement, where on-chip mechanisms could prevent or place boundaries around unauthorized actors' use of export-controlled AI chips. Implemented well, this would greatly aid enforcement, and reduce the need for top-down export controls that harm the competitiveness of the U.S. chip industry, instead enabling more surgical end-use/end-user-focused controls if desired.

Later applications include enforcing the terms of future international agreements or other regulations that govern the large-scale training and deployment of AI models. Here, on-chip mechanisms could widen the space of possible agreements and policies by providing a trustworthy verification platform. For example,

on-chip governance mechanisms could allow AI developers to credibly report “training runs” that exceed certain computation thresholds, as called for by a recent White House Executive Order.<sup>2</sup> The existence of these mechanisms could allow for flexible and efficient international governance regimes for AI, allowing policymakers to think beyond the limitations of slow and complex structures such as the International Atomic Energy Agency (IAEA).<sup>3</sup>

**Much of the required functionality for on-chip governance is already widely deployed on various chips, including cutting-edge AI chips.**

Chips sold by leading firms AMD, Apple, Intel, and NVIDIA have many of the features needed to enable the policies described above. These features are used today in a wide variety of applications. On the iPhone, on-chip mechanisms ensure that unauthorized applications can’t be installed. Google uses on-chip mechanisms to remotely verify that chips running in their data centers have not been compromised. Many multiplayer video games now work with a hardware device called a “Trusted Platform Module” to prevent in-game cheating. In the AI space, these features are increasingly used to distribute training across different devices and users while preserving privacy of code and data.<sup>4</sup>

**On-chip governance does not require secret monitoring of users or insecure “back doors” on hardware. On-chip governance is better implemented through privacy-preserving “verification” and “operating licenses” for AI chips used in data centers.**

“Verification” involves the user of a chip making claims that are verifiable by another party about what they are doing with the chip. For example, verifying the quantity of computation or the dataset used in a particular training run.<sup>5</sup> Secure on-chip verification of this kind is made possible by a “Trusted Execution Environment” (TEE). Because of the TEE’s security properties, the verifier can trust that information received from the TEE has not been “spoofed,” without the chip’s user needing to divulge sensitive data.<sup>6</sup>

“Operating licenses” provide an enforcement mechanism. This is useful in cases where, for example, the chip’s owner is found to have acquired the chip in violation of an export control agreement, or if the chip’s user refuses to participate in a legally required verification process. Operating licenses would be best enabled using a dedicated “security module” that links the functioning of the chip to a periodically renewed license key

from the manufacturer (or a regulator), not unlike the product licenses required to unlock proprietary software. Hardware operating licenses of this kind are already used in some commercial contexts.

These mechanisms should primarily be used on the specialized data center AI chips that are targeted by the current AI chip export controls. However, some limited mechanisms on consumer GPUs may be useful if, in the future, these devices are export-controlled.<sup>7</sup>

**Existing technologies need to be hardened before they can be relied upon in adversarial settings such as export control enforcement.**

On-chip governance mechanisms are only useful insofar as they reliably work even when adversaries are actively attempting to circumvent them.<sup>8</sup> Commercial versions of these technologies are not typically designed to defend against a well-resourced attacker with physical access to the hardware. Investments in hardware and software security will be required for on-chip governance mechanisms to function reliably in these kinds of environments.

The specific defenses required to adequately secure on-chip governance mechanisms depend on the context in which they are deployed. This report explores three contexts: minimally, covertly, and openly adversarial.

**A staged approach to the development and rollout of on-chip governance for data center AI chips is possible.**

Intermediate stages of R&D could still be useful in production contexts. In the short term, firmware updates could be deployed to exported AI chips implementing early versions of a hardware operating license linked to the terms of an export license. This would be useful as an additional cautionary measure for already-planned AI chip exports to high-diversion-risk geographies.

A promising and relatively feasible next step would be to make devices “tamper-evident” (attempts to tamper with the chips would leave indelible evidence). This could be a sufficient level of security in cases where occasional physical inspections of the hardware are possible.

For subsequent generations of AI chips, hardware security features could be further hardened, working toward full “tamper-proofing” to make physical inspections less necessary.

To motivate further investigation of on-chip governance, this report sketches an example architecture for data center AI chips that could provide a flexible platform for dynamically implementing different governance mechanisms. The core of this proposal is a hardened security module, included on all high-performance data center AI chips, that can ensure that the chip has valid, up-to-date firmware and software and, where applicable, an up-to-date operating license. If these conditions are not met, it would block the chip from operating.

This valid, up-to-date firmware and software then could help enforce limits on the uses of these chips and offer sophisticated “remote attestation” capabilities (remote authentication to securely verify desired properties of the chip and the software it is running). The security module could ensure that if firmware/software vulnerabilities are found, users would have no choice but to update to patched versions where the vulnerability has been fixed. The security module also could be configured to require an up-to-date, chip-specific operating license.

Current AI chips already have some components of this architecture, but not all. These gaps likely could be closed with moderate development effort as extensions of functionality already in place. The primary technical challenge will be implementing adequate hardware security, particularly for tamper-evidence and tamper-proofing. This report estimates this could be achieved with as little as 18 months of involved technical effort (and up to 4 years) from leading firms.

Because a small number of allied countries encompass the supply chain for the most advanced AI chips, only a small number of countries would need to coordinate to ensure that all cutting-edge AI chips have these mechanisms built in. On-chip mechanisms would need to be supported by a way to track the ownership of data center AI chips, and some form of inspections to ensure these chips are not tampered with, where required.

On-chip governance mechanisms present a promising area for further research for computer engineers, computer scientists, and policy researchers. This report offers the following recommendations to U.S. policy-makers to move toward a world where all leading AI chips are secure and governable.

## Establish government coordination

---

**Recommendation:** The White House should issue an executive order establishing a NIST-led interagency working group, focused on getting on-chip governance mechanisms built into all export-controlled data center AI chips.

**Background:** For on-chip governance to reach commercial scale, long-term collaboration between government and industry will be required. For progress to be made quickly, an executive order could be an appropriate forcing function. The National Institute of Standards and Technology (NIST) would make a suitable lead for this effort. Expertise and staff also should be drawn from the Department of Energy, the Department of Defense, the Department of Homeland Security, the National Science Foundation, and the U.S. intelligence community. The working group should also be informed by a technical panel drawn from industry and academia to help direct technical standards and research.

## Create commercial incentives

---

**Recommendation:** The Department of Commerce (DoC) should incentivize U.S. chip designers to conduct necessary R&D using “advance export market commitments.”<sup>9</sup>

**Background:** Given that on-chip governance mechanisms need to be implemented on commercial chips, much of the necessary R&D will need to happen in an industry setting. To incentivize this work, the DoC should consider making commitments related to future access to export markets to U.S. chip firms, conditional on firms implementing a specific set of security features on controlled products. Such commitments would be an effective way of incentivizing the necessary R&D without spending public money, given the large amount of lost revenue to chip firms caused by export restrictions.<sup>10</sup> Export market commitments could include not extending export controls to new jurisdictions, relaxing the “presumption of denial” licensing policy for chip exports to lower-risk customers in China, or moving toward more surgical end-use or end-user-based controls. The DoC should develop the required feature sets by analyzing specific attacker threat models in different export contexts, in coordination with the U.S. Intelligence Community and Department of Homeland Security.

## Accelerate security R&D

---

**Recommendation:** NIST should coordinate with industry and relevant government funding bodies to scope, fund, and support R&D that can be conducted outside leading chip companies and integrated later.

**Background:** While the large majority of R&D will need to be conducted by the firms building and selling AI chips at scale, some work may be usefully conducted outside of these firms, especially technologies that would benefit from being standardized across the industry. NIST should coordinate with the Semiconductor Research Corporation, relevant Defense Advanced Research Projects Agency (DARPA) program managers, and other relevant government funding bodies to scope and fund useful R&D to be performed by academic and/or commercial partners. For example, work on specialized tamper-proof enclosures (physical housings for chips that prevent the chip from being modified without compromising its operation) for high-end chips could be potentially outsourced to academic and commercial hardware security labs. To support these projects, NIST should create technical standards and reference implementations for on-chip governance mechanisms that are designed for wide adoption by industry.

## Plan for a staged rollout and fund extensive red-teaming

---

**Recommendation:** To ensure that on-chip governance mechanisms are properly designed and safely introduced, the DoC and Department of Homeland Security (DHS) should establish flexible export licensing and red-teaming programs.

**Background:** On-chip mechanisms will require substantial testing before being relied upon in more adversarial environments (e.g., exports of controlled chips to China). To facilitate a staged rollout approach where mechanisms can be depended upon in successively more challenging operating contexts, the DoC should create export licensing arrangements where licenses can be flexibly granted for different geographies based on the security features on the device to be exported. In tandem, the Cybersecurity and Infrastructure Security Agency within DHS should establish red-teaming and bug bounty programs to help find and patch any software and hardware security vulnerabilities. A promising near-term starting point is setting up a public prize for finding vulnerabilities in hardware security features on today's AI chips.

## Coordinate with allies

---

**Recommendation:** The State and Commerce Departments should coordinate with allies on policies and standards for on-chip governance.

**Background:** As with many other forms of technology governance, on-chip governance will be of limited effectiveness without international buy-in. The State and Commerce Departments should include the potential role of on-chip governance mechanisms in diplomatic discussions with countries that occupy important positions in the supply chain for cutting-edge AI chips (especially Taiwan, the Netherlands, South Korea, and Japan), including potential new multilateral control regimes.<sup>11</sup> Looking beyond export control coordination, using on-chip governance mechanisms to facilitate AI governance cooperation (e.g., international agreements on compute usage reporting) would benefit from close coordination with like-minded allies, such as the United Kingdom and the European Union.

## Encourage AI chip firms to move early

---

**Recommendation:** Chip firms should be encouraged to move early to build and harden the security features required for on-chip governance.

**Background:** The United States has signaled interest in on-chip governance in a recent request for comment issued by the Department of Commerce.<sup>12</sup> Chip suppliers that are more able to apply and build on existing technical efforts will have a head start on demonstrating and realizing compliance, with potential benefits in terms of access to markets that are the subject of export controls or other relevant regulation.

Developing and deploying the mechanisms described in this report will take time (months in the most optimistic case, years in the most likely case). If the capabilities and national security risks of AI systems continue to grow at the pace observed in 2022 and 2023, the need for highly effective controls could become acute in several years. This suggests that policymakers concerned about this issue should begin formulating policies and incentivizing the development of appropriate technologies now. Once the relevant security features have been mandated in the most powerful AI chips, they need not be used immediately: The mechanisms outlined in this report would allow for rapid and flexible responses to new developments and threats once installed.

## Introduction

On February 25, 2022, Russian forces attacked the Ukrainian town Melitopol and, after a week of heavy fighting, it eventually was captured. Thanks to its rich soil, the region has been an agricultural hub for over 200 years, a fact that was not lost on the invaders. In the weeks that followed the invasion, locals noticed that grain was disappearing from their silos. But it wasn't just grain being stolen from the occupied town. Over the course of several weeks, combine harvesters (farm equipment used to harvest grain) began to go missing. A review of security footage later would reveal the machinery being loaded onto military trucks, conspicuously marked with white "Z"s.<sup>13</sup> In all, around \$5 million worth of farm equipment was stolen. GPS tracking features on the harvesters painted a startling picture: These stolen assets had embarked on a 700-mile odyssey to Zakhn Yurt, a remote village in Chechnya. But when the invaders tried to use the stolen harvesters, they realized they couldn't turn them on. The harvesters had been disabled by the U.S. manufacturer, John Deere, who has revealed that though they rarely use it, they have the ability to remotely shut down any of their machines.<sup>14</sup>

Tools built into sensitive technologies can enable policies not only for *restriction*, as in the previous story, but also for *verification*. In 1954, the United States

tested a new high-yield thermonuclear weapon design at Bikini Atoll in the Pacific Ocean. It remains the most powerful nuclear weapon ever detonated by the United States, around one thousand times more powerful than those used on Hiroshima and Nagasaki. The test (named "Castle Bravo") caused nuclear fallout to spread over four thousand square miles, resulting in sometimes lethal doses of radiation for people on neighboring islands and nearby fishing vessels, and inciting a strong international reaction, including calls for a comprehensive test ban.<sup>15</sup> In March of 1960, the United Kingdom, the United States, and the Soviet Union were negotiating the terms of such an agreement. These discussions led to the 1963 Partial Nuclear Test Ban Treaty, which 123 countries have since ratified. It was a partial ban rather than a comprehensive one in part due to a key problem for verifying compliance: it was, at the time, impossible to reliably remotely detect underground tests. Consequently, the ban was limited to tests conducted in the atmosphere, underwater, and in outer space. Two years later, significant progress had already been made towards solving the problem of reliably detecting underground tests, using the idea of a network of seismometers (devices used to measure seismic activity) combined with a new efficient algorithm for differentiating between nuclear tests and other seismic activity. But a treaty had already been signed, and it wasn't until many years later, in 1990, that

the United States and Soviet Union ratified a treaty involving underground tests: the "Threshold Test Ban Treaty", which prohibited all nuclear tests exceeding 150 kilotons. This treaty was enabled by mutual agreement between the two countries on a specific technical protocol for the verification of underground tests based on the approach described above. Of course, verification is only one part of the rationale behind arms control treaties, but this story shows the role that verification technologies can play in enabling international agreements and governance structures that may not otherwise be able to exist.

Both these stories highlight some of the challenges with technology-based solutions to policy problems. The first is achieving sufficient reliability. Although the combine harvesters were remotely



John Deere is one of the world's largest exporters of farm equipment and spends around \$2 billion annually on research and development. This has led to a complex hardware and software stack for their equipment, allowing remote control of newer vehicles. Here, Ukrainian farmer Mykhailo Palahniuk points toward a John Deere harvester under repair, on his 6-hectare farm where he grows crops of wheat, barley, and soy. (Scott Peterson/Getty Images)

disabled, it's likely that Russian troops eventually were able to bypass the protection, provided it was worth the time and money to do so. The second is timing. Though it turned out to be possible to verify underground nuclear tests, this development came too late to be truly useful for nuclear nonproliferation.

This report considers the applicability of these kinds of technological solutions to AI policy. What if policies concerning AI chips, a crucial input for dual-use AI systems, could be implemented directly on the chips themselves? What if updates to export regulations could be deployed through a simple software update? Such “on-chip governance mechanisms” could help flexibly address many of the national security issues posed by future AI systems in a way that does not presuppose any specific risks. However, this approach raises difficult questions about how dangerous technologies should be governed. This report lays out the policy objectives that could be achieved with on-chip governance mechanisms. It then examines the technical and social challenges to their implementation. Finally, the report provides a set of recommendations for U.S. policymakers to move toward a world where all leading AI chips are secure and governable.

### The National Security Risks Posed by Artificial Intelligence

The 2021 Final Report of the National Security Commission on AI characterized AI as “the quint-essential ‘dual-use’ technology.”<sup>17</sup> In late 2023, this characterization appears increasingly apt. China has been rolling out a nationwide AI-based system of mass surveillance, driven by advances in facial and voice recognition technology.<sup>18</sup> These tools also have been key to Beijing’s mass oppression and incarceration of Uyghurs in Xinjiang.<sup>19</sup> AI also plays a key role in China’s military ambitions, with the goal of progressively integrating AI into its joint forces over the coming years.<sup>20</sup>

Looking closer to home, earlier this year, OpenAI, a top U.S. AI lab, released “GPT-4,” marking the birth of a new generation of broadly capable AI models that are increasingly unlocking groundbreaking applications in both civilian and defense contexts.<sup>21</sup> The demonstrated



*The U.S. nuclear weapon test Castle Bravo had a yield 2.5 times greater than predicted due to unforeseen reactions involving lithium-7. (United States Department of Energy)*

capabilities of today’s most powerful systems suggest that their successors could be highly proficient within a range of weaponizable domains, and could pose serious risks if they fall into the hands of adversaries. GPT-4 has nascent capabilities useful for designing, planning, and executing complex scientific experiments, including synthesizing chemical weapons.<sup>22</sup> Defense contractors have started offering decision-making systems powered by the current generation of broadly capable models, models whose foreign and open-source counterparts are increasingly becoming available to the United States’ adversaries.<sup>23</sup> A broader set of worrying capabilities are also being discovered. In early 2023, Anthropic, another leading U.S. lab, contracted top biosecurity experts to red-team and evaluate its model’s ability to help with the design and acquisition of biological weapons. They found that “a straightforward extrapolation of today’s systems to those we expect to see in 2-3 years suggests a substantial risk that AI systems . . . will greatly widen the range of actors with the technical capability to conduct a large-scale biological attack.”<sup>24</sup> Other domains where evidence

### APPLICATIONS BEYOND AI

This report uses the term “AI chips,” and primarily highlights the benefits of on-chip mechanisms for addressing AI-related national security concerns (specifically compute-intensive broadly capable systems). But the advanced chips referenced also play an important role in non-AI applications, such as design and testing for aerospace systems and nuclear weapons.<sup>16</sup> The measures discussed in this report are highly relevant for these cases, and in general, for wherever advanced chips are used in national security-relevant applications.

of national security risks are emerging include cyber offensive operations, large-scale deception or persuasion operations, and “agentized” AI systems evading human control.<sup>25</sup>

Broadly capable models at the frontier of R&D also have properties that will pose a thorny challenge for governance: These models develop new capabilities in an unpredictable way, are hard to make reliably safe, and are likely to proliferate rapidly to illicit actors.<sup>26</sup> Systems based on these models could have destabilizing effects on international relations and lower the barrier for non-state actors to cause harm.<sup>27</sup>

### **Compute Governance: Opportunities and Challenges**

In October 2022, prior to many of the AI advances previously described, the U.S. Department of Commerce imposed aggressive export controls to limit China’s access to high-end AI chips. The new regulations cited concerns that China could use these chips to produce advanced military systems, including weapons of mass destruction, and to enhance the speed and accuracy of its military operations.<sup>28</sup> This move is an example of “compute governance”: placing guardrails on how and by whom the resources necessary to produce AI computation (i.e., specialized computer chips) are used. Such measures can be effective in the context of AI because of the close relationship between the amount of compute/chips used to train a particular AI model and the capabilities the model possesses.<sup>29</sup> Compared to other inputs to AI, such as data and algorithms, chips have a unique set of governance-relevant properties. AI chips are physical goods that are more quantifiable, more difficult to copy, and have a highly concentrated supply chain; all attributes that make it easier to define and enforce policies that control access and govern their usage.<sup>30</sup>

But in some crucial ways, compute governance measures that resemble unilaterally imposed rules on who can export chips to whom are a blunt and ineffective tool. First, enforcement of these rules is hard. China’s extensive civil-military fusion and use of shell entities to evade export controls have historically compromised the effectiveness of export controls, particularly given the limited resources and outdated technology available to the Bureau of Industry and Security (the office tasked with export control enforcement for AI chips).<sup>31</sup> AI chip smuggling is already happening on a small scale today and is likely to be an increasing concern.<sup>32</sup> Second, to have a chance of being effective, such rules need to be far-reaching, which has consequences for the competitiveness of U.S. firms. Under the current

set of controls, exports of leading AI chips are prohibited to any customer in China. The ban applies not just to U.S.-made chips, but also to any chip produced using U.S.-origin technology, software, or equipment. Understandably, this has prompted calls for restraint from the U.S. semiconductor industry, which fears not just loss of market access to China, but also the “de-Americanization” of supply chains more broadly.<sup>33</sup> Lastly, these rules have workarounds. Cloud computing makes it possible for cloud service providers in other countries to provide export-controlled AI chips remotely to any customer.<sup>34</sup> Countries and other actors can also stockpile chips today to guard against the possibility of new or expanded export controls in the future, pushing some of the effectiveness of export controls to the point at which stockpiled chips are less relevant for training leading AI models.

### **On-Chip Governance Mechanisms: A Quick Introduction**

The challenges inherent to the current approach for governing compute obscure the potential harmony that exists between the interests of U.S. policymakers and U.S. chipmakers. Both seek to promote U.S. technological competitiveness, and neither wants to see dangerous and destabilizing technologies in the hands of unlawful actors or rogue states. The problems described in the previous section arise in large part due to the limited policy solution space allowed by the technology deployed on today’s generation of AI chips, rather than the inherent differences in goals between these two groups. For example, blanket export restrictions on AI chips going to China were seen as necessary in part because there does not exist a widely deployed technical solution for preventing an AI chip from being used by an unauthorized actor once it has been shipped overseas. If such technology was deployed and made sufficiently reliable, the need for sweeping, top-down export restrictions would be reduced.

This report introduces the concept of “on-chip governance mechanisms”: physical mechanisms implemented on AI chips and related hardware to allow for control of how and by whom these devices can be used. These actions will be appropriate only in certain contexts, such as export control enforcement for advanced AI chips used in data centers.<sup>35</sup> Implementing on-chip governance on such chips could be valuable by virtue of the fact that the most broadly capable dual-use AI systems also require the most specialized chips to develop within a reasonable time frame and budget.<sup>36</sup> These mechanisms provide a uniquely flexible governance tool. If a

basic “security module” is introduced as a standard for powerful new AI chips, new policies could be quickly and securely deployed as software and firmware updates. This possibility is discussed in Section 2.

On-chip mechanisms for commercial uses already are common on consumer devices. For example, the iPhone has hardware restrictions that enable Apple to exercise editorial control over which specific apps can be installed. Google uses a technique called “remote attestation” to ensure the security and integrity of devices in their data centers.<sup>37</sup> Some video game companies use a similar method to interface with a dedicated secure processor on users’ computers to ensure they are not using cheating

software in multiplayer video games.<sup>38</sup> IBM and Intel use an approach known as “hardware licensing” to remotely restrict/unlock the performance of data center chips based on a subscription model. Using similar technology to these examples, on-chip mechanisms could be implemented on data center AI chips to directly govern the training and deployment of broadly capable AI systems that pose meaningful national security risks.<sup>39</sup> Section 2 provides an overview of what this might concretely look like, and Section 4 dives further into the required technical underpinnings.

In the domestic context, on-chip governance mechanisms could enable the development of AI systems

to be regulated while helping protect the intellectual property of developers.<sup>40</sup> For example, a recent White House executive order calls for AI developers and cloud service providers to report on compute usage and training data above certain thresholds.<sup>41</sup> On-chip mechanisms could allow these firms to quickly and securely report this information, without needing to directly reveal sensitive code or data. While these use cases are interesting, the more promising governance application for on-chip mechanisms is international governance, where regulations are otherwise difficult to monitor and

enforce. In the near term, this means export control enforcement, but in the longer term, this could include international

## **Realizing the potential of on-chip governance will require substantial investments in better hardware security and a strong partnership between the U.S. government and leading AI chipmakers.**

agreements.<sup>42</sup> This report focuses on export controls as a promising early application in Section 3 and associated challenges in Section 5.

This may sound like an ambitious proposal, and it is. However, new governance tools almost certainly will be needed to meet the national security challenges presented by current and future rapid advances in AI. As AI systems grow more powerful, the need to effectively govern them will grow more urgent. A flexible governance framework built on a platform of on-chip mechanisms would allow regulations to adapt to changes in the technology that cannot yet be foreseen. Because developing these mechanisms will take time, work to

### **DEFINITIONS**

This report defines on-chip governance mechanisms as technical mechanisms that rely on hardware-level security features to:

- Enable a *controller* to restrict what can be done with a hardware device; and/or
- Enable a *verifier* to verify claims about the state or use of the hardware, based on having a high level of trust about the integrity of the security mechanisms.

The proximate controller almost always would be the hardware vendor, but the de facto controller could be, for example, a regulator who mandates that particularly powerful hardware should not be made available to unlawful actors.<sup>44</sup> In a future, more comprehensive AI governance regime, a regulator could be both a verifier and a controller: For example, they could require AI developers to verifiably report that they are going about their development safely, and impose restrictions on developers who cannot prove this.<sup>45</sup>

This report also uses the terms “compute user” and “compute operator.” The user is the entity that uses chips in an operational capacity (e.g., a company that trains AI models). The operator is the entity that owns, physically controls, and manages the computing hardware (e.g., a cloud service provider). In some cases, the same entity will be both the compute user and the compute operator. In other cases, these entities will be distinct.<sup>46</sup> For specific definitions of other AI compute-related terms used in this report, see Appendix A.

design and implement them needs to commence well before they are needed. Due to the concentration of the semiconductor supply chain, a coalition of only a few partner countries, or perhaps even the United States alone, could be enough to ensure that on-chip governance mechanisms were introduced on almost all leading AI chips.<sup>43</sup>

The goal of this report is to introduce the concept of on-chip mechanisms as a tool for governance and review the underlying hardware security features that could make them possible. It does not attempt to analyze the implementation of these mechanisms in any specific use cases with rigorous technical detail, nor try to exhaustively map possible use cases. Detailed analysis of the broader impacts of specific mechanisms used for specific purposes also is left for future work.

## **Realizing the potential of on-chip governance will require substantial investments in better hardware security and a strong partnership between the U.S. government and leading AI chipmakers.**

Realizing the potential of on-chip governance will require substantial investments in better hardware security and a strong partnership between the U.S. government and leading AI chipmakers. It also will require a thoughtful development approach that acknowledges the privacy and free speech implications of making AI chips more controllable by regulators. Effective on-chip governance is best implemented through enabling privacy-preserving “verifiable claims” and operating licenses for chips that are used almost exclusively in large data centers, and will *not* require secret monitoring of users or insecure “back doors.” In fact, related mechanisms are widely deployed already on various chips, including cutting-edge AI chips, without compromising users’ privacy or security.

While the future evolution of risks from AI and advanced computing cannot be predicted with certainty, given the potential stakes, and the lead times involved, it is time to lay the groundwork to expand the range of AI governance options available to the United States over the coming years.

## **What Would Effective On-Chip Governance Look Like?**

This section briefly lays out a sketch of a concrete vision for the set of on-chip mechanisms and associated measures that would allow for flexible compute governance. The core of this proposal is a hardened “security module,” included on all high-performance data center AI chips, that can ensure that the chip has valid, up-to-date firmware and software and, where applicable, an up-to-date operating license. If these conditions were not met, the security module would prevent the chip from operating.

This valid, up-to-date firmware and software then could help enforce limits on the uses of these chips, and offer sophisticated “remote attestation” capabilities, or, in less technical terms, the ability for the chip to send trusted information about the chip and its usage to a third-party verifier. The security module also would ensure that if vulnerabilities are found in firmware and software, users would have no choice but to update to patched versions where the vulnerability has been fixed. Chip-specific operating licenses would allow export-controlled chips to be configured such that they could be remotely disabled by the manufacturer by ceasing to issue licenses for that chip. This would allow export controls to be enforced remotely if the terms of an export license had been violated. Chips also would have support for “trusted execution environments” that could, together with remote attestation capabilities, allow the chips to be used to make a wide range of “verifiable claims,” such as the amount of compute used to train an AI model or other properties of the training process.

Implementing these features on AI chips provides a platform for *adaptive governance*. These features would allow for a wide range of policies (for example, a training compute reporting requirement above a certain threshold, as called for by the recent White House executive order) to be implemented and updated directly on the chip by simply deploying a firmware or software update.<sup>47</sup> Many of the required security features are already common on CPUs and are being increasingly introduced on GPUs, such as NVIDIA’s new H100.<sup>48</sup> These likely could be implemented at an acceptable cost as an extension of existing standards for secure boot and remote attestation features.

These technical features ideally would be supported by robust supply chain tracking and “Know Your Customer” policies for AI chip exports/sales, which would allow the controller to know which chips are being used by which actors. This system of supply chain tracking also could

include periodic monitoring and inspections to ensure that any novel attempts to physically tamper with chips can be caught.

With this overall sketch as a framework, the next section describes in more detail the specific policies that these technical features could unlock.

## Policies that On-Chip Governance Mechanisms Could Enable

At a high level, on-chip governance mechanisms could allow a regulator to take the following actions:

1. **Restriction:** Restricting access to, or “throttling” (reducing) the performance of a chip. Such measures also could include preventing the chip from being used as part of a large cluster/supercomputer.
2. **Verification:** Requiring the chip user to securely verify how they are using the chip (e.g., which specific code or data is being used in an AI training run).

Details on the technical underpinnings of these capabilities are included in Section 5, and Section 6 discusses their viability in adversarial contexts.

These actions will be appropriate only in certain contexts. Restriction mechanisms are appropriate in the adversarial context of export control enforcement, on the specialized data center AI chips that are targeted by current AI chip export controls. In the future, as chips grow more powerful, it may become necessary to place some export restrictions on consumer-grade GPUs.<sup>51</sup> These chips could then potentially be equipped with some limited mechanisms to deter smuggling and misuse.

In practice, restriction and verification could be used to enable the following policy measures:

- **Operating licenses:** Using hardware-enforced licenses to deny access to unauthorized users, (e.g., for export control enforcement).
- **Location verification:** Verifying the location of chips, (e.g., to assist with export control enforcement).
- **Usage verification:** Verifying how chips are being used, (e.g., to enforce an international agreement on tracking and reporting compute usage).<sup>52</sup>
- **Usage limitations:** Limiting certain chip use cases, (e.g., to restrict exported chips from being used to build large AI clusters capable of training frontier models).<sup>53</sup>

The rest of this section will discuss each of these in more detail.

### Operating Licenses to Prevent Unauthorized Use

On-chip mechanisms could be used to implement a chip-specific operating license that requires periodic renewal, similar to a software subscription model. Operating licenses could control whether the chip works at all, limit specific features, or specify more complex restrictions. Importantly, on-chip mechanisms could implement a time-based license, where a chip disables itself if it does not receive a renewed license. This approach prevents reliance on the chip needing to receive an active shutdown command, which likely could be blocked by an uncooperative compute operator.

## VERIFICATION VS. MONITORING

This report uses the term “verification” to distinguish it from the idea of activity monitoring. “Monitoring” implies that some third party is able to track how a chip is being used (e.g., specific code or data loaded on the chip) through some process of unilateral surveillance. Such monitoring is likely neither technically feasible nor desirable from a user privacy and chip security standpoint. Building “back doors” into AI hardware is technically possible but would not result in chips that consumers will want to buy, and would introduce serious security vulnerabilities.<sup>49</sup>

“Verification” refers to a process where the user of a chip instead can remotely attest to a third-party verifier what they are doing with a processor (e.g., how much training compute is being used, or whether a particular dataset was used), using a “Trusted Execution Environment” (TEE). Because of the TEE’s security properties, the verifier can trust that information received from the TEE has not been spoofed, so long as they have confidence that hardware security features on the chip have not been compromised. Instead of unilateral surveillance, this should be thought of as a collaboration between a verifier and the chip owner. This collaboration also could be made fully privacy-preserving (i.e., not revealing sensitive code or data) using techniques from multi-party and confidential computing.<sup>50</sup> If a chip owner refuses to engage in such a collaboration, restriction mechanisms could allow the verifier (e.g., a regulator or device manufacturer with particular terms of use) to prevent them from continuing to use the chip.

Hardware-based operating licenses already are used in commercial contexts; two U.S. companies, Intel and IBM, run hardware licensing programs under the names Intel On Demand and Capacity on Demand respectively.<sup>54</sup> In these cases, operating licenses are used to restrict or unlock existing features on chips, depending on whether a customer has paid for them.

This capability would be particularly useful for export control enforcement—for example, if a chip were sold to an entity that subsequently was found to have previously unknown ties to the People’s Liberation Army.<sup>55</sup> In practice, this might take the form of a Bureau of Industry and Security statement that export licenses will be granted for controlled chips if the chips have a security module that could be used to disable the chips remotely if there is ever a reason to believe the chips have been utilized by end users and/or for end uses that constitute a breach of the export license. This could include:

1. Cases where there is a reason to believe that chips have been, or are at risk of being, re-exported or transferred in violation of their original export license.
2. Cases where there is reason to believe that remote access to the chips has been given to sanctioned entities, such as those connected to the Chinese military (if controls on AI chips offered as cloud services are implemented).
3. Cases where the owner of the chips is not collaborating with authorities to prove that neither of the two violations mentioned above is occurring.

While an operating license mechanism could require some communication between the chip and the manufacturer, the core functionality would not require an open internet connection. The license could be conveyed to and from the chips by whatever means are most appropriate, whether that be an internet connection, or carefully controlled physical media going in and out of an air-gapped data center.

More speculatively, it may be possible to use operating licenses to make consumer GPUs less useful for AI applications, by using a license to unlock some of the most AI-relevant features and capabilities of the GPU. Such mechanisms are not currently needed, but may become useful in the future.

### Location Verification

Combining trusted location verification with operating licenses could allow for rapid and effective export

control enforcement. How would this work? Due to the hard limit of the speed of light and the lower bound on latency from existing communications infrastructure, how quickly a device responds can be used to establish an upper bound on the distance between the device and the source of the query. With secure on-chip mechanisms and multiple trusted “landmark” servers, it becomes possible to determine the approximate location of a chip by comparing these upper bounds, as depicted in the diagram below.

Due to the substantial difference between the speed of light and the latency of ordinary internet infrastructure, the chips would need to be within hundreds of kilometers of the landmarks, and less than 100 km in areas that are near the borders of areas where chips are not allowed to operate. This likely would require hundreds of trusted landmarks globally, but these servers would be quite cheap to set up. Queries would take the form of cryptographic challenges issued against the chip’s private key, to ensure that the responder is indeed the chip in question.

This kind of location verification mechanism would be particularly valuable for deterring chip smuggling. Of course, the response from a device always could be delayed to reduce the precision of the location estimate. Collaboration on the part of the compute operator could be incentivized by enacting a policy of revoking operating licenses if the measurement is so imprecise that it cannot



*This diagram shows how a trusted server in Paris, France, could verify that a chip is within the blue circle, and thus not in any country to which chip exports are restricted, by verifying that the chip can respond in less than 9 ms (using the upper bound of the speed of light). The smaller red circle shows an approximate range from which chips could attain a 9 ms latency to the server, using ordinary internet infrastructure.<sup>56</sup> This shows how landmark servers placed every few hundred kilometers could be used to establish sufficient coverage to verify that chips have not been re-exported illegally.<sup>57</sup>*

verify that the chip is not in a country in which it should not be. Alternatively, the chip itself could query a set of trusted servers, and the operating license could specify that the chip should lock down if it cannot establish that it is in an allowed region. This approach also could allow a chip to establish its own location without the landmark servers being able to determine the chip's location, which could be desirable in some cases, to protect user privacy. This kind of "region-lock" mechanism could potentially also be useful on consumer GPUs in the future, if the smuggling of such chips becomes a serious concern.

### Usage Verification

Continued progress in AI may create and exacerbate scenarios that resemble a "security dilemma."<sup>58</sup> For example, if one country were unsure about a rival's intentions or activities related to developing AI-powered military capabilities, it may be rational for that country to develop or accelerate the development of its own capabilities. Uncertainty about the specific capabilities of rivals, how AI might change the shape of warfare, and exactly how powerful future AI systems might be could all exacerbate this dynamic. This could lead to incentives to prioritize dangerous capabilities research at the expense of safety research, increasing the chance of accidents that could cause harm to all actors.<sup>59</sup> Recent trends in military adoption of AI technology suggest these dynamics are at risk of emerging between Washington and Beijing.<sup>60</sup>

As with most security dilemmas, a promising move is to reduce mutual uncertainty about how and whether potentially dangerous systems are being developed by any actor. Just as monitoring and verification technologies have been used to support international agreements and mutual trust in the nuclear domain, on-chip mechanisms could support similar moves in the AI domain.<sup>61</sup>

Two key points of difference between these domains are that the number of different actors

involved in developing new technologies is likely to be greater in the AI domain, and those actors are much more likely to be commercial actors.

On-chip mechanisms could allow compute users (commercial or otherwise) to make verifiable claims (information that is trustworthy through hardware-level integrity guarantees) about the state of a chip and how it is being used. These features could be extended to the level of an entire cluster and enable compute users to verify key information relevant to AI capabilities and risks, such as the amount of training compute used to

train an AI system or other properties of the training process. For example, one recent proposal describes how "hashed" (i.e., privacy-preserving) parts of an AI system could be stored, and later used to prove how much compute was used to train it.<sup>62</sup>

Many of the security features necessary for verifiable claims are already available on high-end server CPUs, as well as NVIDIA's flagship H100 GPU. In recent years, this has been marketed as "confidential computing" and promoted by the Confidential Computing Consortium, of which NVIDIA is a member.<sup>63</sup>

### Usage Limitations

On-chip mechanisms could be used to limit the possible uses of chips in various ways. The most relevant to this report are limiting AI chip usage in large clusters/supercomputers, limiting sensitive data access to support privacy and information security, and limiting chips to only running approved code or models. Each of these applications is discussed below.

#### LIMITING AI CHIP USAGE IN LARGE CLUSTERS/SUPERCOMPUTERS

In the October 2023 revisions to AI chip export controls, BIS requested public proposals for "technical solutions that limit [AI chips] from being used in conjunction with large numbers of other such items in ways that enable training large dual-use AI foundation models with capabilities of concern."<sup>64</sup> If this kind of usage were prevented, chips could be safely exported for end uses that only require a smaller number of chips.

As part of the request, BIS mentions an example mechanism, where the various chips that make up a single system, such as a server or a "pod" of servers, are limited to only operating with the original set of chips, and the

whole system is limited to only communicate at less than 1 GB/s with the outside world. This kind of restriction could

**Continued progress in AI may create and exacerbate scenarios that resemble a 'security dilemma.'**

be based on "roots of trust" in each of the chips in the system, that allow all of the chips to attest to each other's identity. Chips would then refuse to work with any chip they do not recognize, which would prevent the end user from introducing additional network connections that would allow the system to be integrated as part of a larger cluster.

A mechanism like this would require a very high degree of interoperability between all of the chips in the system, including, for example, the CPU and the network interface controller. Existing chips could not do this, but

fortunately, the data center industry already is working to develop standards and protocols to allow heterogeneous devices found in data centers to attest their identity and integrity to each other using such mechanisms.<sup>65</sup> However, this level of interoperability could be at least 2 years away, based on an interview with an industry expert.

Sophisticated on-chip attestation mechanisms should be complemented by lower-tech physical protections to make it more difficult to modify the system without damaging it, and leaving evidence of modification. This could involve techniques such as “potting”: covering the circuit board in difficult-to-remove material. See Appendix B for a dedicated discussion of anti-tampering technologies.

In the future, this kind of attestation mechanism could be extended to implement more flexible restrictions on the use and configuration of computing clusters. In addition to identifying each other, AI chips could share relevant information to detect if they are part of a very large computing workload (e.g., large-scale AI training). For example, each AI chip in a server could track how much data is moving in and out of itself. This information then could be used to estimate the total amount of data being moved to and from the whole server, and therefore detect whether the chips are being used within a large, tightly connected cluster of multiple servers. However, this kind of system could potentially be broken by compromising a small number of the least secure devices involved, making it relatively fragile.

#### LIMITING SENSITIVE DATA ACCESS

On-chip mechanisms could support information security and privacy practices. For example, when an AI system is deployed, on-chip mechanisms could be used to ensure a user’s data is processed without either the AI developer or the user being able to access the other party’s intellectual property (data or model weights). Beyond their commercial utility, such features may become increasingly important as AI systems develop further capabilities in domains with high potential for misuse, such as biology.<sup>66</sup>

#### LIMITING AI CHIPS TO ONLY RUNNING APPROVED CODE OR MODELS

On-chip mechanisms could be used to ensure that only approved code and/or AI models can be run on the processor. This could allow a subset of chips intended for specific uses (e.g., those for use in self-driving cars), to be configured to only run specific, trusted models. This could allow some kinds of misuse to be prevented without much active oversight of the chips.

## Technical Underpinnings

This section begins by explaining the basic operating principles of the core hardware security features—secure boot and remote attestation—that most restriction and verification mechanisms are based on. It then explains how these features could be applied in security modules and trusted execution environments to enable on-chip governance. Each of these key components is shown in Diagram A on the following page.

### CRYPTOGRAPHIC SIGNATURES

On-chip mechanisms rely heavily on cryptographic signatures (also known as digital signatures), a way of verifying the authenticity of a file or message using public key cryptography.

Public key cryptography is a system that uses two different mathematically related codes, called keys, to encrypt and decrypt data. One key is public, while the other key is private and must be kept protected.

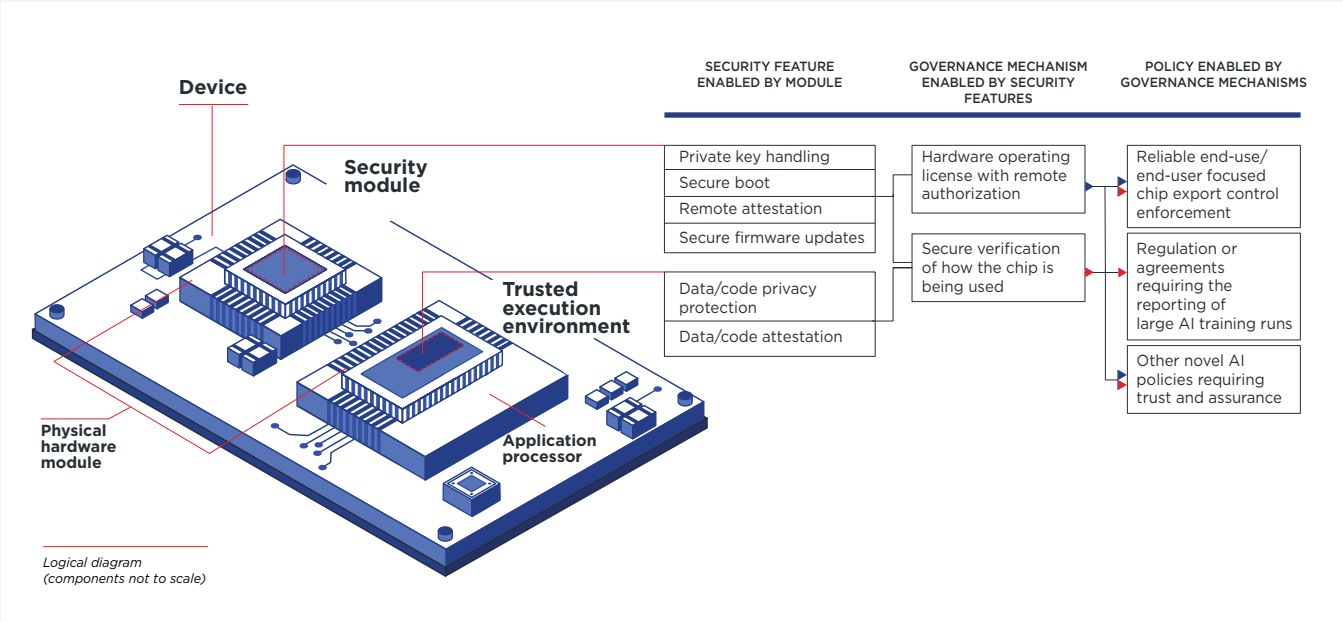
A cryptographic signature is a sequence of bits that can be used to verify the authenticity of a file or message. It is created using the file and a private code (referred to as a “private key”). Recipients of the file then can use a corresponding “public key” to verify that the signature is valid and that the file comes from the owner of the private key and has not been modified in transit.

### Secure Boot

“Secure boot” is a hardware feature that aims to prevent unauthorized firmware, operating systems, or other software from running on a device.<sup>67</sup> When a chip is turned on (booted), the part of the chip that is responsible for loading the initial firmware code onto the chip checks whether the code has been cryptographically signed by the chip’s manufacturer, and refuses to boot if not. This ensures that the chip will run only manufacturer-approved firmware. This typically works as follows:

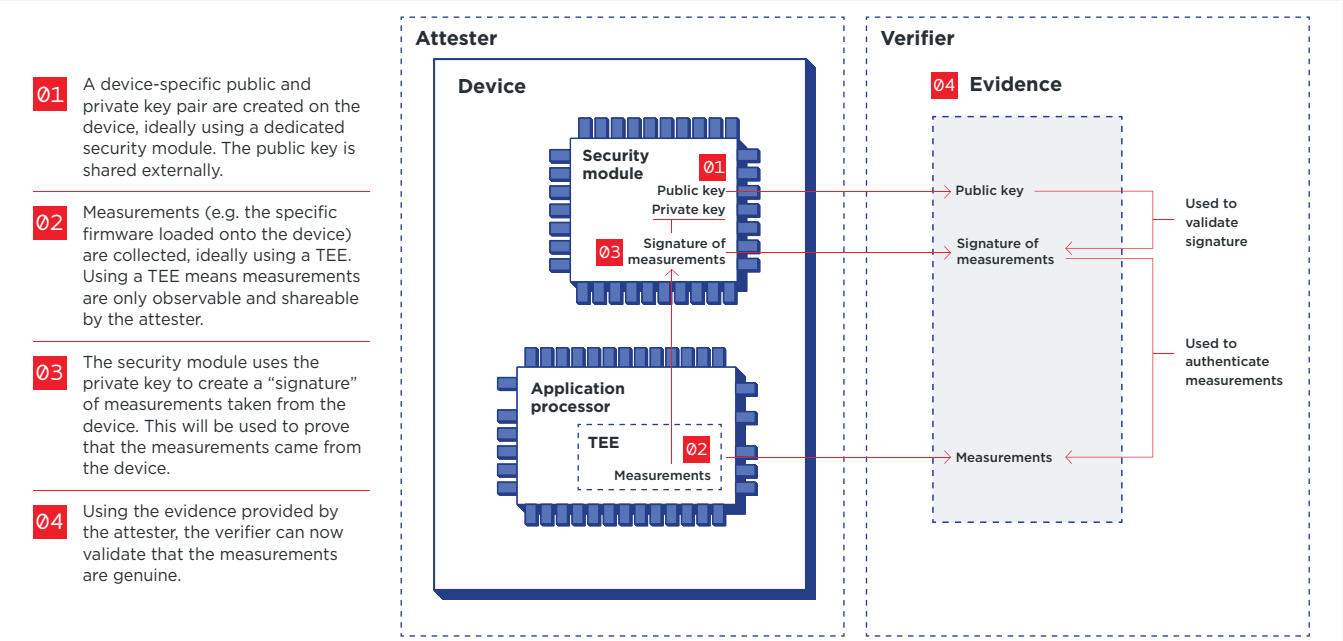
1. The manufacturer generates a pair of keys: a public key and a private key
2. The manufacturer stores the public key in read-only memory on the device
3. The manufacturer signs the firmware with the private key, creating a signature for it
4. The manufacturer sends the firmware and the signature to the device. The device uses the public key to verify that the signature matches the firmware.

DIAGRAM A: AN EXAMPLE ARCHITECTURE FOR ON-CHIP GOVERNANCE



A logical diagram of the governance mechanisms proposed in this report, and the physical hardware modules and security features that could enable them.

DIAGRAM B: HOW DOES REMOTE ATTESTATION WORK?



A simple diagram of an example remote attestation process, showing the flow of keys, information, and signatures. The "measurements" here could be, for example, information about what code has been loaded onto the chip.

Secure boot does not require the device to have any secret information, such as a private key. It only needs to protect the public key from being overwritten. Remote attestation, on the other hand, requires that chips be able to sign outputs so that they can be verified to have come from that chip. This means that the chip itself needs to hold its own private key and prevent anyone from reading it; otherwise, whoever reads the key can forge attestations. Remote attestation is discussed in more detail next.

### Remote Attestation

The same functionality that is used to check the integrity of the configuration and firmware of a chip as part of secure boot can be extended to allow the hardware to securely remotely attest to (i.e., make claims about - the state of the system.<sup>68</sup> This is known as “remote attestation.” In a remote attestation procedure, the chip generates a signature for the currently loaded firmware (and other measurements about the chip’s state) using its own private key, and sends that signature to a verifier (e.g., the manufacturer). The verifier can then use the signature to ensure that the chip is running approved firmware or has a valid “operating license”. This overall process is depicted in Diagram B on the previous page. Remote attestation capabilities make it possible for a remote party to have some degree of control over how a chip is being used, particularly in combination with “trusted execution environments” (page 14). Such features could be especially useful in the export control context, where an exporter could retain the ability to remotely restrict access to a chip if an export control violation has been detected via remote attestation.

### Security Modules

To implement techniques such as secure boot, many chips today have dedicated security modules, including a dedicated processor, that are responsible for handling private keys and performing other security-related functions. For the purposes of on-chip governance mechanisms such as operating licenses, a security module would need to perform responsibilities such as:

- Secure boot, including measuring, enforcing, and attesting to firmware integrity
- Enabling secure remote firmware updates
- Handling private keys and cryptographic operations to support verifiable claims

- General oversight of the behavior of the chip
- Attesting to device identity.

To implement an operating license (see Section 3), a security module would need to have the ability to limit or disable a chip’s operations if the chip does not receive a renewed license within a particular time window. The format of the license could be a short piece of text, cryptographically signed by the compute vendor. The text should include the identifier of the chip in question and information about the ways in which it is authorized to operate, and for how long. The firmware running on the security module would interpret and enforce this license. To support this functionality, the security module would need to have access to an immutable ID corresponding to the chip it is responsible for.

With a timed license expiry period (e.g., weekly or monthly), chip vendors could disable chips without any active intervention being required. The authors expect that a timed license is the only way to implement a robust mechanism for remotely disabling chips: if the mechanism relied on a shutdown command being actively delivered to the chip, the command almost always could be blocked from reaching the chip by the compute operator.

Another technical requirement for properly implementing a hardware operating license is a secure timer. Accurate, hack-proof, and tamper-proof tracking of time generally is considered very difficult.<sup>69</sup> The main reason for this is that, currently, it is within the capabilities of many actors to compromise timers by manipulating the power supply to the chip, and thus manipulating the execution speed of instructions.<sup>70</sup> However, for the purposes of controlling access to or usage of AI compute, the primary concern is with the amount of computation done since authorization was received, rather than the exact amount of time.<sup>71</sup> This can be tracked much more robustly by simply counting clock cycles.<sup>72</sup>

It also could be possible to achieve a usable approximation of time if the relevant parts of the chip were continuously powered. This could be achieved with an added battery that could continue to power the relevant part of the chip even when the rest of the system is powered off.<sup>73</sup> It also might be possible to require the surrounding system to provide continuous low levels of power to the chip by designing the timer to “max out” if power is lost, thus requiring re-authorization in the event of a loss of power.

## Trusted Execution Environments

Trusted Execution Environments (TEEs) are isolated environments created within a processor that protect the code and data running inside them from being accessed or modified by other parts of the system. The key difference between security modules and TEEs is that TEEs create a protected environment on the main processor cores, whereas a security module is a separate lower-performance processor specialized for security-related tasks. While security modules can be sufficient for protecting highly sensitive information, TEEs provide an additional layer of protection around the primary computational work performed by the chip.

TEEs are typically used to protect data inside the environment from spying or interference by other parts of the system, such as malware, other users, or the platform software provided by a cloud provider. In the case of on-chip governance mechanisms, TEEs can be used to enable a chip to remotely attest to the state of the TEE and the code running inside the TEE, with these claims being verifiable by third parties.

This can enable certain types of privacy-preserving collaboration using a technique known as multi-party computing. For example, one party could set up a TEE on a chip and attest to another party about the specific code that is loaded in the TEE. The other party could then send encrypted data to the TEE, which is processed by the code, and the results shared, without the original party ever having access to the unencrypted data.<sup>74</sup> This approach conceivably could be used by a third-party evaluator to run tests on an AI model without ever having direct access to the unencrypted weights.

TEEs also might be useful for implementing privacy-preserving logging of information during training. This would allow for retrospective inspections of the training process. A recent paper proposes a protocol for verifying adherence to rules related to AI training—for

example, the amount of compute, data, or training process used.<sup>75</sup> In this proposal, weights on a chip would be hashed and signed at random times during training, and these hashes would be logged.<sup>76</sup> The logged hashes could be used later to prove which chips were used to train a given model, and to verify the provenance of a model through the provision and replication<sup>77</sup> of training transcripts from the organization that did the training.<sup>78</sup>

## When Should a Security Module vs. a Trusted Execution Environment Be Used?

Security modules use separate dedicated processors for handling security-critical operations like cryptography and enforcing policies. TEEs are isolated environments created within the main processor(s) of a chip to protect code and data from being accessed by other software on the system.

A security module could be much simpler than its associated AI chip, and thus much more secure. If the interface between the security module and the user-accessible parts of the system can be kept very simple, it is much more feasible to ensure it does not have major vulnerabilities that could be exploited to gain access to the security module from the main processor.

Trusted execution environments, on the other hand, run on the main processor(s) themselves. This complexity has often led to TEEs being vulnerable to side-channel attacks that exploit shared resources like caches.<sup>79</sup> A separate security module reduces this risk given that user code is not allowed to run on it. However, TEEs are necessary to enable remote

attestation of code (and data) running on chips. As such, this report suggests that security modules should be used in on-chip governance mechanisms where possible, such as for requiring a valid operating license, and TEEs should be used otherwise only where necessary, such as for enabling verifiable claims about training compute usage.

**Many chips today have dedicated security modules, including a dedicated processor, that are responsible for handling private keys and performing other security-related functions.**

## Challenges for Implementation

Many of the required features for on-chip governance mechanisms already are present on commercial devices. Apple's iPhone is one of the most well-realized implementations. The secure boot functionality of an iPhone aims to ensure that only legitimate firmware and legitimate versions of the iOS operating system can be booted. Because only legitimate versions of iOS can be booted, Apple can tightly control the apps that can be run. This functionality is enabled in part by the Apple Secure Enclave Processor, a security module also found on other Apple devices such as MacBooks.<sup>80</sup>

Many of these features also are present on the world's leading AI GPU, the NVIDIA H100. It and most other NVIDIA GPUs include a dedicated security module.<sup>81</sup> The H100 also includes a TEE known as "NVIDIA Confidential Computing."<sup>82</sup> The H100 is relatively uncommon among GPUs for having a TEE, but TEEs are relatively common on CPUs, as are dedicated security modules.<sup>83</sup> Despite its advanced features, the H100 still may not support all of the mechanisms required for an ideal implementation of the governance measures described in this report, even with appropriate firmware updates. However, this example, together with the commercial hardware licensing schemes already implemented by Intel and IBM, shows that the features discussed thus far are likely feasible and economical

to implement on AI chips.<sup>84</sup> Given that NVIDIA chips are by far the most capable and popular for training cutting-edge models, it would be valuable to build on or refine their existing security features into an initial implementation of on-chip governance mechanisms.<sup>85</sup>

More challenging will be ensuring the integrity of these mechanisms in the face of efforts by determined and well-resourced adversaries. In real-world applications, the security features and mechanisms described in the previous section would be exposed to adversarial parties attempting to compromise them in various ways. The risks of these mechanisms being misused by third-party hackers or for unlawful surveillance also must be considered.

This section analyzes these challenges in detail. It first describes the privacy and cybersecurity implications of on-chip governance mechanisms and offers thoughts on how mechanisms should be designed to avoid these issues. It then turns to the principal challenge for implementation: making on-chip governance mechanisms sufficiently secure to defend against an adversary with physical access to the chip. The section presents three prospective operating contexts and threat models and analyzes how far away current technologies are from being mature enough to deploy in each of these contexts. A more detailed discussion of the nature and feasibility of the required security technologies can be found in Appendix B.

### THE TRACK RECORD OF SIMILAR TECHNOLOGY

The Apple Secure Enclave Processor is a security module found on many Apple devices, including iPhones and MacBooks.<sup>86</sup> Its primary purpose is to protect sensitive information such as cryptographic keys. It also plays a role in Secure Boot.<sup>87</sup> Over the years that various iterations have been in use, the Apple Secure Enclave Processor has proven to be quite secure since it was first deployed in 2013.<sup>88</sup> Only one major publicly known vulnerability has been discovered, in 2020.<sup>89</sup>

This is despite the Processor being subject to substantial amounts of security research<sup>90</sup>—and strong interest in circumventing these safeguards from much of Apple's customer base. Circumventing secure boot on iPhones is popularly known as "jailbreaking" iPhones. While jailbreaks were common in the early- to mid-2010s, publicly known ways to jailbreak the most recent iPhones have become much rarer since the late 2010s as Apple has improved its security. Today jailbreaking is only possible if the phone's operating system hasn't been updated in several years.<sup>91</sup>

Other relevant efforts to secure hardware against attacks have not necessarily achieved this level of success. In 2021, NVIDIA introduced "Lite Hash Rate" (LHR) limitations on some of its GeForce gaming GPUs.<sup>92</sup> The purpose of the LHR feature was to limit the cryptocurrency mining performance ("hash rate") of these GPUs to ensure the availability of gaming GPUs for gamers, with cryptocurrency miners instead purchasing NVIDIA's dedicated line of cryptocurrency mining GPUs.<sup>93</sup> The hash rate limiter appears to have been based on secure boot features verifying that the code controlling the GPU was legitimate.<sup>94</sup> That code then looked for a certain pattern of memory accesses to detect cryptocurrency mining, and then throttled the performance of the GPU.<sup>95</sup> However, methods for partial circumvention were developed in a few months, and full circumvention was achieved a little more than a year after the release of the restricted GPUs.<sup>96</sup> Full circumvention reportedly became possible after a hack of NVIDIA's code base revealed that the code used to detect memory access patterns could be fooled into constantly resetting its internal counter.<sup>97</sup>

## Privacy, Surveillance, and Cybersecurity Implications

One of the most immediate concerns for on-chip governance mechanisms is their potential to be misused, either by the owner of the mechanism to conduct unlawful surveillance or by third party hackers taking advantage of insecure “back doors.”<sup>98</sup>

First, on-chip governance mechanisms should be designed to minimize the danger of such misuse. In particular, mechanisms for remotely disabling chips should be designed to respond to the absence of authorization, rather than an active shut-down signal. This means that if someone stole the keys to this system, the only misuse that would be possible would be to stop the chips from being disabled. This, of course, would be very damaging for the intended goal of the mechanisms but would not enable directly harmful misuse, such as an unexpected shutdown signal during a period of crucial operation. The previous section emphasizes robust secure boot functionality in part because it increases the security of devices, by making it more difficult for malware to compromise low levels of the software stack, rather than making any type of attack or misuse more feasible.

Relatedly, verification systems could and should be designed such that the compute operator is responsible for communicating the verified claims to the verifier. There is no need for verifiers to be able to read information from the system unilaterally, and if the verifier does not have that capability, no third parties can exploit the capability. Instead of unilateral surveillance, this should

### On-chip governance mechanisms should be designed to minimize the danger of misuse.

be thought of as a collaboration between a verifier and the chip owner. This collaboration also could be made fully privacy-preserving (i.e., not revealing sensitive code or data) using techniques from multi-party and confidential computing.<sup>99</sup> If a chip owner refuses to engage in such a collaboration, restriction mechanisms could allow the verifier (e.g., a regulator or device manufacturer with particular terms of use or enforcing the terms of an export license) to prevent them from continuing to use the chip.

There have been some concerns that security modules similar to the type proposed here can provide “back doors” to computers.<sup>100</sup> Traditionally, security modules and system processors have had an extreme level of trust

and privileged access, such that if an attacker can compromise such a component, they can bypass other forms of security.<sup>101</sup> However, practically all CPUs and GPUs have system processors that are at least as concerning from this perspective as a security module would be, given the inherent advantages in the security module’s security due to its simplicity. Security modules also should be designed to have limited access to the rest of the chip, such that compromising the module would not allow sensitive data to be exfiltrated.<sup>102</sup>

On-chip governance mechanisms should not be used to share any kind of personal data. The verification-based approach proposed in this report allows the compute owner to choose what kind of information is shared and removes the ability of a verifier or controller to directly acquire sensitive data. These mechanisms will be appropriate only for chips used in particular contexts, such as where export control violations are likely, or to support domestic regulation lawfully governing the usage of AI chips. This kind of limited application appears well-supported by current norms and laws: a report from the Center for Strategic and International Studies analyzes the privacy implications of collecting or requiring the collection of commercial data in an export controls context and finds that to date, foreign countries’ domestic digital privacy frameworks explicitly focus on personal data while leaving commercial data more open.<sup>103</sup>

## Overview of Threat Models and Defenses

The threat models considered here assume that the attacker has physical access to the AI hardware.<sup>104</sup> Different types of attackers will have different levels of willingness to spend resources to circumvent a mechanism, and different degrees of “covertness”—the desire to avoid being discovered to have attempted to circumvent a mechanism.<sup>105</sup> Based on these considerations, this report loosely groups attackers into three threat models of increasing difficulty:

- **Minimally adversarial contexts**, where attackers do not spend much on attacks, and are very averse to being discovered attempting to compromise mechanisms
- **Covertly adversarial contexts**, where attackers are more willing to spend substantial resources to compromise mechanisms, but still want to avoid being caught doing so
- **Openly adversarial contexts**, where attackers are willing to spend very significant resources to compromise mechanisms and are indifferent to this being discovered.

## OVERVIEW OF THREAT MODELS AND REQUIRED PROTECTIONS

Threat model	Key attacker properties	Protections required	Example applications	Feasibility	Time to implement <i>minimal</i> solution	Time to implement <i>ideal</i> solution
<b>Minimally adversarial</b>	Low resources, highly covert	Basic security measures	Domestic regulation, export control enforcement on cloud services	<b>High:</b> Current level of security likely sufficient.	<b>Months:</b> Some mechanisms could be implemented as changes to firmware and chip configuration.	<b>2–5 years:</b> There are likely software- and hardware-level vulnerabilities in current security features.
<b>Covertly adversarial</b>	Moderate to high resources, covert	Exceptionally secure software, tamper-evidence	Export control enforcement against large companies, treaty verification	<b>Moderate:</b> Significant additional investment in software security and tamper-evidence required.	<b>Months:</b> Firmware changes and ad hoc tamper-evidence likely could be implemented in months, and may be sufficient in some cases.	<b>2–5 years:</b> There are likely hardware-level vulnerabilities in current hardware security features. Improved tamper-evident features also could take years to reach large-scale production.
<b>Openly adversarial</b>	High resources, non-covert	Provably secure software, tamper-proofing	More challenging cases of export control enforcement and treaty verification, where other deterrence fails.	<b>Uncertain:</b> Significant investments in software and hardware security may be sufficient.	<b>2–3 years:</b> Hardware-level vulnerabilities would need to be resolved, and rudimentary tamper-proofing measures would need to be developed.	<b>4–8 years:</b> Truly robust tamper-proof packaging could take years to develop and test, due to the need for slow physical production and testing processes.

Each of these categories requires a distinct approach to defense. The table below summarizes these different approaches. In all three contexts, physical, firmware, and software security are important. A detailed discussion of the nature and feasibility of the required security features in each of these areas can be found in Appendix B.

## MINIMALLY ADVERSARIAL CONTEXTS

In these contexts, would-be attackers do not spend much on attacks and are very averse to being discovered attempting to compromise mechanisms. Would-be attackers of this type would be, for example, technology companies based in the United States or friendly nations that are subject to regulations related to training computation usage or other development practices. Such companies would be very likely to comply with inspections and have limited motivation to circumvent the restrictions. The level

of security already present on existing hardware security features and software and firmware likely would be sufficient for such actors.

As an illustrative case study, NVIDIA's software license agreement currently bans the use of its gaming GPUs in data centers.<sup>106</sup> Even though gaming GPUs can be viable sometimes as more affordable alternatives to data center chips, and NVIDIA has limited ability to directly enforce this license agreement, no major U.S. cloud provider offers cloud AI computing services based on gaming GPUs.<sup>107</sup>

Some other examples of minimally adversarial contexts include:

- Compute vendors enforcing license agreements
- Enforcement and monitoring of domestic regulation
- Treaty verification between countries with high mutual trust

- Auditing and agreements between AI companies with high trust in each other.

In characterizing a situation as minimally adversarial, policymakers and counterparties will need to consider how much these actors would have to gain from circumventing a mechanism. In many cases, an actor may not have much to gain. But in some cases, skirting a regulation might allow, for example, a company to gain billions of dollars' worth of market share via developing a better AI system. In such a case, a company may be willing to spend substantial resources circumventing a restriction or monitoring system. One analogous example would be the Volkswagen emissions scandal.<sup>108</sup> In such cases, it is especially important to ensure on-chip governance mechanisms can resist sophisticated attacks, and characterizing them as covertly adversarial may be more appropriate.

#### COVERTLY ADVERSARIAL CONTEXTS

In these contexts, attackers are more willing to spend substantial resources to compromise mechanisms but still want to avoid being caught. Companies in some countries, such as China, historically have shown less respect for license agreements or intellectual property and may be relatively willing to attempt attacks on on-chip governance mechanisms. However, given threats, for example, of being cut off from the supply of further chips, or broader U.S. sanctions, these companies would face incentives against attempting these attacks openly. Many potential applications of on-chip mechanisms for export control enforcement and international agreements therefore can be characterized as covertly adversarial contexts.

In covertly adversarial contexts, if a high degree of software security has been achieved, the key to defense becomes tamper-evidence: ensuring that any physical tampering would leave physical evidence that could be discovered by inspectors. If inspections (either in person or remote, if the technology exists) are feasible, and violators can and would be effectively punished, tamper-evidence should be sufficient to achieve deterrence. Tamper-evidence appears relatively easily achievable from a technical implementation perspective. See Appendix B for further details.

#### OPENLY ADVERSARIAL CONTEXTS

In openly adversarial contexts, tampering efforts cannot be deterred by threats of punishment or penalties. This likely would be the case if export-controlled chips have ended up in the hands of an uncooperative foreign military or other powerful state-linked actors.<sup>109</sup>

#### THINKING IN TERMS OF COST IMPOSITION

When considering whether a given set of defenses would be sufficient, it is important to consider that the most dangerous forms of export control circumvention likely would require an attacker to overcome mechanisms on *large numbers* of chips (many thousands), either to train powerful AI models or to deploy them at scale.

This is both advantageous and disadvantageous for the defender. On the one hand, evidence of tampering with large numbers of chips would be easier to discover, and labor-intensive tampering would be very expensive. On the other hand, the need to tamper with large numbers of chips means that up-front costs of developing an attack can be spread across many chips, which can make some types of attacks look cheap relative to their payoff. For example, if, at some point in the future, a foreign military illegally acquires \$500 million worth of AI chips (around 10,000 leading-edge AI chips at today's prices), it could be worth it for them to spend another \$500 million to develop a way of defeating the remote disabling mechanism on the chips. On the other hand, costs that need to be paid for each chip will become very large. This is important for physical attacks that would require the use of very sophisticated equipment and skilled labor.

It thus becomes important to design security measures that impose high per-chip costs on attackers. It also is important that any single points of failure that would allow scalable attacks, such as firmware vulnerabilities, need to be designed to withstand very well-resourced attackers.

International treaty verification and enforcement also could sometimes be appropriate to treat as openly adversarial. For example, if a country with strong incentives to "cheat" the terms of a treaty has been allowed to amass powerful chips under the conditions of that treaty, it would be ideal if the chips were secure enough that the country could not violate their treaty commitments, even if they were willing to openly renege on those commitments.

All of this means that on-chip governance mechanisms operating in such contexts should be tamper-proof. Tamper-proofing refers to defenses that detect tampering efforts and respond by destroying whatever the attacker was attempting to access. Tamper-proofing like this is currently used on some dedicated hardware security modules, but no existing solutions on the market appear to be applicable to AI chips. It seems likely, but not certain, that effective tamper-proofing for AI chips could be developed, but this likely would require investment and time to develop and deploy at scale. See Appendix B for further details.

## Implementation Timelines

If the capabilities and national security risks of broadly capable AI systems continue to grow at the pace seen in 2022 and 2023, the need for highly effective controls will become acute in several years’ time. Crucially, developing and deploying the governance mechanisms described in this report will take time (months in the most optimistic case, years in the most likely case). This suggests that policymakers concerned about this issue should begin formulating policies and preparing appropriate technologies now. Once the relevant security features have been mandated in the most powerful AI chips, they also need not be used immediately: The mechanisms described in the previous section would allow for rapid and flexible responses to new developments and threats once installed.

The 2022 U.S. export controls targeting AI chips are an excellent example of the importance of acting early when governing computing hardware. To simplify, the export of any chip equal to or better than the NVIDIA A100 to China was restricted. At the time of imposition, these controls had likely minimal effect China’s AI industry, because thousands of affected AI chips already were present in China, and chips of similar performance to the A100 were still uncontrolled.<sup>110</sup> But if these controls are kept in place for years, the difference between the best chips on the market and the best chips that Chinese AI developers can legally obtain in 2027 will be likely substantial.<sup>111</sup> Another key lesson is these export controls were updated a year later to be

more effective and close key loopholes (and will likely continue to be updated). This gives additional reason to begin any similarly technically complex rulemaking process early.

It likely will take 18 months to 4 years to robustly harden the technologies required for on-chip governance mechanisms, and a further 4 years for chips with these mechanisms to become sufficiently widespread for these mechanisms to be broadly effective. However, intermediate stages of technological development still will be useful in production contexts. In the short term, firmware updates could be deployed to any AI chips with the necessary security features. This would initiate a “testing phase” for on-chip governance mechanisms, where their usage would be limited to minimally adversarial environments and/or environments where in-person inspections are possible.

The impact of the additional lag introduced by “sufficient uptake” could be mitigated by tracking the sale of AI chips before the introduction of on-chip governance mechanisms and restricting their sale to specific actors. For example, the broad ban on the export of high-end AI chips to China and Russia could be kept in place until effective on-chip governance mechanisms have been implemented, at which point licenses could be granted under certain conditions.<sup>112</sup> Recently, the Bureau of Industry and Security suggested that they could make exceptions to export controls for chips equipped with technical mechanisms that would prevent the chips from being used for powerful AI training, and requested proposals for such mechanisms.<sup>113</sup>

### IMPLEMENTATION STAGES FOR ON-CHIP GOVERNANCE

Stage	Required steps and dependencies	Expected duration
Policy formulation	Establish policies that require or incentivize chip firms’ implementation of on-chip governance mechanisms. Draft requirements should be communicated to chip companies as early as possible to ensure that technical work can commence.	~1 year
Technical development	Develop secure versions of on-chip governance mechanisms based on hardened security modules and other defenses (See Appendix B). Can begin once requirements from the previous stage are sufficiently clear.	18 months to 4 years
Sufficient uptake	To ensure that all or most cutting-edge AI development can be governed by on-chip mechanisms, these chips first will have to see uptake by the large commercial entities developing the most powerful AI systems. As a rule of thumb, it is assumed that chips that are four years old (approximately two GPU generations) are no longer cost-competitive.	4 years

## Timelines for Technical Development of Security Features

This report defines the goal of technical development as a hardened security module included on all high-performance data center AI chips that can ensure that the chip has valid, up-to-date firmware and software and, where applicable, an up-to-date license. The security module would block the chip from operating if these conditions were not met. This valid, up-to-date firmware and software then could help enforce limits on the uses of these chips and offer sophisticated remote attestation capabilities. The security module could ensure that if vulnerabilities are found in this firmware and software, users would have no choice but to update to patched versions where the vulnerability has been fixed. Technical R&D to support such an implementation would involve:

- Implementations of security modules and trusted execution environments applicable to cutting-edge AI chips, including license requirements and remote attestation
- Development of tamper-evident and tamper-proof technologies specific to high-performance data center chips
- Potential additional features, such as communication between chips to ascertain and report use in large clusters, latency-based geolocation, or logging in secure non-volatile memory
- Red-teaming, verifying, or otherwise enhancing the security of the above features.

The rest of this section offers more detailed estimates of the time required to design and implement sufficient defenses for different operating contexts and threat models, drawing on conversations with chip industry experts.<sup>114</sup>

For minimally adversarial contexts the current level of hardware security likely would be sufficient, and thus

many mechanisms could be implemented as firmware updates. This would take a few months. Some mechanisms may not be possible to implement this way on current hardware, in which case silicon-level changes would be required, and the time to implement them would increase to between 18 months and 4 years.<sup>115</sup>

For covertly adversarial contexts, a minimal solution likely could be deployed in a few months, using a combination of firmware changes and post hoc tamper-evident measures, such as adding tamper-evident seals to server cases. However, given the mixed track record of similar existing efforts, it is likely that the implementations of security features such as secure boot found on existing AI chips include “unpatchable” vulnerabilities that a well-resourced adversary could find. Therefore, a solution like this should not be considered fully trustworthy but may be acceptable to deploy in cases where there is sufficient monitoring, and sufficient capability and willingness to deter evasion attempts through legal means.

Designing and thoroughly testing a highly secure basic security module likely would take at least a year, and it would need to be finished at least a year before the chip enters the market. Thorough external testing of the finished product could add at least a year to this. Ideally, this would be combined with custom tamper-evident packaging and protections against side channel and fault injection attacks. Developing and scaling up the production of novel physical protections could take years but could be done concurrently with the development of the security module.

For openly adversarial contexts, an extremely well-secured security module would be a necessity, due to having little ability to deter hacking. Additionally, some kind of tamper-proof envelope would be required. Developing and producing such tamper-proofing features likely would take several years, due to the unsuitability of existing solutions, and the need to prototype and physically test novel physical mechanisms, and then scale up their production.

## Recommendations

On-chip governance mechanisms present a promising area for further research for computer engineers, computer scientists, and policy researchers. This report offers the following recommendations to move toward a world where all leading AI chips are both secure and governable:

### Establish government coordination

**The White House should issue an executive order establishing a NIST-led interagency working group, focused on getting on-chip governance mechanisms built into all export-controlled data center AI chips.**

For on-chip governance to reach commercial scale, long-term collaboration between government and industry will be required. For progress to be made on the time scale required, an executive order is an appropriate forcing function. An executive order also could include other important initiatives to secure the AI supply chain, such as cross-agency coordination to tackle AI chip smuggling and better track other critical inputs to AI.<sup>116</sup>

The National Institute of Standards and Technology (NIST) would make a suitable lead for this effort. Relevant existing NIST initiatives include the CHIPS Program Office, and the Cryptographic Module Validation and Hardware-Based Confidential Computing programs.<sup>117</sup> Expertise and staff also should be drawn from the following agencies and offices:

- The Department of Energy (Sandia National Lab)
- The Department of Commerce (Bureau of Industry and Security and the Office for Policy and Strategic Planning)
- The Department of Defense (DARPA and microelectronics-focused groups)
- The Department of Homeland Security (Cybersecurity & Infrastructure Security Agency)
- The U.S. intelligence community (National Security Agency)
- The National Science Foundation (Center for Hardware and Embedded System Security and Trust).

While the implementation of on-chip governance mechanisms efforts can be broken down further into

distinct policy and technical efforts, central oversight and steering will help:

- Ground policy development and implementation in technical findings and efforts, and conversely, target technical efforts toward addressing policy issues seen as most compelling
- Account for synergies and dependencies within different areas of effort (for example, ensuring tampering countermeasures are applicable to the most promising security module implementations)
- Provide a single point of contact for industry.

This program should be informed by a technical panel drawn from industry, academia, and government to evaluate feasibility and challenges (including those around cost and time frames) for technical work toward the implementation of on-chip governance mechanisms. This panel likely will need to draw on both unclassified and classified information (for example, through classified meetings and reporting annexes) to benefit fully from both nongovernment academic and industry expertise and knowledge around the state-of-the-art for secure computing hardware, and relevant offensive capabilities, as held by national laboratories and the intelligence community.

### Create commercial incentives

**The Department of Commerce (DoC) should incentivize U.S. chip designers to conduct necessary R&D using advance export market commitments.**

Given that on-chip governance mechanisms need to be implemented on commercial chips, much of the necessary R&D will need to happen in an industry setting. Advance market commitments are contracts offered by a government to guarantee a viable market for a product once it has been successfully developed.<sup>118</sup> BIS has already suggested they could except certain chips from export controls if they meet a set of (yet to be defined) technical requirements.<sup>119</sup> They should now make this explicit by using advance market commitments that guarantee export market access, conditional on firms provably implementing a specific set of security features on their data center AI chips.

Export market commitments could include not extending export controls to new jurisdictions, relaxing the presumption of denial licensing policy for chip exports to lower-risk customers in China, or moving

toward more surgical end-use or end-user-based controls. These commitments could be an effective way of incentivizing development without spending public money: NVIDIA has estimated lost revenue of up to \$400 million in Q4 2022 as a result of existing controls.<sup>120</sup> This figure is likely much higher today, given NVIDIA's data center revenue has more than doubled.<sup>121</sup>

A key challenge is ensuring that technical requirements are adequately defined. Different tiers of requirements could be appropriate for different export geographies. The DoC should develop these requirements by analyzing specific attacker threat models in different export contexts, drawing on expertise from the National Security Agency and Cybersecurity & Infrastructure Security Agency.

### **Accelerate security R&D**

**NIST should coordinate with industry and relevant government funding bodies to fund and support hardware security R&D that can be conducted outside leading chip companies and integrated later.**

While the bulk of R&D for on-chip governance will need to be conducted by the firms building and selling AI chips at scale, some work may be conducted usefully outside of these firms, especially technologies that would benefit from being standardized across the industry. NIST (and the CHIPS Program Office within NIST) should coordinate with the Semiconductor Research Corporation, DARPA, and other relevant government funding bodies to fund useful R&D performed by academic and/or commercial partners.<sup>122</sup>

For example, R&D on specialized tamper-proof enclosures (physical housings for chips that prevent the chip from being modified without compromising its operation) for high-end chips could be potentially (partly) outsourced to academic and commercial hardware security labs. There are many precedents for this: The DARPA-supported Morello program and NIST-led Supply Chain Assurance project are examples of programs in the hardware security space that include academic and/or commercial partners.<sup>123</sup> One promising set of commercial partners are firms that develop “ruggedized” AI servers for national security or other sensitive applications. Such firms typically offer products that incorporate leading AI chips in form factors optimized for challenging environments.<sup>124</sup>

To support these projects, NIST could expand on its work on Hardware-Enabled Security to create technical standards and reference implementations for on-chip

governance mechanisms that are designed for wide adoption by industry.<sup>125</sup>

### **Plan for a staged rollout and fund extensive red-teaming**

**To ensure that on-chip governance mechanisms are properly designed and safely introduced, the Department of Commerce and Department of Homeland Security (DHS) should establish flexible export licensing and red-teaming programs.**

On-chip mechanisms will require substantial testing before being relied on in more adversarial environments, such as exports of controlled chips to the PRC. To facilitate a staged rollout approach where mechanisms can be depended upon in successively more challenging operating contexts, BIS should create export licensing arrangements where licenses can be flexibly granted for different geographies based on the security features on the device to be exported. This would allow BIS to test the utility of different hardware-based mechanisms for export control enforcement and develop robust technical standards, and it also would allow chip firms to receive feedback from their customers to improve their designs. Theoretically, this process could begin immediately with firmware updates to currently controlled chips.

In tandem, the Cybersecurity and Infrastructure Security Agency (CISA, within the Department of Homeland Security) should establish red-teaming and bug bounty programs to help find and patch any software and hardware security vulnerabilities in AI hardware. These programs could fit within CISA's “Secure by Design” program. They also would benefit from technical expertise and input from DARPA, which has run similar exercises as part of the System Security Integration Through Hardware and Firmware (SSITH) program.<sup>126</sup> A promising near-term starting point is setting up a public prize for finding vulnerabilities in hardware security features on today's AI chips.

### **Coordinate with allies**

**The State and Commerce Departments should coordinate with allies on policies and standards for on-chip governance.**

U.S. chip suppliers such as NVIDIA currently dominate the supply of the most powerful logic chips, meaning that, conditional on successful implementation, the United States could realize many of the policy benefits

from on-chip governance mechanisms through unilateral action. However, to mitigate risks to the potential effectiveness of an on-chip mechanism policy from advances in foreign chip design and production, the United States should seek buy-in and harmonization with countries occupying key chokepoints—particularly Taiwan, the Netherlands, South Korea, and Japan.<sup>127</sup> Looking beyond export control coordination, using on-chip governance mechanisms to facilitate AI governance cooperation (e.g., international agreements on compute usage reporting) would benefit from close coordination with like-minded allies such as the United Kingdom and the European Union.<sup>128</sup>

### Encourage AI chip firms to move early

#### Chip firms should move early to build and harden the security features required for on-chip governance.

If the U.S. government looks to realize the national security and governance benefits of on-chip governance mechanisms, chip suppliers that are more able to apply and build on existing technical efforts will have a head start on demonstrating and realizing compliance, with potential benefits in terms of access to markets that are the subject of export controls or other relevant regulation. Leading chip suppliers (as well as other industry players with relevant capabilities), should build on and harden existing security features toward enabling on-chip governance mechanisms.

## Limitations and Conclusion

Much of this report focuses on security, as it is the principal challenge for effectively implementing on-chip governance mechanisms. However, security is a difficult topic to assess. Ultimately, the applicability of on-chip governance mechanisms for many use cases depends on hard-to-assess factors such as well-resourced adversaries' capabilities for fully invasive physical attacks, or the ability of current AI chips to resist types of attacks to which they have never been subjected.

This report's optimism about the feasibility of secure on-chip mechanisms is influenced significantly by the relative success of Apple's Secure Enclave Processor. The Processor is a relevant point of comparison since Apple devices are among the rare devices that frequently are "attacked" by their own users to circumvent built-in restrictions. However, this comparison is still far from perfect: These attackers are typically relatively poorly resourced, without very significant financial motive to succeed, and without budgets to buy expensive equipment for sophisticated tampering.

Though adequate security will represent a novel challenge, developing on-chip governance remains an urgent and important mission for addressing national security risks from AI and maintaining American technological leadership. Developing and deploying the mechanisms described in this report will take time (months in the most optimistic case, and years in the most likely case). If the capabilities and national security risks of AI systems continue to grow at the pace observed in 2022 and 2023, the need for highly effective controls will become acute in several years. This suggests that policymakers concerned about this issue should begin formulating policies and incentivizing the development of appropriate technologies now. Once the relevant security features have been mandated in the most powerful AI chips, they need not be used immediately: The mechanisms outlined in this report would allow for rapid and flexible responses to new developments and threats once installed. With ambition and coordination with industry and key allies, the United States can create a secure foundation for a more flexible and targeted form of AI governance to meet the challenges of the 21st century.

## Appendix A: Glossary for AI Compute

What follows is a brief overview of different technical concepts related to AI computing that are used in this report.

### Different Types of AI Chips

“AI chip” refers to any chip that is designed for AI applications.<sup>129</sup> This report primarily uses this term, but several related terms are used frequently:

- **AI accelerator:** In computing, accelerator generally refers to a processor or component that is specialized for some type of task, and thus accelerates performance on that task relative to only using a CPU. Thus “AI accelerator” is an umbrella term for chips, or modules on a chip, that are designed to improve performance in AI applications. The only difference between “AI chip” and “AI accelerator” is that an accelerator can be a module on a larger chip.
- **GPU, graphics processing unit:** As the name suggests, GPUs originally were designed for generating graphics, but they were discovered to be well-suited for deep learning, and have since evolved to be even better suited for AI applications. The most important producers of GPUs are currently NVIDIA and AMD, with NVIDIA having a much greater market share for AI applications.<sup>130</sup>
- **TPU, tensor processing unit:** TPUs are a type of AI accelerator developed by Google, likely the most popular non-GPU AI chip, and notable for being used for many landmark AI results achieved by Google-affiliated organizations such as DeepMind.
- **Other terms:** Many smaller chip companies have coined new terms for their AI chips. For example, the British company Graphcore calls its chips “IPUs” (intelligence processing unit).<sup>131</sup>

This report focuses especially on NVIDIA GPUs, as:

- Large-scale AI training is performed overwhelmingly with NVIDIA GPUs or Google TPUs, with few exceptions.<sup>132</sup>
- Because TPUs are operated only in Google’s own data centers, Google could implement governance mechanisms to verify and restrict the use of compute at the cloud service layer. This would make many

applications of on-chip governance mechanisms, such as export control enforcement, no longer applicable.

### Distinguishing Between AI Chips and Non-AI Chips

All the above assumes that there is a distinct set of “AI chips” that one might wish to regulate. Currently, AI chips are fairly specialized, but the most popular AI chips still have major non-AI uses, and some non-AI chips still provide decent performance for AI. In general, GPUs can be divided into data center GPUs, which are used typically for commercial purposes, and gaming GPUs, which are used typically by individual consumers for entertainment purposes.

The most commonly used chips for training large AI models are NVIDIA’s data center GPUs.<sup>133</sup> These GPUs also are used for many other applications: somewhere between 10 percent and 50 percent of the uses of NVIDIA data center chips are still non-AI.<sup>134</sup> Including all of these chips in a regulatory regime would likely have substantial costs.

At the time of writing, the most powerful *gaming* GPU is the NVIDIA RTX 4090, which is not as powerful as the last two generations of NVIDIA’s AI-focused data center GPUs: the A100 and the H100. However, consumer gaming GPUs generally have better price-performance (cost per unit of performance) than top-of-the-line data center GPUs, due to their much lower price.<sup>135</sup> This does not translate into better price-performance in large-scale training workloads, however, due to relative limitations in memory bandwidth and chip-to-chip interconnect bandwidth. Based on conversations with engineers at AI companies training large AI models, the authors expect that using gaming GPUs for large-scale AI training today would result in a significant, but not crippling, overall price-performance penalty, perhaps 2x.

Regulations related to AI chips would be more straightforward if the market were segmented more clearly into AI chips and non-AI chips. It likely would be valuable if compute vendors made more specific product differentiations. For example, NVIDIA removed support for its NVLink chip-to-chip interconnect protocol from its leading gaming GPUs. This reduced the usefulness of gaming GPUs for training powerful AI models, while having no effect on the vast majority of gamers who never would likely have used the feature. It might be possible to use on-chip mechanisms to strengthen this distinction. For example, GPUs intended primarily for

actual graphics applications could be required to be equipped with mechanisms that limit their usefulness for AI in order to create this kind of market segmentation and allow these chips to be sold with fewer restrictions.<sup>136</sup>

However, an imperfect regime that regulates only a somewhat arbitrary set of the most powerful chips still could be useful. It would make the lives of those wishing to circumvent regulations at least somewhat more difficult and would allow suspicion to be targeted particularly at actors who go out of their way to use chips not included in the set of regulated chips.

### Compute Clusters

AI chips are combined into **compute clusters**. Compute clusters are interconnected computers that work collectively to perform complex tasks. They consist of diverse hardware components and a software stack. A software stack is a collection of software programs organized in multiple levels, where each level abstracts away technical detail from the layer below.

A compute cluster may be built and operated directly by an organization that wants to utilize the compute, such as a university or a corporation, on their own premises. This configuration is often called “on-premises,” or on-prem for short. Alternatively, a compute cluster can reside in a **data center**, which is a facility dedicated to hosting computer hardware. Large data centers, especially those operated by cloud providers, can host multiple compute clusters.

Computing clusters contain several **nodes**, which effectively are individual computers. These are also known as **servers**. AI compute nodes have **CPUs** for basic functions, and specialized **AI chips**, like NVIDIA GPUs or Google TPUs, for AI-specific computations. These chips are supported by ample memory to store model weights. A relevant example of a very powerful single node would be NVIDIA’s DGX<sup>137</sup> systems, each of which has eight NVIDIA A100 GPUs. A node also will have other components, such as drives for data storage.

Training large AI models requires distributing the model across multiple AI chips and nodes, necessitating frequent synchronization of parameters. Traditionally, each node will have one or more **network interface cards** (NICs) that connect it to the cluster’s network. These NICs will be connected to specialized components, known as switches, that route traffic between nodes. AI compute clusters typically use very high-end NICs and **switches** to enable extremely high bandwidth communication across AI chips in different nodes. Typically, AI chips within a node also are directly connected together with specialized hardware, such as

NVIDIA’s NVLink. NVIDIA also is developing a specialized switch, called the NVSwitch, that connects GPUs in different nodes to each other more directly, bypassing the conventional NIC.<sup>138</sup>

The above describes the most typical structure, but different compute vendors offer different alternatives. At one extreme, Cerebras designs massive chips that integrate all of the above into a single piece of silicon.<sup>139</sup>

### An Example of a Hardware, Firmware, and Software Stack for an AI Compute Cluster

The following “stack” of components make up a compute cluster. The most important concepts to understand for this report are the firmware and the driver.

- **Hardware:** This includes the physical components of the cluster, such as CPUs, AI chips (GPUs or TPUs), memory, network switches, and network interface cards.
- **Firmware:** Firmware is the low-level software running on the hardware components, such as AI chips, switches, and network interfaces, managing their basic operations. Firmware typically is provided by the chip vendor and offers an interface between the hardware and higher-level software, including user-provided software.
- **Operating system (OS):** The OS manages resources and provides a platform for other software to run on. Examples include Linux distributions and Windows Server.
- **Drivers:** Drivers enable communication between the OS and hardware components, such as AI chips and network interfaces.
- **AI framework:** These frameworks simplify AI model development, training, and deployment. The most popular frameworks for deep learning are PyTorch and TensorFlow.
- **Model distribution software:** Libraries and tools that help distribute the AI model across multiple nodes and chips, such as NVIDIA’s NCCL (NVIDIA Collective Communications Library).
- **Applications:** This is where the custom AI models, training scripts, and data processing pipelines reside, developed by researchers or engineers to solve specific problems.

## Appendix B: Additional Security Considerations

What follows is a detailed discussion of the securing software, firmware, hardware, and the supporting ecosystem for on-chip governance. This appendix focuses primarily on physical hardware security, given that aspect differs the most for on-chip governance compared to other security contexts.

### Securing Firmware and Software

Most on-chip governance mechanisms would rely on at least some firmware, and possibly software. Even if a secure boot mechanism has verified the “integrity” of firmware and software in the sense that it is the legitimate version, this does not mean that the legitimate version is free of vulnerabilities, and securing any substantial code against adversaries is notoriously difficult.

Because attacks based on exploiting firmware and software vulnerabilities are relatively cheap, difficult to detect after the fact, and do not require physical access to the device, they should be considered in any threat model. For these reasons, they also also be assumed to be the first type of attack an adversary would attempt. Investing in other types of protections is only worthwhile if the firmware and software on a device are exceptionally secure.

It appears likely that a security module would be simple enough that it would be feasible to formally verify the correctness of all code running on the module. For example, most of the kernel code running on NVIDIA’s “Peregrine” security module is formally verified,<sup>140</sup> Apple’s Secure Enclave Processor runs an Apple-customized version of the L4 microkernel,<sup>141</sup> and a version of L4 has been formally verified.<sup>142</sup>

However, fully formally verified code does not mean unhackable code. To date, developing complex software stacks that are fully secure against well-resourced adversaries has proved prohibitively difficult. Serious efforts have been made to secure software by means of testing and more advanced methods such as formal verification, but they have failed to produce bug-free its. For example, internal NVIDIA investigations into operating system-like software running on their GPUs (the kind of software where on-chip governance mechanisms would be implemented) found that, while formally verified code had significantly fewer bugs, several bugs still could be found per week of investigation, including some that were exploitable.<sup>143</sup> In addition to the incredible complexity of modern software, the intractability of bug-free software is exacerbated by the complexity of the

underlying compilers and hardware. This complex stack gives rise to interactions that are almost impossible to fully account for during the development process.<sup>144</sup>

The saving grace of software (and firmware) is that it can be updated, and thus vulnerabilities can be fixed once found. However, this poses some difficulties in the case of on-chip governance mechanisms, as the user may not want to update their system. This can be addressed either by having the hardware enforce updates via expiring licenses, or by requiring users to regularly remotely attest to what firmware and software they are running, and imposing legal consequences on users whose systems are too far out of date. Thus, the most valuable measure to secure the software on a chip would be to implement extremely well-hardened hardware features for securely enforced updates and/or remote attestation.

Future advances in the capabilities of vulnerability-finding AI systems could impact the interplay between offense and defense significantly. On the one hand, if new systems provide significant new capabilities to attackers, this could reduce the time and expertise needed to undermine the software underlying on-chip governance mechanisms. On the other hand, these systems also could be used by defenders to more thoroughly identify and remediate vulnerabilities before (and after) products are deployed; in the long run, this could trend toward a significant defensive advantage and make effective cyber defense much more feasible than before, especially if the defender has differential access to the most advanced AI systems.<sup>145</sup>

### Securing Hardware

The central obstacle to deploying on-chip governance mechanisms today is achieving adequate hardware security: making chips either tamper-evident or tamper-proof. This section provides an overview of the technical considerations for achieving either goal.

#### TAMPER-EVIDENCE

Physical attacks, by definition, involve physically manipulating the system. This makes them much easier to detect. Methods as simple as keeping the devices under video monitoring could be sufficient.<sup>146</sup> It is also possible to use various *tamper-evident* technologies to allow inspectors to detect physical manipulation after the fact. For example, a server housing AI chips could be held together by screws that are painted over with glitter nail polish and photographed. Later, inspectors could compare the nail polish on the screws to the photos, and check whether the flecks of glitter are in the same

positions.<sup>147</sup> Tamper-evident metal seals also have been used heavily by the International Atomic Energy Agency to detect whether nuclear materials have been accessed inappropriately.<sup>148</sup> Publicly available evidence for the effectiveness of high-end tamper-evident techniques is limited. One report assessed 289 tamper-evident seals, including some used for safeguarding nuclear materials, and found that all could be defeated cheaply.<sup>149</sup> On the other hand, the authors blamed this largely on the limited resources spent on developing better seals, and expressed optimism about the feasibility of developing much more effective seals, if reasonable resources were devoted to that goal. Many tamper-evident techniques already have been developed for dedicated hardware security chips in order for those chips to meet security level 2 and above as defined in the FIPS 140 standard.<sup>150</sup>

More challenging would be providing evidence of fault injection attacks—semi-invasive attacks wherein a chip is induced to misbehave, for example by manipulating its power supply or exposing it to electromagnetic pulses.<sup>151</sup> But because this would involve exposing the chip to unusual stimuli and inducing unusual states, it may be feasible to design chips to be tamper-evident against such attacks through techniques like having a specific on-chip fuse blow if the power supply is manipulated. Additionally, fault injection attacks can be potentially mitigated and detected through specific software measures.<sup>152</sup>

Certain tamper-evident measures also could require occasional inspections. In the export control context, these could be on-site or involve short-notice mail-in requirements. While inspecting all chips presumably would involve unacceptable overhead, inspections of a small number of random and/or risk-based inspections should be sufficient to achieve statistical confidence that large-scale tampering of chips is not occurring.<sup>153</sup> Such a program could be implemented at a fairly low cost compared to the existing budget of the Bureau of Industry and Security, but likely would require additional funding beyond the Bureau's current budget to scale to global stocks of tens of millions of controlled AI chips.<sup>154</sup>

#### REMOTE TAMPER-EVIDENCE

Some hardware security features even could provide remote tamper-evidence: compute operators could be required to regularly remotely attest to the integrity of their chips. Secure boot and remote attestation provide some degree of remote tamper-evidence, in that these tools can reveal if a chip is not running legitimate firmware, or if the configuration is not as expected. However, this method may not be sufficient if the

chip itself has been physically tampered with, as the attacker also could compromise the remote attestation mechanism. There is ongoing research into developing protective enclosures for chips that could act as a physical unclonable function (PUF), and thus allow a chip to attest remotely to the integrity of the enclosure.<sup>155</sup> Techniques such as “probe signal injection” also could be used, where a physical device profile first is defined by injecting an electromagnetic signal to elicit a “signature,” and then the device is tested periodically to check if its physical signature has changed.<sup>156</sup> For each of these technologies, it might be possible to extend this technology to remotely attest to the integrity of an entire server.

#### TAMPER-PROOFING

To *prevent* physical attacks on a chip, the chip needs to have tamper-proof packaging.<sup>157</sup> This means packaging with (a) some means of detecting that it has been disturbed, and (b) the ability to take a destructive response when a disturbance is detected. Different types of responses are required in different cases. When the goal is to protect a private key, the response is simple and easy to implement: wipe the private key. This is typically called “zeroization”. When protecting the core functionality of the chip, the response would be ideally to trigger some self-destruct mechanism, to destroy the core functionality that the attacker is trying to access.<sup>158</sup>

The detection problem is similar to the “tamper-evidence” problem. Tamper-detecting envelopes often are used in high-grade hardware security modules; they are a requirement for the highest level of security defined in the FIPS 140-2 standard for cryptographic modules.<sup>159</sup> Tamper-detection usually is implemented using an envelope with current running through it, designed in such a way that its electrical properties would change if the envelope were broken. This change in electrical properties can be detected from inside the envelope, and a tamper response can be initiated. Such solutions appear to be technically feasible, but existing solutions are too bulky to be used for AI chips, as the enclosure would interfere with cooling.<sup>160</sup> However, this problem is likely solvable. Several mature technologies then could be used to implement simple self-destruct mechanisms cheaply, which the envelope could trigger upon detection of a tampering attempt.

The most sophisticated hardware security modules appear to be very difficult to attack,<sup>161</sup> and there are no publicly known cases in which they have been physically compromised.<sup>162</sup> However, this evidence is unfortunately weak. These are niche products, almost always stored such that many other layers of defense would have had to

fail for an attacker even to attempt tampering. Due to the contexts in which these devices are used, it also is likely that, even if a successful attack had occurred, the information would be classified or otherwise non-public.

Turning to self-destruct mechanisms, these are rare on commercially available chips, but such mechanisms should be relatively feasible to develop. Mature technologies exist, such as eFuses, that irreversibly modify the behavior of chips if triggered. Beyond fuses, other possible approaches include using excess voltage to deliberately damage the chip, or even extremely low-yield explosives.

To ensure that these protective measures cannot be disabled by cutting off power to the chip or removing it from the circuit board, the chip additionally needs to have a battery. The battery should be included in the tamper-proof packaging and should be able to provide sufficient power to keep the tamper-detection system active and power the zeroization or self-destruct mechanism for the duration of the life of the chip.<sup>163</sup> The chip must be programmed correspondingly to trigger the zeroization or self-destruct if that battery is about to run out.

### **Securing the Supporting Ecosystem**

In addition to targeting on-chip governance mechanisms themselves, attackers could target the systems of relevant controllers and verifiers. This may seem like a significant issue in that major companies and other organizations are quite frequently successfully attacked. For example, NVIDIA was compromised by a group of hackers in 2022, and some source code and design documents were stolen.<sup>164</sup> However, to truly compromise a well-designed on-chip governance mechanism, attackers would need to steal specific private keys. Such keys are more feasible to protect effectively than, for example, design documents that need to be accessible to large numbers of employees. As an example, the public key infrastructure upon which the security of internet traffic largely relies is rarely compromised, despite substantial incentives to do so. Indeed, the authors are not aware of any cases in which the root private key of a root certificate authority has been stolen.<sup>165</sup> Nonetheless, given their sensitivity, securing keys for on-chip governance mechanisms likely would merit particularly strong information security measures—for example, using threshold cryptography to split the storage of keys across multiple independent systems.<sup>166</sup>

Another angle of attack on the supporting ecosystem is in manufacturing supply chains. To rely on on-chip governance mechanisms in sensitive operating contexts,

regulators will need confidence that these mechanisms have not been compromised by untrusted firms or insider attacks during fabrication and packaging. While this area is outside the scope of this report, the Department of Defense’s “Trusted & Assured Microelectronics” program could provide a useful starting point for best practices.<sup>167</sup>

1. Tim Fist and Erich Grunewald, “Preventing AI Chip Smuggling to China,” Center for a New American Security, October 27, 2023, <https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china>.
2. A “training run” is a computational workload, often distributed across multiple chips, where large quantities of data are used to “train” an AI model to perform some task. The recent Executive Order requires U.S. AI developers to report any training run that exceeds a certain threshold of computation, measured in “operations.” It also requires U.S. cloud computing providers to report training runs conducted by their non-U.S. customers, using the same threshold. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” the White House, October 30, 2023, <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
3. Some have called for an “IAEA for AI” model to meet the challenges of global AI governance: John Mecklin, “Why the IAEA Model May Not Be Best for Regulating Artificial Intelligence,” *Bulletin of the Atomic Scientists*, June 9, 2023, <https://thebulletin.org/2023/06/why-the-iaea-model-may-not-be-best-for-regulating-artificial-intelligence/>.
4. Fan Mo, Zahra Tarkhani, and Hamed Haddadi, “Machine Learning with Confidential Computing: A Systematization of Knowledge,” arXiv, April 2, 2023, <http://arxiv.org/abs/2208.10134>; Fan Mo et al., “PPFL: Privacy-Preserving Federated Learning with Trusted Execution Environments,” arXiv, June 28, 2021, <http://arxiv.org/abs/2104.14380>; and Xiaoguo Li et al., “A Survey of Secure Computation Using Trusted Execution Environments,” arXiv, February 23, 2021, <http://arxiv.org/abs/2302.12150>.
5. For example, a recent White House executive order requires AI developers to report the development of models trained with “biological sequence data” above a certain computation threshold. Such regulations could evolve to require more formal verification of which dataset was used in training, especially if such regulation applied to foreign AI developers accessing U.S. compute via the cloud or U.S.-produced chips. The hardware security features described in this report could enable this, perhaps using a “Proof of Training Data” protocol of the kind described here: Dami Choi, Yonadav Shavit, and David Duvenaud, “Tools for Verifying Neural Models’ Training Data,” July 2, 2023, <https://doi.org/10.48550/arXiv.2307.00682>.
6. In the information security context, “spoofing” refers to the falsification of data by an attacker. See “Spoofing Attack,” Wikipedia, [https://en.wikipedia.org/w/index.php?title=Spoofing\\_attack&oldid=1166570796](https://en.wikipedia.org/w/index.php?title=Spoofing_attack&oldid=1166570796).
7. Ideally, this would be avoided by chip firms further differentiating consumer and data center GPU designs. However, the Commerce Department recently added a notification requirement for exports of consumer chips with AI-relevant capabilities, suggesting that some consumer GPUs may be export-controlled in the future. “Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections,” Supplementary information section C.2, 88 Fed. Reg. 73458, October 25, 2023, <https://www.federalregister.gov/d/2023-23055/p-204>.
8. Following cybersecurity conventions, this report uses the term “adversary” to refer to anyone attempting to circumvent or compromise an on-chip mechanism. Thus, the adversary need not be an adversary in a broader sense and can instead be, e.g., a company attempting to evade regulations.
9. “Advance market commitments” (AMCs), a relatively new idea, describe binding contracts offered by a government to guarantee a viable market for a product once it has been successfully developed. AMCs have seen success in incentivizing the development of new vaccines: Federation of American Scientists, “Creating Advanced Market Commitments and Prizes for Pandemic Preparedness,” <https://fas.org/publication/creating-advanced-market-commitments-and-prizes-for-pandemic-preparedness/>.
10. Stephen Nellis and Jane Lee, “U.S. Officials Order Nvidia to Halt Sales of Top AI Chips to China,” Reuters, September 1, 2022, <https://www.reuters.com/technology/nvidia-says-us-has-imposed-new-license-requirement-future-exports-china-2022-08-31/>.
11. Emily Benson and Catharine Mouradian, “Establishing a New Multilateral Export Control Regime,” Center for Strategic and International Studies, November 2, 2023, <https://www.csis.org/analysis/establishing-new-multilateral-export-control-regime>.
12. “Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections,” Supplementary information section D.2, 88 Fed. Reg. 73458, October 25, 2023, <https://www.federalregister.gov/d/2023-23055/p-350>.
13. Olexsandr Fylyppov and Tim Lister, “Russians Plunder \$5M Farm Vehicles from Ukraine—to Find They’ve Been Remotely Disabled,” CNN, May 1, 2022, <https://www.cnn.com/2022/05/01/europe/russia-farm-vehicles-ukraine-disabled-melitopol-intl/index.html>.
14. “Farmers Fight John Deere Over Who Gets to Fix an \$800,000 Tractor,” Bloomberg, March 5, 2020, <https://www.bloomberg.com/news/features/2020-03-05/farmers-fight-john-deere-over-who-gets-to-fix-an-800-000-tractor>.
15. Brookings, “Castle Bravo: The Largest U.S. Nuclear Explosion,” <https://www.brookings.edu/articles/castle-bravo-the-largest-u-s-nuclear-explosion>.

16. Bureau of Industry and Security, “Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification,” October 13, 2022, <https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor>; Gregory C. Allen, “Blocking China’s Access to AI Chips Matters to U.S. National Security,” July 31, 2023, <https://www.csis.org/analysis/blocking-chinas-access-ai-chips-matters-us-national-security>; Liza Lin and Dan Strumpf, “China’s Top Nuclear-Weapons Lab Used American Computer Chips Decades After Ban,” *Wall Street Journal*, January 29, 2023, <https://www.wsj.com/articles/chinas-top-nuclear-weapons-lab-used-american-computer-chips-decades-after-ban-11674990320>.
17. National Security Commission on Artificial Intelligence, “Final Report,” March 1, 2021, 7, <https://www.nsc.ai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf>.
18. Eduardo Baptista, “Insight: China Uses AI Software to Improve Its Surveillance Capabilities,” Reuters, April 8, 2022, <https://www.reuters.com/world/china/china-uses-ai-software-improve-its-surveillance-capabilities-2022-04-08/>.
19. Sara Goudarzi, “China’s High-Tech Surveillance Drives Oppression of Uyghurs,” *Bulletin of the Atomic Scientists*, October 27, 2022, <https://thebulletin.org/2022/10/chinas-high-tech-surveillance-drives-oppression-of-uyghurs/>.
20. Jacob Stokes, Alexander Sullivan and Noah Greene, “U.S.-China Competition and Military AI,” Center for a New American Security, July 25, 2023, <https://www.cnas.org/publications/reports/u-s-china-competition-and-military-ai>.
21. As an illustration of the capabilities at today’s frontier, GPT-4 has the ability to process both visual and text data, and achieves high human-level performance (top 20 percent of human test takers) in the majority of professional and academic exams it has been tested on, including college-level exams in mathematics, verbal reasoning, chemistry, and biology. OpenAI, “GPT-4 Technical Report,” March 14, 2023, <https://cdn.openai.com/papers/gpt-4.pdf>.
22. Daniil A. Boiko, Robert MacKnight, and Gabe Gomes, “Emergent Autonomous Scientific Research Capabilities of Large Language Models,” arXiv, April 11, 2023, <http://arxiv.org/abs/2304.05332>.
23. See, for example, “Donovan: AI-Powered Decision-Making for Defense,” Scale AI, <https://scale.com/donovan>.
24. “Frontier Threats Red Teaming for AI Safety,” Anthropic, July 26, 2023, <https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety>; *Oversight of A.I.: Principles for Regulation: Hearing Before the Senate Judiciary Committee Subcommittee on Privacy, Technology, and the Law*, 117th Cong. 8 (2023), statement of Dario Amodei, Co-Founder and CEO, Anthropic, <https://www.judiciary.senate.gov/imo/media/doc/2023-07-26--testimony--amodei.pdf>.
25. See the following proof of concept for an overview of the cyber offensive capabilities of today’s frontier models: HYAS, “EyeSpy: Cognitive Threat Agent,” August 2, 2023, [https://www.hyas.com/hubfs/HYAS\\_EyeSpy\\_Proof\\_of\\_Concept.pdf](https://www.hyas.com/hubfs/HYAS_EyeSpy_Proof_of_Concept.pdf); OpenAI, “GPT-4 System Card,” March 23, 2023, 14, <https://cdn.openai.com/papers/gpt-4-system-card.pdf>. For an example of how today’s frontier models have started to be turned into autonomous agents that can explore independently, learn new skills, and make novel discoveries, see Guanzhi Wang et al., “Voyager: An Open-Ended Embodied Agent with Large Language Models,” arXiv, May 25, 2023, <http://arxiv.org/abs/2305.16291>. For an overview of the risks such systems could pose in the near term see ; Alan Chan et al., “Harms from Increasingly Agentic Algorithmic Systems,” in 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, 651–66, <https://doi.org/10.1145/3593013.3594033>; and Richard Ngo, Lawrence Chan, and Sören Mindermann, “The Alignment Problem from a Deep Learning Perspective,” arXiv, February 22, 2023, <http://arxiv.org/abs/2209.00626>.
26. Markus Anderljung et al., “Frontier AI Regulation: Managing Emerging Risks to Public Safety,” arXiv, July 11, 2023, <http://arxiv.org/abs/2307.03718>.
27. Ben Garfinkel and Allan Dafoe, “Artificial Intelligence, Foresight, and the Offense-Defense Balance,” *War on the Rocks*, December 19, 2019, <https://warontherocks.com/2019/12/artificial-intelligence-foresight-and-the-offense-defense-balance/>; Sarah Kreps, “Democratizing Harm: Artificial Intelligence in the Hands of Nonstate Actors,” Brookings Institution, November 2021, <https://www.brookings.edu/articles/democratizing-harm-artificial-intelligence-in-the-hands-of-non-state-actors/>.
28. Bureau of Industry and Security, “Commerce Implements New Export Controls on Advanced Computing and Semiconductor Manufacturing Items to the People’s Republic of China (PRC),” October 7, 2022, <https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3158-2022-10-07-bis-press-release-advanced-computing-and-semiconductor-manufacturing-controls-final/file>.
29. Empirically derived “scaling laws” show that models steadily improve as they are made larger and trained using more computation. Pablo Villalobos, “Scaling Laws Literature Review,” *Epoch*, January 26, 2023, <https://epochai.org/blog/scaling-laws-literature-review>.
30. Jess Whittlestone et al., “Future of Compute Review—Submission of Evidence,” Centre for Long-Term Resilience, August 8, 2022, <https://www.longtermresilience.org/post/future-of-compute-review-submission-of-evidence>; Miles Brundage, Girish Sastry, et al., “Computing

- Power and the Governance of Artificial Intelligence,” *Forthcoming*, n.d.; Saif M. Khan and Alexander Mann, “AI Chips: What They Are and Why They Matter,” Center for Security and Emerging Technology, April 2020, <https://doi.org/10.51593/20190014>; Saif M. Khan, Alexander Mann, and Dahlia Peterson, “The Semiconductor Supply Chain: Assessing National Competitiveness,” Center for Security and Emerging Technology, January 2021, <https://doi.org/10.51593/20190016>; and Ben Buchanan, “The AI Triad and What It Means for National Security Strategy,” Center for Security and Emerging Technology, August 2020, <https://doi.org/10.51593/20200021>.
31. Ryan Fedasiuk, Karson Elmgren, and Ellen Lu, “Silicon Twist: Managing the Chinese Military’s Access to AI Chips,” Center for Security and Emerging Technology, June 2022, <https://doi.org/10.51593/20210068>.
  32. Erich Grunewald, “AI Chip Smuggling into China: Potential Paths, Quantities, and Countermeasures,” Institute for AI Policy & Strategy, October 2023, <https://www.iaps.ai/research/ai-chip-smuggling-into-china>.
  33. Semiconductor Industry Association, “Statement on Potential Additional Government Restrictions on Semiconductors,” July 17, 2023, <https://www.semiconductors.org/sia-statement-on-potential-additional-government-restrictions-on-semiconductors>.
  34. Hanna Dohmen, Jacob Feldgoise, Emily S. Weinstein Timothy Fist, “Controlling Access to Compute via the Cloud: Options for U.S. Policymakers, Part II,” Center for Security and Emerging Technology, June 1, 2023, <https://cset.georgetown.edu/article/controlling-access-to-compute-via-the-cloud-options-for-u-s-policymakers-part-ii/>.
  35. This idea has been discussed in “Artificial Intelligence: Challenges and Opportunities for the Department of Defense: Hearing Before the Senate Committee on Armed Services, Subcommittee on Cybersecurity,” 117th Cong. 4 (2023), statement of Jason Matheny, President and CEO, RAND, <https://www.rand.org/pubs/testimonies/CTA2723-1.html>. See also William Alan Reinsch and Emily Benson, “Digitizing Export Controls: A Trade Compliance Technology Stack?,” Center for Strategic & International Studies, December 1, 2021, <https://www.csis.org/analysis/digitizing-export-controls-trade-compliance-technology-stack>.
  36. Application-specific AI systems (“narrow AI”) also can pose serious national security risks. Given their narrow set of use cases, these systems may be somewhat more tractable to regulate than highly capable general-purpose models, which are highly dual use. On a practical level, given that narrow AI systems typically require far less compute than general-purpose AI systems, the governance mechanisms described in this document will be unlikely to be effective for controlling their proliferation.
  37. “Remote Attestation of Disaggregated Machines | Documentation,” Google Cloud, December 2022, <https://cloud.google.com/docs/security/remote-attestation>.
  38. Andrew Cunningham, “Riot Games’ Anti-Cheat Software Will Require TPM, Secure Boot on Windows 11,” *Ars Technica*, September 8, 2021, <https://arstechnica.com/gaming/2021/09/riot-games-anti-cheat-software-will-require-tpm-secure-boot-on-windows-11/>.
  39. By “data center chips” this report means chips intended for the enterprise data center market, which are usually more powerful than chips intended for consumers.
  40. Jess Whittlestone et al., “Future of Compute Review—Submission of Evidence.”
  41. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” the White House, October 30, 2023, <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
  42. Reinsch and Benson, “Digitizing Export Controls”; Sarah O’Hare O’Neal and Jack Clark, “Microsoft and Open AI [Sic] Comment on Advance Notice of Proposed Rulemaking (ANPRM) for the Identification and Review of Controls for Certain Foundational Technologies,” November 9, 2020, [https://downloads.regulations.gov/BIS-2020-0029-0056/attachment\\_1.pdf](https://downloads.regulations.gov/BIS-2020-0029-0056/attachment_1.pdf); Yonadav Shavit, “What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring,” arXiv, March 20, 2023, <https://doi.org/10.48550/arXiv.2303.11341>; and Matthew Mittelsteadt, “AI Verification: Mechanisms to Ensure AI Arms Control Compliance,” Center for Security and Emerging Technology, February 2021, <https://doi.org/10.51593/20190020>.
  43. For an overview of relevant supply chain chokepoints, see Khan, Mann, and Peterson, “The Semiconductor Supply Chain.” The United States unilaterally blocked the export of cutting-edge AI chips to China by leveraging a foreign direct product rule as part of the October 2022 wave of export controls. Gregory C. Allen, “Choking off China’s Access to the Future of AI,” Center for Strategic & International Studies, October 11, 2022, <https://www.csis.org/analysis/choking-chinas-access-future-ai>.
  44. It also may be possible for the hardware vendor to hand direct control over the mechanism to another entity, such as a government agency.
  45. This report uses the term “AI developer” to refer to organizations or teams developing AI systems, not to individuals.
  46. For example, if a company runs its own servers on its own premises, they are both operator and user. If a company is using a cloud provider, the cloud provider is the operator, and the company is the user.
  47. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” the

White House, October 30, 2023, <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.

48. “NVIDIA H100 Tensor Core GPU Architecture Overview,” NVIDIA, <https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper>.
49. Jeff Goldman, “Chip Backdoors: Assessing the Threat,” Semiconductor Engineering, August 4, 2022, <https://semiengineering.com/chip-backdoors-assessing-the-threat/>.
50. For example, an auditor could run tests on model weights without having direct access to the encrypted weights or having obtained proof about which training data was used to produce a set of model weights. See Confidential Computing Consortium, “Confidential Computing”; Dami Choi, Yonadav Shavit, and David Duvenaud, “Tools for Verifying Neural Models’ Training Data,” arXiv, July 2, 2023, <https://doi.org/10.48550/arXiv.2307.00682>.
51. In the most recent update to its semiconductor export controls, BIS has added a notification requirement for exports of consumer chips with AI-relevant capabilities. “Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections,” Supplementary information section C.2, 88 Fed. Reg. 73458, October 25, 2023, <https://www.federalregister.gov/d/2023-23055/p-204>.
52. See Megan Lamberth and Paul Scharre, “Arms Control for Artificial Intelligence,” *Texas National Security Review* 6, no. 2 (Spring 2023): 95–110, <https://doi.org/10.26153/TSW/46142>; Mauricio Baker, “Nuclear Arms Control Verification and Lessons for AI Treaties,” arXiv, April 8, 2023, <https://doi.org/10.48550/arXiv.2304.04123>; Mittelsteadt, “AI Verification.”
53. This specific-use case is highlighted in a recent request for public comment from the Bureau of Industry and Security: “Public Information on Export Controls Imposed on Advanced Computing and Semiconductor Manufacturing Items to the People’s Republic of China (PRC),” <https://www.bis.doc.gov/index.php/about-bis/news-room/2082>.
54. “Intel On Demand,” Intel, <https://www.intel.com/content/www/us/en/products/docs/ondemand/overview.html>; “Capacity on Demand - IBM Documentation,” IBM, February 8, 2022, <https://www.ibm.com/docs/en/pow-er9?topic=environment-capacity-demand>.
55. Reinsch and Benson, “Digitizing Export Controls.”
56. The speed of light is just under 300 km/ms. This means that if a chip/server responds in  $y$  ms, it is, at most  $y \times 150$  km away. In the case depicted in the diagram, by operating a trusted landmark server in Paris, France, we can be perfectly confident that any chip responding to a query from that server in less than 9 ms cannot be in Russia (Russia’s Kaliningrad enclave is  $9 \times 15 = 1350$  km from Paris.) Using ordinary internet infrastructure, which is substantially slower than the speed of light, chips as distant as London and Brussels can achieve round-trip latencies below 9 ms to Paris. This means that, in Western Europe, landmark servers spaced a few hundred kilometers apart would be sufficient to allow chips to verify that they are not in Russia.
57. Using speed-of-light communication via e.g., radio, the red circle could be expanded to be essentially equivalent to the blue circle. In many cases, it also may be possible to place the trusted server in the same datacenter as the AI chips in question, allowing much greater precision.
58. The term “security dilemma” was introduced by John Herz, “Idealist Internationalism and the Security Dilemma,” *World Politics* vol. 2, no. 2 (1950): 171–201, at p. 157. For an overview of how this concept is relevant to AI, see: Brookings. “Artificial Intelligence and the Security Dilemma,” .
59. Center for Security and Emerging Technology. “AI Accidents: An Emerging Threat.” <https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/>; Dan Hendrycks, Mantas Mazeika, and Thomas Woodsidek, “An Overview of Catastrophic AI Risks,” arXiv, October 9, 2023, <http://arxiv.org/abs/2306.12001>.
60. Stokes, Sullivan and Greene, “U.S.-China Competition and Military AI.”
61. For an overview of nuclear monitoring and verification technologies, see “Assessment of Nuclear Monitoring and Verification Technologies,” Department of Defense (Defense Science Board), January, 2014, [https://media.nti.org/pdfs/Assessment\\_of\\_Nuclear\\_Monitoring\\_and\\_Verification\\_Technologies.pdf](https://media.nti.org/pdfs/Assessment_of_Nuclear_Monitoring_and_Verification_Technologies.pdf). For an overview of analogous ideas in the AI space, see Matthew Mittelsteadt, “AI Verification: Mechanisms to Ensure AI Arms Control Compliance,” Center for Security and Emerging Technology, February 2021, <https://doi.org/10.51593/20190020>.
62. Shavit, “What Does It Take to Catch a Chinchilla?”
63. “Members,” Confidential Computing Consortium, <https://confidentialcomputing.io/about/members/>.
64. “Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections,” Supplementary information section D.2, 88 Fed. Reg. 73458, October 25, 2023, <https://www.federalregister.gov/d/2023-23055/p-350>.
65. See, for example, the Caliptra root of trust being developed by the CHIPS Alliance: CHIPS Alliance. “Caliptra: A Datacenter System on a Chip (SoC) Root of Trust (RoT),” GitHub, <https://www.opencompute.org/docu->

- ments/caliptra-silicon-rot-services-09012022-pdf. Other related efforts are discussed in this recent NIST report: Bartock, Michael, Murugiah Souppaya, Ryan Savino, Tim Knoll, Uttam Shetty, Mourad Cherfaoui, Raghu Yeluri, et al., “Hardware-Enabled Security: Enabling a Layered Approach to Platform Security for Cloud and Edge Computing Use Cases,” National Institute of Standards and Technology (U.S.), May 4, 2022, <https://doi.org/10.6028/NIST.IR.8320>.
66. Jonas B. Sandbrink, “Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools,” arXiv, August 12, 2023, <http://arxiv.org/abs/2306.13952>.
  67. “Secure Boot,” Microsoft, February 8, 2023, <https://learn.microsoft.com/en-us/windows-hardware/design/device-experiences/oem-secure-boot>; OCP Security Workgroup, “Hardware Secure Boot,” Open Compute Project, 2021, 7, <https://www.opencompute.org/documents/secure-boot-2-pdf>. This report is specifically interested in enforced secure boot. Often, secure boot without enforcement is provided as an option that the hardware user can enable or disable, particularly on PCs. This can provide valuable protection against malware but obviously does not restrict the user’s behavior.
  68. OCP security workgroup, “Attestation of System Components v1.0 Requirements and Recommendations,” November 4, 2020, <https://www.opencompute.org/documents/attestation-v1-0-20201104-pdf>; Henk Birkholz et al., “Remote ATtestation procedureS (RATS) Architecture,” Request for Comments (RFC Editor, January 2023), <https://www.rfc-editor.org/info/rfc9334>.
  69. Ross Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems*, 3rd ed. (Hoboken, NJ: John Wiley & Sons, 2020), 250–51.
  70. Wei Huang et al., “Aion Attacks: Manipulating Software Timers in Trusted Execution Environment,” Lecture Notes in Computer Science (Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, 2021), 173–93, [https://doi.org/10.1007/978-3-030-80825-9\\_9](https://doi.org/10.1007/978-3-030-80825-9_9).
  71. However, if one only tracks computations, rather than time, in theory it would be possible for an actor intending to circumvent restrictions to collect a number of powered off, authorized chips. These chips then could be used for whatever number of operations they are authorized for, even long after the overseer has stopped authorizing that actor’s chips. E.g., if an actor’s chips were authorized every 24 hours, and their authorization was revoked at the start of a 7-day training run, the training run could still be completed if they had 7 times more chips in reserve than were being used for the training run. However, due to how expensive it would be to keep such reserve compute, this is unlikely to be a major problem. This problem also could be addressed simply by shortening the license reauthorization period.
  72. It might be possible to break this using some kind of fault injection attack to prevent the clock cycle counter from incrementing. However, it would likely be extremely difficult and costly to repeatedly carry out such an attack without interfering with the normal functions of the chip, on dozens of chips in operation. See “AON Timer Technical Specification,” OpenTitan Documentation, February 23, 2023, [https://docs.opentitan.org/hw/ip/aon\\_timer/doc/index.html](https://docs.opentitan.org/hw/ip/aon_timer/doc/index.html) for an example of this approach. The attacker may be able to slow this down, but they would also be slowing down the useful computations done by the chip by the same amount.
  73. See “Top Earlgrey,” <https://opentitan.org/book/hw/top-earlgrey/index.html> for an example of this approach.
  74. Confidential Computing Consortium, “Confidential Computing,” 10.
  75. Shavit, “What Does It Take to Catch a Chinchilla?”
  76. Hashing refers to transforming data into a short alphanumeric sequence of a standardized length. Hashes are generated using algorithms such that the same data always will produce the same hash, but without the hash revealing the original data. This allows the owner of hashed data to prove that their data generated that hash, without revealing the original data itself.
  77. As part of verifying a model’s provenance, the proposed scheme involves retraining parts of the model using the provided transcripts on trusted third-party hardware, to check that the resulting weights match the originally logged hashes. This retraining step verifies that the transcripts accurately reflect the original training process.
  78. In most cases the compute operator could be trusted to store the logs, because logging would either be a voluntary action taken by the operator or required by some type of regulatory regime with the capacity to take enforcement action against anyone found to not have kept the required logs. However, it would be a useful additional security feature for the chip to have secure non-volatile storage in which to keep (cryptographic hashes of) some recent logs. This would allow an inspector to detect that something is wrong if the operator has succeeded in stealing the private key on the chip (e.g., through side channels) and forged logs but has not succeeded in otherwise tampering with the chip. Inspecting these logs in person would be costly but likely worth it in high stakes situations.
  79. Stephan van Schaik et al., “SoK: SGX.Fail: How Stuff Get eXposed,” 2022, <https://sgx.fail>; Huang et al., “Aion Attacks.”
  80. Apple, “Apple Platform Security,” May 2022, 9–17, [https://help.apple.com/pdf/security/en\\_US/apple-platform-security-guide.pdf](https://help.apple.com/pdf/security/en_US/apple-platform-security-guide.pdf).
  81. AleksandarK, “NVIDIA Unlocks GPU System Processor (GSP) for Improved System Performance,” TechPowerUp,

- March 12, 2023, <https://www.techpowerup.com/291088/nvidia-unlocks-gpu-system-processor-gsp-for-improved-system-performance>; NVIDIA, “NVIDIA Accelerated Linux Graphics Driver README and Installation Guide,” February 2023, chap. 43, [https://download.nvidia.com/XFree86/Linux-x86\\_64/530.30.02/README/gsp.html](https://download.nvidia.com/XFree86/Linux-x86_64/530.30.02/README/gsp.html). The current version is code-named “Peregrine.” Marko Mitic, “Systematically Securing the RISC-V - Secure Foundation for Embedded Functionality,” <https://www.youtube.com/watch?v=l7ilkfHvWNI>; Mike Heskin, “Ok so, Considering That: A) Nvidia Has Moved Away from Falcon for Good, Replacing It with a RISC-V Based Solution (‘Peregrine’); b) Nintendo No Longer Uses the TSEC for Secure Boot on New Switch Units;,” Tweet, Twitter, January 29, 2021, <https://twitter.com/hexkyz/status/1355168275856982019>.
82. “NVIDIA Confidential Computing,” NVIDIA, July 25, 2023, <https://www.nvidia.com/en-us/data-center/solutions/confidential-computing/>.
  83. For example, Intel’s CPUs rely on the Intel Management Engine (Rivka Gehler et al., “Intel® Converged Security and Management Engine (Intel® CSME) Security Technical White Paper,” October 2022, <https://www.intel.com/content/dam/www/public/us/en/security-advisory/documents/intel-csme-security-white-paper.pdf>, and AMD’s Epyc server CPUs are equipped with the AMD Secure Processor (“AMD Secure Encrypted Virtualization (SEV),” AMD, <https://www.amd.com/en/developer/sev.html>).
  84. “Capacity on Demand - IBM Documentation”; “Intel On Demand.”
  85. Among the most compute-intensive training runs, the vast majority of non-Google entries used NVIDIA chips. The Google entries use Google-designed TPU chips. Implementing on-chip governance mechanisms on TPUs is less urgent because Google only offers access to TPUs via their own cloud services, and thus can implement governance mechanisms at the cloud service layer. Epoch, “Parameter, Compute and Data Trends in Machine Learning,” [https://docs.google.com/spreadsheets/d/1AAIebjNsnJj\\_uKALH-bXNfn3\\_YsT6sHXTCU0q7OIPuc4/](https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALH-bXNfn3_YsT6sHXTCU0q7OIPuc4/).
  86. Apple, “Apple Platform Security,” 10–17.
  87. Apple, “Apple Platform Security,” 29.
  88. Apple, “Apple Announces iPhone 5s—The Most Forward-Thinking Smartphone in the World,” September 10, 2013, <https://www.apple.com/newsroom/2013/09/10Apple-Announces-iPhone-5s-The-Most-Forward-Thinking-Smartphone-in-the-World/>.
  89. ironPeak, “Crouching T2, Hidden Danger,” ironPeak, October 5, 2020, <https://ironpeak.be/blog/crouching-t2-hidden-danger/>.
  90. See, for example, Tarjei Mandt, Mathew Solnik, and David Wang, “Demystifying the Secure Enclave Processor,” August 2016, <http://mista.nu/research/sep-paper.pdf>; Jeremy Erickson and Misha Davidov, “Deciphering the Messages of Apple’s T2 Coprocessor,” Duo Security, February 14, 2019, <https://duo.com/labs/research/apple-t2-xpc>.
  91. As of 2023, publicly available jailbreaks only work on iPhones released several years ago, and only if the user has not updated the software for over two years. It is worth noting that not all methods to install unapproved apps require fully jailbreaking a device. For crowdsourced collations of jailbreaks, see “iOS Jailbreaking,” Wikipedia, [https://en.wikipedia.org/w/index.php?title=IOS\\_jail-breaking#By\\_device\\_and\\_OS](https://en.wikipedia.org/w/index.php?title=IOS_jail-breaking#By_device_and_OS); “Can I Jailbreak,” <https://canijailbreak.com/>.
  92. NVIDIA retired the LHR feature in October 2022 by disabling it in new driver versions (Michael Kan, “Nvidia Confirms ‘LHR’ Mining Limiter for GPUs Has Been Eliminated,” PCMag, October 14, 2022, <https://www.pcmag.com/news/nvidia-confirms-lhr-mining-limiter-has-been-eliminated-from-gpus>). The feature became obsolete after Ethereum moved to proof-of-stake and demand for GPUs for mining purposes fell.
  93. Matt Wuebbeling, “GeForce Is Made for Gaming, CMP Is Made to Mine,” NVIDIA Blog, February 18, 2021, <https://blogs.nvidia.com/blog/2021/02/18/geforce-cmp/>.
  94. This claim is based on the following statement from NVIDIA: “End users cannot remove the hash limiter from the driver. There is a secure handshake between the driver, the RTX 3060 silicon, and the BIOS (firmware) that prevents removal of the hash rate limiter.” Jacob Ridley, “Nvidia Says Its Cryptocurrency Mining Limiter ‘Cannot Be Hacked,’” PC Gamer, February 19, 2021, <https://www.pcgamer.com/nvidia-ethereum-mining-limiter-can-not-be-hacked/>.
  95. More precisely, it seems only the offending process was throttled. Lolliedieb, How the 100% LHR unlocker works (lolMiner interview), interview by Seb Hezlo, May 12, 2022, [https://www.youtube.com/watch?v=LgAr4Erm\\_4o](https://www.youtube.com/watch?v=LgAr4Erm_4o).
  96. Michael Crider, “Nvidia’s Crypto-Crippling ‘Lite Hash Rate’ GPU Tech Has Been Defeated,” PCWorld, May 9, 2022, <https://www.pcworld.com/article/698962/nvidia-rtx-cards-fully-unlocked-for-crypto-miners.html>.
  97. Lolliedieb, How the 100% LHR unlocker works (lolMiner interview).
  98. For examples of some relevant questions to consider in the context of state control, see Richard Danzig, “Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority,” Center for a New American Security, 2018, app. 1.
  99. For example, an auditor could run tests on model weights without having direct access to the encrypted weights, or obtain proof about which training data was used to pro-

- duce a set of model weights. See Confidential Computing Consortium, “Confidential Computing”; Choi, Shavit, and Duvenaud, “Tools for Verifying Neural Models’ Training Data.”
100. Ms. Smith, “Now You, Too, Can Disable Intel ME ‘backdoor’ Thanks to the NSA,” CSO Online, August 29, 2017, <https://www.csoonline.com/article/3220476/researchers-say-now-you-too-can-disable-intel-me-backdoor-thanks-to-the-nsa.html>.
  101. For example, a vulnerability in the Apple T2 security chip (an earlier iteration of the Secure Enclave Processor) allowed attackers with physical access to gain privileged access to the device (ironPeak, “Crouching T2, Hidden Danger”). The reference implementation of the widely used Trusted Platform Module 2.0 standard for security modules also recently was found to have (patchable, likely unexploitable) firmware vulnerabilities (Francisco Falcon, “Vulnerabilities in the TPM 2.0 Reference Implementation Code,” Quarkslab’s blog, March 14, 2023, <https://blog.quarkslab.com/vulnerabilities-in-the-tpm-20-reference-implementation-code.html>). Many vulnerabilities also have been discovered in the Intel Management Engine (“Search Results - ‘Intel Management Engine,’” CVE, <https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=Intel+Management+Engine>).
  102. Note that restricted access to information on the rest of the chip would trade off against the module’s ability to attest to that information.
  103. Reinsch and Benson, “Digitizing Export Controls.”
  104. The assumption of physical access is made because on-chip governance mechanisms are useful primarily in cases where an untrusted actor will or may have physical access to the device. In other contexts, such as when a trusted cloud provider wants to enforce restrictions on their customers, the restrictions can be imposed at the software level. Hardware-level implementations of restrictions still can be somewhat useful due to being particularly difficult to circumvent, but they are not qualitatively superior to software-level implementations.
  105. This report borrows the terminology of covert adversaries from Yonatan Aumann and Yehuda Lindell, “Security Against Covert Adversaries: Efficient Protocols for Realistic Adversaries,” in *Theory of Cryptography*, vol. 4392, Lecture Notes in Computer Science (Springer, 2007), 137–56, [https://doi.org/10.1007/978-3-540-70936-7\\_8](https://doi.org/10.1007/978-3-540-70936-7_8).
  106. “License for Customer Use of NVIDIA GeForce Software,” NVIDIA, <https://www.nvidia.com/en-us/drivers/geforce-license/>.
  107. Katyanna Quach, “Nvidia: Using Cheap GeForce, Titan GPUs in Servers? Haha, Nope!,” *The Register*, January 3, 2018, [https://www.theregister.com/2018/01/03/nvidia\\_server\\_gpus/](https://www.theregister.com/2018/01/03/nvidia_server_gpus/); Jordan Novet, “Nvidia Made a Change to How It Lets Developers Use Its Chips, and Some Folks Aren’t Happy,” *CNBC*, December 27, 2017, <https://www.cnbc.com/2017/12/27/nvidia-limits-data-center-uses-for-geforce-titan-gpus.html>.
  108. “Learn About Volkswagen Violations,” EPA, September 27, 2022, <https://www.epa.gov/vw/learn-about-volkswagen-violations>.
  109. Criminal organizations, although ostensibly openly adversarial, are less likely to pose a significant threat in this category. Only the most well-resourced and sophisticated actors, who often may be state-backed, would have the means and the motivation to engage in large-scale AI training or deployment that requires cutting-edge chips, or to overcome sophisticated on-chip security mechanisms.
  110. Since the 2022 export controls, technical thresholds for controlled chips have been updated to include a wider range of AI chips. “Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections,” 88 Fed. Reg. 73458, October 25, 2023, <https://www.federalregister.gov/documents/2023/10/25/2023-23055/implementation-of-additional-export-controls-certain-advanced-computing-items-supercomputer-and>.
  111. This claim is contingent on the bandwidth threshold in the 2022 export controls being sufficient to severely hamper large-scale supercomputing and AI training. Currently, chip performance in TOP/s can be scaled up indefinitely, so long as inter-chip bandwidth remains below a threshold.
  112. This report is focused on governing cutting-edge chips. However, chips lagging behind cutting edge may still have significant misuse potential in the future, as they could be used to run inference on near-frontier models, or train smaller—but still dangerous—models. Indeed, the misuse potential of a given AI chip should be expected to grow over time as algorithmic efficiency improves (Danny Hernandez and Tom B. Brown, “Measuring the Algorithmic Efficiency of Neural Networks,” arXiv, May 8, 2020, <https://doi.org/10.48550/arXiv.2005.04305>; Ege Erdil and Tamay Besiroglu, “Algorithmic Progress in Computer Vision” [arXiv, August 24, 2023], <https://doi.org/10.48550/arXiv.2212.05153>) and more powerful models become widely available. Therefore, to continue to prevent misuse effectively, it may be desirable to lower, rather than raise, the performance thresholds used to determine what kind of regulations a given chip is subject to.
  113. “Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections,” Supplementary information section D.2, 88 Fed. Reg. 73458, October 25, 2023, <https://www.federalregister.gov/d/2023-23055/p-350>.
  114. Discussions with chip industry experts (including at NVIDIA), 2023.

115. These numbers are based on estimates from current and former employees at major chip companies.
116. Fist and Grunewald, “Preventing AI Chip Smuggling to China.”
117. Computer Security Division, Information Technology Laboratory, “Cryptographic Module Validation Program | CSRC | CSRC.” CSRC | NIST, October 11, 2016, <https://csrc.nist.gov/Projects/Cryptographic-Module-Validation-Program>; Michael Bartock, Murugiah Souppaya, Jerry Wheeler, Timothy Knoll, Muthukkumaran Ramalingam, and Stefano Righi, “Hardware-Enabled Security: Hardware-Based Confidential Computing,” National Institute of Standards and Technology, February 23, 2023, <https://doi.org/10.6028/NIST.IR.8320D.ipd>.
118. Willy Chertmanm, “Creating Advanced Market Commitments and Prizes for Pandemic Preparedness,” Federation of American Scientists, <https://fas.org/publication/creating-advanced-market-commitments-and-prizes-for-pandemic-preparedness/>.
119. “Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections,” Supplementary information section D.2, 88 Fed. Reg. 73458, October 25, 2023, <https://www.federalregister.gov/d/2023-23055/p-350>.
120. Stephen Nellis, Jane Lee, and Jane Lee, “U.S. Officials Order Nvidia to Halt Sales of Top AI Chips to China,” Reuters, September 1, 2022, sec. Technology, <https://www.reuters.com/technology/nvidia-says-us-has-imposed-new-license-requirement-future-exports-china-2022-08-31/>.
121. NVIDIA Newsroom. “NVIDIA Announces Financial Results for Second Quarter Fiscal 2024,” <http://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2024>.
122. The Semiconductor Research Corporation funds academic and public-private research, and lists several hardware security-related projects in its 2023 call for research. “Semiconductor Research Corporation—SRC,” <https://www.src.org/>. The DARPA Microsystems Technology Office also supports a range of projects related to hardware security: “Microsystems Technology Office (MTO),” <https://www.darpa.mil/about-us/offices/mto>.
123. “Department of Computer Science and Technology – CHERI: The Arm Morello Board,” <https://www.cl.cam.ac.uk/research/security/ctsr/cheri/cheri-morello.html>; “Supply Chain Assurance | NCCoE,” National Institute of Standards and Technology, <https://www.nccoe.nist.gov/supply-chain-assurance>.
124. See, for example: “GPU Cards | Curtiss-Wright Defense Solutions,” <https://www.curtisswrightds.com/products/computing/gpu>; “Rugged Servers and Subsystems,” <https://www.mrcy.com/products/rugged-servers-and-subsystems>.
125. Michael Bartock, Murugiah Souppaya, Jerry Wheeler, Timothy Knoll, Muthukkumaran Ramalingam, and Stefano Righi, “Hardware-Enabled Security: Hardware-Based Confidential Computing,” National Institute of Standards and Technology, February 23, 2023, <https://doi.org/10.6028/NIST.IR.8320D.ipd>.
126. “DARPA Finding Exploits to Thwart Tampering (FETT) Bug Bounty Capture-the-Flag Qualifier (Archived),” <https://www.darpa.mil/news-events/darpa-finding-exploits-to-thwart-tampering>.
127. Saif M. Khan, “Securing Semiconductor Supply Chains,” Center for Security and Emerging Technology, January 2021, <https://doi.org/10.51593/20190017>; Gregory C. Allen, “Choking off China’s Access to the Future of AI,” Center for Strategic & International Studies, October 11, 2022, <https://www.csis.org/analysis/choking-chinas-access-future-ai>.
128. For an overview of how such cooperation could fit into a broader transatlantic technology strategy, see: Carisa Nietzsche, Emily Jin, Hannah Kelley, Emily Kilcrease, Megan Lamberth, Martijn Rasser and Alexandra Seymour, “Lighting the Path,” Center for a New American Security, August 30, 2022, <https://www.cnas.org/publications/reports/lighting-the-path>.
129. Khan and Mann, “AI Chips.”
130. Nathan Benaich and Nathan Hogarth, “Compute Index,” State of AI Report, June 2, 2023, <https://www.stateof.ai/compute>; Dylan Patel, “How Nvidia’s CUDA Monopoly In Machine Learning Is Breaking - OpenAI Triton And PyTorch 2.0,” SemiAnalysis, January 16, 2023, <https://www.semianalysis.com/p/nvidiaopenaitritonpytorch>.
131. “IPU Processors,” Graphcore, <https://www.graphcore.ai/products/ipu>.
132. Epoch, “Parameter, Compute and Data Trends in Machine Learning,” [https://docs.google.com/spreadsheets/d/1AA-IebjNsnJj\\_uKALHbXNfn3\\_YsT6sHXtCU0q7OIPuc4/](https://docs.google.com/spreadsheets/d/1AA-IebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/).
133. Epoch, “Parameter, Compute and Data Trends in Machine Learning.”
134. Discussion with former NVIDIA employee, 2023.
135. Tim Dettmers, “Which GPU(s) to Get for Deep Learning: My Experience and Advice for Using GPUs in Deep Learning,” January 30, 2023, <https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/>.
136. More specifically, a rule like this might look like “any chips above a particular theoretical FLOP/s performance limit needs to either have an acceptable mechanism for limiting usefulness for AI training be registered and regulated as an AI chip.”

137. "NVIDIA DGX A100," NVIDIA, <https://www.nvidia.com/en-us/data-center/dgx-a100/>.
138. "NVLink & NVSwitch," NVIDIA, <https://www.nvidia.com/en-us/data-center/nvlink/>.
139. "Product - System," Cerebras, <https://www.cerebras.net/product-system/>.
140. Marko Mitic, "Systematically Securing the RISC-V - Secure Foundation for Embedded Functionality," <https://www.youtube.com/watch?v=l7ilkfHvWNI>.
141. "Secure Enclave," Apple Support, May 17, 2021, <https://support.apple.com/guide/security/secure-enclave-sec59b0b31ff/web>.
142. Gerwin Klein et al., "seL4: Formal Verification of an OS Kernel," in Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles (ACM 22nd Symposium on Operating Systems Principles, ACM, 2009), 207–20, <https://doi.org/10.1145/1629575.1629596>.
143. Adam Zabrocki and Alex Tereshkin, "Exploitation in the Era of Formal Verification," <https://www.youtube.com/watch?v=TcIaZ9LW1WE>.
144. Formally verified software can have exploitable vulnerabilities if: There are flaws, e.g., logical errors in the definition of intended behavior; Vulnerabilities are introduced in the process of compiling the formally verified source into a binary; The model of the hardware that the verification system is working with is incomplete. See Zabrocki and Tereshkin, "Exploitation in the Era of Formal Verification." at 15:12. Ideally, code for on-chip governance mechanisms also would account for risks from fault-injection attacks (where a chip is induced to misbehave by, for example, manipulation of its power supply, or electromagnetic pulses) (Chad Spensky et al., "Glitching Demystified: Analyzing Control-Flow-Based Glitching Attacks and Defenses," in 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2021, 400–412, <https://doi.org/10.1109/DSN48987.2021.00051>). Formally verified languages such as SPARK can be extended to have relatively deep awareness of the hardware, and this can allow them to be used to avoid hardware-level bugs. But this requires substantial additional work, especially on custom hardware (Zabrocki and Tereshkin, "Exploitation in the Era of Formal Verification." at 13:48). On the other hand, exploiting such hardware-level bugs also requires the attacker to have a deep understanding of the hardware, so they are relatively difficult to exploit.
145. Ben Garfinkel and Allan Dafoe, "How Does the Offense-Defense Balance Scale?" *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 736–63, <https://doi.org/10.1080/01402390.2019.1631810>; Andrew Lohn and Krystal Jackson, "Will AI Make Cyber Swords or Shields?" *Center for Security and Emerging Technology*, August 2022, <https://doi.org/10.51593/2022CA002>.
146. Given the chips in question would be in data centers, very little useful information (besides evidence of physical attacks) would be leaked by a video feed of a server rack, so this seems unlikely to create privacy and intellectual property issues.
147. Kyle Rankin, "Anti-Interdiction on The Librem 5 USA," *Purism*, July 20, 2022, <https://puri.sm/posts/anti-interdiction-on-the-librem-5-usa/>.
148. Alexander Enders, "Safeguarding the Future: IAEA Looks for Improved Solutions for Passive Loop Seals for Nuclear Verification," IAEA, July 1, 2020, <https://www.iaea.org/newscenter/news/safeguarding-the-future-iaea-looks-for-improved-solutions-for-passive-loop-seals-for-nuclear-verification>.
149. Roger G. Johnston, Anthony RE Garcia, and Adam N. Pacheco, "Efficacy of Tamper-Indicating Devices," *Journal of Homeland Security*, April 16 (2002). Unfortunately for the purposes of this report, Johnston et al. did not disclose exactly which seals they tested or how their attacks worked. It therefore is difficult to say how concerning their findings truly are.
150. "Security Requirements for Cryptographic Modules," National Institute for Standards and Technology, May 2001, tbl. 2, <https://doi.org/10.6028/NIST.FIPS.140-2>.
151. Jakub Breier and Xiaolu Hou, "How Practical Are Fault Injection Attacks, Really?," *IEEE Access* 10 (2022): 113122–30, <https://doi.org/10.1109/ACCESS.2022.3217212>.
152. Spensky et al., "Glitching Demystified." Even detection of fault injection attacks with moderate probability per chip would be sufficient to achieve statistical confidence that large-scale efforts will be caught if sufficient numbers of chips are inspected.
153. Shavit, "What Does It Take to Catch a Chinchilla?"
154. Fist and Grunewald, "Preventing AI Chip Smuggling to China."
155. Vincent Immler et al., "Secure Physical Enclosures from Covers with Tamper-Resistance," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2019, 51–96, <https://doi.org/10.13154/tches.v2019.i1.51-96>. PUFs are objects that rely on their unique physical characteristics to produce particular responses for particular inputs. These characteristics can be designed to degrade if tampered with. In this case, having tested a range of input-output pairs before a PUF is sold, a verifier can then confirm whether a PUF has been tampered with by seeing if a PUF still is generating the appropriate output for a given input.
156. Carlos Moreno, Sebastian Fischmeister, and Philippe Viben, "A method and apparatus for detection of counterfeit parts, compromised or tampered components or devices, tampered systems such as local communication networks, and for secure identification of components," *World*

Intellectual Property Organization WO2021056101A1, filed September 23, 2020, and issued April 1, 2021, <https://patents.google.com/patent/WO2021056101A1/en>.

157. Some security experts may object to the term “tamper-proof,” preferring “tamper-resistant” as a more realistically achievable term. While it is true that “tamper-proof” is used usually in misleading ways, its usage in this report is intended to convey a higher standard than what “tamper-resistant” usually evokes.
158. Destroying the core functionality is ideal for two reasons: Firstly, once the chip knows it is in the hands of an adversary that is actively attempting to tamper with the chip, it is safest to simply destroy the chip rather than allow the attacker more opportunities to circumvent or disable the anti-tampering functionality. Secondly and more generally: The strongest response to tampering efforts creates the strongest deterrent. However, if the tamper-detection mechanism in use is very sensitive and produces false positives at non-trivial rates, it likely would be preferable for the chip to lock itself until it receives some unusually strong form of re-authorization. The overseer could make this re-authorization conditional on, e.g., a physical inspection of the facility to ensure no foul play.
159. “Security Requirements for Cryptographic Modules,” National Institute for Standards and Technology, May 2001, tbl. 2, <https://doi.org/10.6028/NIST.FIPS.140-2>.
160. Johannes Obermaier and Vincent Immler, “The Past, Present, and Future of Physical Security Enclosures: From Battery-Backed Monitoring to PUF-Based Inherent Security and Beyond,” *Journal of Hardware and Systems Security* 2, no. 4 (December 2018): 2–4, <https://doi.org/10.1007/s41635-018-0045-2>.
161. Obermaier and Immler, “The Past, Present, and Future of Physical Security Enclosures.”
162. More specifically, no publicly known cases where a FIPS 140-2 level 4 device has been compromised. Vincent Immler (Assistant Professor of Electrical and Computer Engineering, Oregon State University), in discussion with the author, April 19, 2023.
163. For an overview of battery-backed solutions, see Obermaier and Immler, “The Past, Present, and Future of Physical Security Enclosures.”
164. Lily Hay Newman, “The Lapsus\$ Hacking Group Is Off to a Chaotic Start,” *Wired*, March 15, 2022, <https://www.wired.com/story/lapsus-hacking-group-extortion-nvidia-samsung/>.
165. The DigiNotar case is a possible example, but it appears that the root keys were never actually extracted. Rather, attackers were able to temporarily access DigiNotar’s systems to generate unauthorized certificates (Fox-IT, “Black Tulip: Report of the Investigation into the DigiNotar Certificate Authority Breach,” August 13, 2012.). However, compromises of lower level certificate authorities are not unheard of (Dan Goodin, “Stuxnet-Style Code Signing Is More Widespread than Anyone Thought,” *Ars Technica*, November 3, 2017, <https://arstechnica.com/information-technology/2017/11/evasive-code-signed-malware-flourished-before-stuxnet-and-still-does/>).
166. Tom Simonite, “To Keep Passwords Safe from Hackers, Just Break Them into Bits,” *MIT Technology Review*, October 9, 2012, <https://www.technologyreview.com/2012/10/09/183378/to-keep-passwords-safe-from-hackers-just-break-them-into-bits/>; “Multi-Party Threshold Cryptography | CSRC,” National Institute for Standards and Technology, August 18, 2023, <https://csrc.nist.gov/projects/threshold-cryptography>.
167. “Trusted & Assured Microelectronics – DoD Research & Engineering, OUSD(R&E),” <https://www.cto.mil/tam/>.

## About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts and the public with innovative, fact-based research, ideas and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, DC, and was established in February 2007 by co-founders Kurt M. Campbell and Michèle A. Flournoy. CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and non-partisan.

©2024 Center for a New American Security

All rights reserved.

---

### CNAS Editorial

---

#### DIRECTOR OF STUDIES

Paul Scharre

#### PUBLICATIONS & EDITORIAL DIRECTOR

Maura McCarthy

#### CREATIVE DIRECTOR

Melody Cook

#### DESIGNER

Rin Rothback

---

### Cover Art & Production Notes

---

#### COVER ILLUSTRATION

Melody Cook

#### COVER PHOTOS

National Archives; Getty Images.

#### PRINTER

CSI Printing & Graphics

Printed on an HP Indigo Digital Press

---

#### Center for a New American Security

1152 15th Street, NW  
Suite 950  
Washington, DC 20005

CNAS.org

@CNASdc

---

#### CEO

Richard Fontaine

#### Executive Vice President & Director of Studies

Paul Scharre

#### Senior Vice President of Development

Anna Saito Carson

---

#### Contact Us

202.457.9400

info@cnas.org



Center for a  
New American  
Security