August 2024, v1.2

Memo: Mechanisms for Flexible Hardware-Enabled Guarantees (flexHEG)

Successful AI governance requires compliance guarantees. As AI advances, the potential for catastrophic misuse or accidental deployment of dangerous capabilities increases [1] - [4]. This poses novel and as-of-yet unsolved governance challenges. Hardware-enabled governance has emerged as a promising pathway to enable international AI governance [5] - [7].

Whether governments (or coalitions of governments) can effectively impose rules requiring safety mechanisms (of whatever kind) depends on their ability to assure that no credible efforts would be able to flout such rules. Without such an assurance, the dominant considerations would remain shaped by concerns about who can release more powerful capabilities sooner. After-the-fact penalties for rule violations, as is commonly done in law, may not on its own be a viable approach if the far-reaching consequences of an AI catastrophe could render any such enforcement action moot. If these assumptions are true, it is crucial to develop technological mechanisms which undergird the deployment and monitoring of agreed-upon rules and safety mechanisms in a way that affords high confidence among all strategic players that *no* player has the technical option of unilaterally violating any such rules.

We believe that such technological mechanisms are possible. With a consolidated R&D effort, we believe we would be able to demonstrate and document compelling prototypes of such *guaranteeable chips* by the end of calendar year 2025. The goal is to lower the barriers to adoption of such a technology–both for the hardware firms involved in producing it, and for the governmental or intergovernmental processes involved in requiring its use in some contexts. Demonstrating flexHEG's technological viability would be of substantial strategic and economic interest, bolstering the belief that the implementation of international AI governance is feasible.

Regulatory capabilities this may enable, if implemented in all high-performance AI accelerators, include:

- Limiting the size of training runs in terms of total FLOP
- Limiting the size of datasets that can be used in a training run
- Privacy-preserving verification that certain types of training data are not being used
- Privacy-preserving verification that certain model architectures or training methods are or are not used
- Requiring the possession of a non-expired license to run computations of a certain size range
- Requiring a standardized evals protocol to be incorporated into the computation graph of training for sufficiently large training runs
- Requiring model weights to be encrypted in such a way that only allows them to be used on (specific) other flexHEG-capable devices, optionally according to specific rules, thereby enabling improved security and effective governance of even distributed AI training & inference

A further benefit of a technological solution as envisioned here is that it would enable significantly more targeted and reliably governance mechanisms than, for example, export controls alone. Furthermore, it enables local and privacy preserving compliance verification, removing the need for any centralised "chip registries", geolocation capabilities or even human inspections, which could reduce strategic concerns with AI governance proposals.

We envision a technology stack enabling flexible on-chip verification of governance compliance. The central capability we wish to deliver is for a cryptographically certified firmware layer to be loaded onto a high-performance computing device, which can enable on-chip checking for assured compliance with a flexible set of rules resulting from a multilateral decision-making process. Such rules might impose, for example, limitations on the type and size of computations a chip is permitted to run, compliance with a regulatory licensing scheme, etc. There are two critical desiderata for such a set-up, which together set it apart from other compute governance proposals:

(1) First, it is essential that there be **no technically viable means to bypass the check**, including by highly sophisticated physical tampering, without rendering the computing device permanently inoperable.

(2) Second, the mechanism must **enable maximum flexibility for the decision-makers without retooling the hardware**. To achieve this, the on-device compliance processor should be capable of making its compliance determinations using general-purpose computation with inputs from various kinds of information about the code and its context.

A potential approach for designing a technology stack with the stated desiderata consists of four functional layers to be implemented through several technological mechanisms.[1][2]

- **Secure processor**: an auxiliary computing device for determining compliance, integrated into a securely assembled chassis, with sufficient hardware capabilities and software libraries to securely and flexibly implement and update the desired safety mechanisms, e.g. through:
  - Static analysis of compute graphs (e.g. maximum FLOP count) and inputs (e.g. cryptographically verified timestamp)
  - Multilateral cryptographic compliance-firmware renewed with periodic broadcast updates, e.g. using multisignature contracts
- **Compliance inputs**: peripheral connections to the compliance processor which supply it with the context which might be needed to make compliance determinations under a broad variety of potential regulatory scenarios, e.g. through:
  - Accelerator interlock, e.g. executable instructions must pass through the compliance processor in order to be loaded into the primary compute processor
  - Dynamic performance counters and/or runtime trace analysis
  - Incorruptible inputs from other compliance processors, e.g. authenticated with continuous group key agreement (CGKA)
  - Real-time clock which is covered by tamper-detection mechanisms, and resistance to any clock-specific tampering attacks
  - Optionally: provable geolocation, e.g. with time-of-flight measurements from new space-based assets, or maybe Iridium STL
- **Tamper-detection and on-device enforcement**: to detect any attempts of physical tempering and disable the chip if any such attempts were detected, e.g. via:
  - Hard-wired chassis-intrusion detection mechanism, e.g. FIPS 140-style active mesh

[1] We are indicating the technological readiness level of these different mechanisms with the following colour code: red for TRL 2, orange for TRL 3, yellow for TRL 4, purple is TRL 5, and blue is TRL 6. The desired readiness level for all of these by the end of such an effort is TLR 7, and above.

[2] What remains out of scope for this specific proposal is solutions to secure the assembly chain, e.g. through fab detection and inspection, and tracing components while in transit between manufacturing facilities.

    - RF-based chassis-intrusion detection mechanism, e.g. anti-tamper radio
    - Permanent self-disabling mechanism, e.g. nanothermite
- **Offline power source:** to ensure the functional integrity of the entire stack, it is essential that the above mechanisms do not rely on external power sources, to be be achieved e.g. through:
    - Power supplied during shipment (analogous to refrigerated shipping)
    - Batteries
    - Fail-safes in the event of power loss, e.g. a circuit inhibiting the ignition of the nanothermite itself relying on energy, such that the absence of power automatically activates the self-destructive mechanism

Call to action:

- If you're a funder and interested in accelerating work in this area, feel free to reuse ideas and language used in this memo e.g. in your own funding calls.
- Teams interested in directly pursuing work in this area may work on one or several of the following types of projects:
    - (i) 'on-paper' system design and analysis,
    - (ii) design of one or more of the components,
    - (iii) prototyping and testing implementations of one or more of the components,
    - (iv) system integration and demonstration of components,
    - (v) complete design, implementation and demonstration of the full stack.
- You can send us a 1-2 page expression of interest outlining the team composition, anticipated timelines and a rough budget.
- Note that this memo is **not** a funding call. However, we anticipate one or several funders to open such calls in the near future, and would be happy to keep you informed about these.

References

[1] Hendrycks, D., Hinton, G., Bengio, Y., Hassabis, D., Altman S., et al. (2023). **Statement on AI risk**. URL: https://www.safe.ai/statement-on-ai-risk
[2] OpenAI. (2024). **Disrupting a covert Iranian influence operation.** URL: https://openai.com/index/disrupting-a-covert-iranian-influence-operation/
[3] NCSC. (2024). **The near-term impact of AI on the cyber threat.** URL: https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat
[4] Urbina F. et al. (2022). **Dual use of artificial-intelligence-powered drug discovery**. Nature machine intelligence. DOI: https://doi.org/10.1038/s42256-022-00465-9
[5] Kulp, G. et al. (2024). **Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control**. URL: https://www.rand.org/content/dam/rand/pubs/working_papers/WRA3000/WRA3056-1/RAND_WRA3056-1.pdf
[6] Aarne, O. et al. (2024). **Secure, Governable Chips.** URL: https://www.cnas.org/publications/reports/secure-governable-chips
[7] Sastry G. et al. (2024). **Computing Power and the Governance of Artificial Intelligence**. 2024. arXiv preprint:2402.08797.