FLI Response to Bureau of Industry and Security (BIS): Request for Comments on Implementation of Additional Export Controls

21st February 2024

Hamza Chaudhry & Landon Klein
policy@futureoflife.org

Request for Comments on Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use (RIN 0694–AI94)

Organization

Future of Life Institute

Point of Contact

Hamza Tariq Chaudhry, US Policy Specialist. hamza@futureoflife.org

We would like to thank the Bureau of Industry and Security for the opportunity to provide comments on the October 7 Interim Final Rule (IFR), or the Rule on 'Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use' (hereafter referred to as 'AC/S IFR'). The Future of Life Institute (FLI) has a long-standing tradition of work on AI governance to mitigate the risks and maximize the benefits of artificial intelligence. For the remainder of this Request for Comment (RfC) document, we provide a brief summary of our organization's work in this space, followed by comments on the AC/S IFR. Our primary comment responds to the RfC on developing technical solutions to exempt items otherwise classified under ECCNs 3A090 and 4A090, and recommends a pilot program for a technical solution. The comment includes arguments for how the pilot program could help improve BIS export controls and mitigate threats to US economic and national-security interests. In the final section, we offer general comments to the AC/S IFR.

We look forward to continuing this correspondence and to serve as a resource for BIS efforts pertaining to AI in the months and years to come.

About the Organization

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence. Since its founding ten years ago, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, participated in the UK AI Summit, and connected leading experts in the policy and technical domains to policymakers across the US government.

FLI's wide-ranging work on artificial intelligence and beyond can be found at futureoflife.org.

Primary Comment On Hardware Governance

On the Request for Comment on Developing technical solutions to exempt items otherwise classified under ECCNs 3A090 and 4A090.

We welcome the request for technical solutions on this issue. FLI has recently been involved in multiple initiatives to create and improve technical solutions for the governance of AI hardware, including semiconductors. In this primary comment, we offer arguments in favor of technical solutions for hardware governance, and introduce a new project from FLI which seeks to improve on-chip governance.

Arguments for Technical Solutions for Hardware Governance

Technical solutions for hardware governance, and specifically chip governance, offer many benefits that can supplement top-down export controls as currently instated by BIS.

A. **Generic Export Controls More Vulnerable to Lack of Enforcement than Hardware Governance:** Export controls, especially those with a wide and expansive purview, are likely to suffer from serious gaps in enforcement. A growing informal economy around chip smuggling has already emerged over the last few years, and it is likely to grow as BIS rules become more expansive. A solution focused on hardware governance is less liable to this gap in enforcement.

B. **Hardware Governance as Less Blunt Instrument and Less Likely to Hurt US Economic Interests:** Export controls most directly target state actors, leading to conflation between 'actor' vs 'application' that may foreclose benefits and exacerbate risks to United States interests. For instance, broadly-applied export controls targeted at the People's Republic of China (PRC) do not distinguish between harmless and harmful use cases within the PRC, the former of which can be economically beneficial to the United States and reduce geo-strategic escalation. For instance, relaxing restrictions on chip exports to demonstrably low-risk customers in China helps drive the economic competitiveness of US firms. These economic benefits are integral to guaranteeing continuing US leadership in the technological frontier, and help preserve global stability. Hardware governance, a more targeted instrument, side-steps these issues with export controls, focusing on applications as opposed to actors.

C. **Hardware Governance is Privacy Preserving and Compatible with Existing Chips Technology:** New and innovative hardware governance solutions are completely compatible with the current state of the art chips sold by leading manufacturers. All relevant hardware (H100s, A100s, TPUs, etc.) have some form of "trusted platform module", a hardware device that generates random numbers, holds encryption keys, and interfaces with other hardware modules to provide security. Some new hardware (H100s in particular) has an additional hardware "secure enclave" capability, which prevents access to chosen sections of memory at the hardware level. TPM and secure enclaves already serve to prevent iPhones from being "jailbroken," and to secure biometric and other highly sensitive information in modern phones and laptops. Hence, a technical solution to hardware governance would not impose serious costs on leading chip companies to modify the architecture of chips currently in inventory or in production. Critically, as the project described below demonstrates, it is possible to use these technical solutions without creating back-channels that would harm the privacy of end-users of the chip supply chain. Accordingly, hardware governance solutions such as the one proposed below are less likely to face resistance to implementation from concerned parties.

Technical Project - Secure Hardware Solutions for Safe AI Deployment

BACKGROUND

Modern techniques in cryptography and secure hardware technology provide the building blocks to create verifiable systems that can enforce AI governance policies. For example, an un-falsifiable cryptographic proof can be created to attest that a model comes from the application of a specific code on a specific dataset. This could prevent copyright issues, or prove that a certain number of training epochs were carried out for a given model, verifying whether a threshold in compute has or has not been breached. The field of secure hardware has been evolving and has reached a stage where it can be used in production to make AI safer. While initially developed for users' devices (e.g. Apple's use of secure enclaves to securely store and process biometric data on iPhones), large server-side processors have become mature enough to tackle securely governed AI workloads.

While recent cutting-edge AI hardware, such as Intel Xeon with Intel SGX or Nvidia H100s with Confidential Computing, possess the hardware features to implement technical mechanisms for AI governance , few projects have emerged to leverage them to build AI governance tooling. The Future of Life Institute has partnered with Mithril Security, a startup pioneering the use of secure hardware with enclave-based solutions for trustworthy AI. This collaboration aims to demonstrate how AI governance policies can be enforced with cryptographic guarantees. In our first joint project, we created a proof-of-concept demonstration of confidential inference. We provide details of this work here because a crucial step to potential adoption of these mechanisms is demonstration that various use cases are practical using current technology.

DESCRIPTION OF PROJECT

Consider here two parties:

4. an AI custodian with a powerful AI model
5. an AI borrower who wants to run the model on their infrastructure but is not to be trusted with the weights directly

The AI custodian wants technical guarantees that:

1. the model weights are not directly accessible to the AI borrower.
2. trustable telemetry is provided to know how much computing is being done.
3. 3. a non-removable off-switch button can be used to shut down inference if necessary.

Current AI deployment solutions, where the model is shipped on the AI borrower infrastructure, provide no IP protection, and it is trivial for the AI borrower to extract the weights without awareness from the custodian.

Through this collaboration, we have developed a framework for packaging and deploying models in an enclave using Intel secure hardware. This enables the AI custodian to lease a model, deployed on the infrastructure of the AI borrower, while the hardware guarantees that the weights are protected, and the trustable telemetry for consumption and off-switch are enforced. While this proof-of-concept is not necessarily deployable as is due to performance (we used Intel CPUs[1],) and specific hardware attacks that need mitigation, it serves as a demonstration of how secure enclaves can enable collaboration under agreed terms between parties with potentially misaligned interests.

By building upon this work, one can imagine how the US could lease its advanced AI models to allied

1 While we used CPUs in this case, a variation of this proof of concept would also work for GPUs, as they also support the Trust Platform Module (TPM) and secure enclave architectures.

countries while ensuring the model's IP is protected and the ally's data remains confidential and not exposed to the model provider. By developing and evaluating frameworks for hardware-backed AI governance, FLI and Mithril hope to encourage the creation and use of such technical measures so that we can keep AI safe without compromising the interests of AI providers, users, or regulators.

FUTURE PROJECTS PLANNED

Many other capabilities are possible, and we plan to rollout demos and analyses of more technical governance approaches in the coming months. The topic of BIS's solicitation is one such approach: hardware could require remote approval if it identifies as part of a cluster satisfying some set of properties including size, interconnection throughput, and/or certificates of authorization. The objectives of the AC/S IFR could be further achieved through a secure training faculty, whereby an authority metes out authorized training compute cycles that are required for large training runs to be able to take place.[2] This secure training faculty could include a training monitor where all ML training runs above a threshold cluster size require, by law, licensing and compute training monitoring. In this process, licensing could be required via regulation requiring cluster limiting in all GPUs, and commitment to training monitoring could be required to obtain a license for training.

Many of these solutions can be implemented on existing and widely deployed hardware to allow AI compute governance to be backed by hardware measures. This addresses concerns that compute governance mechanisms are unenforceable or enforceable only with intrusive surveillance. The security of these measures needs testing and improvement for some scenarios, and we hope these demonstrations, and the utility of hardware-backed AI governance, will encourage both chip-makers and policymakers to include more and better versions of such security measures in upcoming hardware. Thus, while initially relying heavily on export restrictions and cooperation of data centers and cloud providers, eventually in principle on-chip mechanisms could carry the lion's share of the responsibility for enforcement.

In this spirit, we recommend that:

(a) BIS consider requiring the more robust secure enclaves on advanced ICs, rather than just TPMs, which can serve similar functions less robustly.

(b) BIS encourage and support engagement with chipmakers and other technical experts to audit, test, and improve security levels of hardware security measures.

We welcome more engagement and collaboration with BIS on this front.

2 This mechanism also facilitates various future auditability affordances.

General Comments

#1 - The Inclusion of Civil Society Groups in Input for BIS Rules

In its responses to comments to the IFR, BIS has made clear that the input of Technical Advisory Committees (TACs) is an important aspect of deliberating and instating new rules on export controls. It is also clear that the BIS annual report allows industry players to offer feedback on export control trade-offs for semiconductors, as outlined in ECRA Sections 1765 and 1752, and on national security issues under Sections 4812 and 4811. However, it's not evident if civil society actors have the same opportunities for comment and input, aside from this specific Request for Comment. There is now a significant and diverse set of AI policy groups in the civil society ecosystem, populated - as in the case of FLI - by some of the world's leading experts from academia, government and industry. These actors possess a vital viewpoint to share on export control beyond the perspectives typically shared by industry. We invite BIS to clarify and make explicit the requirement for considerable input from civil society actors when it comes to the opportunities listed above, and those in the years to come.

#2 Clarifying Rule Applying to Those States which Facilitate Third Party WMD Activities

We commend the actions outlined within the AC/S IFR to ensure that export controls facilitate the restriction of weapons of mass destruction (WMD) related activities. FLI has published multiple reports on cyber, nuclear, chemical, and biological risks that intersect with the development of advanced AI systems. However, this risk does not emanate from state actors alone. In fact, several reports published over the last year demonstrate these same threats from non-state actors. We invite that BIS clarify that the IFR applies both to states listed in Country Group D:5 (and elsewhere) that use semiconductor technology for indigenous WMD-related activities, and those that are liable to share these technologies with allied and sponsored non-state actors, which in turn could use them for furthering WMD-activities.

#3 Preventing a Chilling Effect on Friendly US-China Cooperation

We support that BIS has clarified its position on §744.6 in light of concerns that an overreach of the AC/S IFR might have a chilling effect on academic and corporate cooperation between Chinese and American persons and entities, cooperation with which may in fact forward the economic and national-security interests of the United States. We ask that the BIS expand on this acknowledgement by positively affirming in a separate section that such cooperation is welcome and within the remit of the AC/S IFR. The absence of clarity over this rule could detrimentally impact the current balance of US-China cooperation, threatening global stability and harming US national-security interests.

#4 On National Security Updates to the IFR

We welcome the BIS decision to introduce a density performance parameter to ensure that less powerful chips cannot be 'daisy-chained' into more powerful technologies and hence circumvent the principle purpose of the BIS rule. We also commend the use of a tiered approach when it comes to control of advanced integrated circuits. We hope that the BIS takes further account of emerging technological developments in hardware governance. For instance, new and innovative secure training hardware governance mechanisms (see point #5) can be required of IC makers in order to help prevent training of dual use models using unauthorized, heterogeneous distributed training.

#5 On addressing access to "development" at an infrastructure as a service (IaaS) provider by customers developing or intending to develop large dual-use AI foundation models with potential capabilities of concern, such as models exceeding certain thresholds of parameter count, training compute, and/or training data.

We welcome discussion on thresholds and potential capabilities of concern with regards to large dual-use foundation models. However, it is important to underscore that there should be explicit authority to change (and likely lower) these thresholds over time. This is because large dual-use AI foundation models with a certain set of thresholds held constant may become more powerful and dangerous to due to other factors.

For instance, algorithmic improvements in an AI model may significantly drive dual-use risk even if parameter count and training compute are held constant. In addition, the threshold of training data cannot just be quantitative but also qualitative - a model trained on higher quality or more dangerous (albeit smaller) training datasets can still present capabilities of concern.

Finally, the IFR would benefit from explicit discussion of the unique risk profile for capabilities of concern presented by dual-use AI models with widely available model weights . Models with widely available model weights at the same threshold as closed models will likely present greater potential capabilities of concern, as the guardrails from these models are more easily removed (if there are guardrails in the first place) and the models can be fine-tuned, using relatively little compute resources, to improve specific capabilities of concern.

Closing Remarks

We appreciate the thoughtful approach of BIS to the development of the AC/S IFR and are grateful for the opportunity to contribute to this important effort. We hope to continue engaging with this project and subsequent projects seeking to ensure AI does not jeopardize the continued safety, security, and wellbeing of the United States.