VERIFICATION OF FRONTIER AI

BRIEF OF THE SCIENTIFIC ADVISORY BOARD[1]

Scientific Advisory Board

WHAT IS "FRONTIER AI"? WHAT IS "AI VERIFICATION"?

The term "frontier AI" refers to highly capable general-purpose AI models that can perform many tasks matching or exceeding the capabilities of today's most advanced models.[2] Frontier AI is poised to transform a wide range of industries, including healthcare, agriculture, finance, transportation, and security. Given the transformative potential of this technology, major powers increasingly see frontier AI as a national security priority, raising the risk of a "frontier AI arms race".[3]

Noting this growing competition, AI experts and scientists have highlighted a wide range of associated risks, including erosion of longstanding international norms and laws, acceleration of AI integration into military systems, and strategic instability that could bring major powers into more direct forms of confrontation.[4]

One of the most important ways to reduce the risks surrounding frontier AI could be to develop a trusted, effective system of verification. In general terms, "verification" means the ability of a party to attest to the actions of another party. In the context of frontier AI, it refers to the ability to attest to a wide range of relevant claims about the development and use of AI systems, such as properties of a frontier AI training run; the application of specified safety testing and mitigations; evaluation results assessing a system's capabilities and propensities; its uses at inference time; the size and speed of inference and training compute resources; or other relevant properties—all without compromising the AI system's security (confidentiality or integrity) and proprietary information.

An internationally accepted technical approach that allowed parties to verify the claims of AI developers globally could help build a common, trusted source of information about AI capacities and trajectories. Such a step would reduce the risks of geopolitical tensions and build greater trust in AI development more generally. It could also facilitate the sharing of AI technology around the world and a more equitable distribution of AI's benefits.

This science brief provides an assessment of the technological prospects for developing an AI verification system and some general considerations on the topic.

WHY IT MATTERS:

There is a real risk that frontier AI developments could become a source of heightened tension and lead to an "AI arms race." Trusted, visible, and confidential ways to verify claims about frontier AI could help mitigate geopolitical escalation and share AI's benefits.

OPTIONS FOR AI VERIFICATION

Many properties of frontier AI models are difficult to verify directly, even by their own developers. The scientific and technological techniques to understand the inner workings of frontier AI systems are commonly referred to as "AI interpretability". A key challenge is that these techniques are currently inadequate to reliably verify the inner workings of frontier AI systems to other parties.[5]

Despite this challenge there are some properties of AI development and deployment that are possible to verify today or in the near future, such as the amount and allocation of compute resources (to training and/or inference), the (coarse or fine) geolocation of compute resources, the provenance of data inputs to training, and compliance with safety evaluation and mitigation protocols (such as those already declared voluntarily by many AI developers). AI developers and operators could be subjected to verification in four ways:

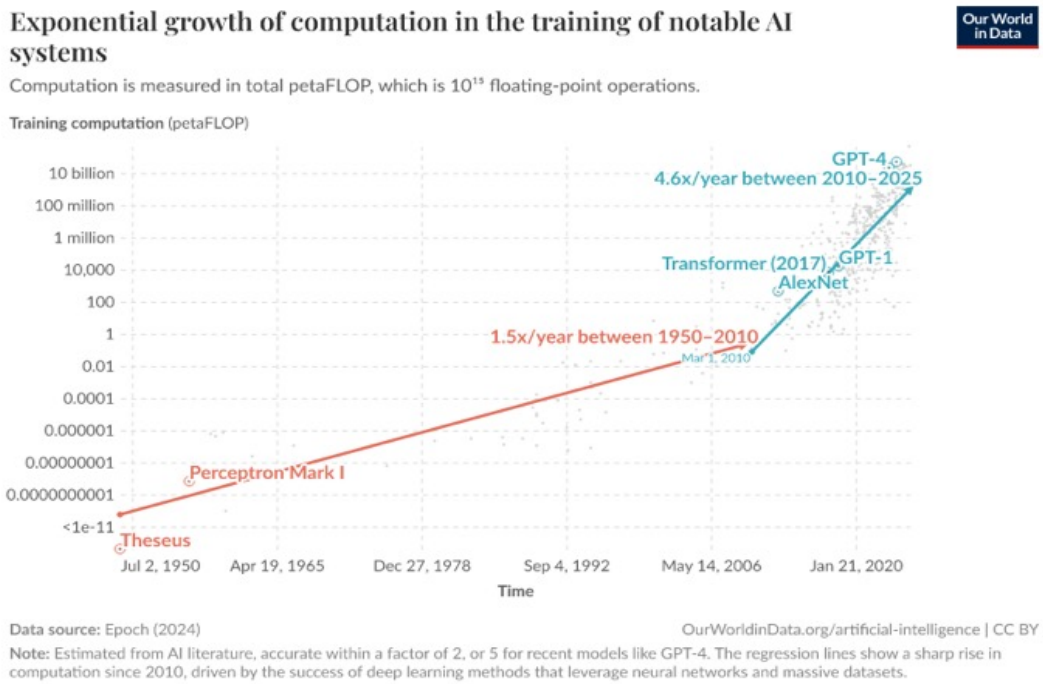
1. **Human-led Audits:** Institutions conducting audits that require some access and trust.
2. **Software Mechanisms:** Creating audit trails that are difficult to falsify, possibly using cryptographic techniques.
3. **Hardware Mechanisms:** Attesting to the actions of frontier AI developers and operators at the hardware level.[6]
4. **Whistleblower Protections:** Enabling employees to credibly attest to their employers' claims about AI development and operations.

Recent regulations by national governments and regional organizations have focused on the institutional aspects of verification.[7] The EU, US, and China, for example, have passed legislation and/or initiatives where oversight bodies conduct audits to ensure AI safety standards are met.[8] These have self-reporting requirements where AI developers must document and report on the capacities of their AI systems. Self-reporting, however, does not replace the need for external oversight and verification of frontier AI systems.[9]

VERIFICATION VIA "COMPUTE"

One of the most promising areas for verifying actions of frontier AI developers and operators is their computing power ("compute" for short).[15] By verifying compute, it would be possible to use the computing hardware of AI systems to monitor, regulate, and control frontier AI development and deployment. This can be done by controlling the movement of AI hardware and/or by modifying AI hardware designs in various ways (the latter category of approaches are often known in the literature as Hardware-Enabled Mechanisms or HEMs). The main technological arguments in favor of compute as an intervention point for verification are:

1. **Compute is necessary.** As AI systems develop wider-ranging capabilities, the reliance on computing power is set to grow. Frontier AI requires huge computing resources for training, an amount which has roughly doubled every six months for the past decade.[11]
2. **Compute is detectable:** Large-scale AI development requires substantial physical compute resources, often housed within large clusters and data centers. The physical resource intensity means they are easier to monitor than AI inputs like data and algorithms.
3. **Compute is excludable:** Unlike data and algorithms, which are easily copied and shared, compute is physical and can be controlled and restricted. This would mean that verification also could be tied to enforcement and regulation. It should be noted that this feature of compute also makes it contested, as some countries already enforce export controls on key hardware.
4. **Compute is quantifiable:** Compute resources can be measured and verified in technically viable ways, such as by examining operations per second, memory, or the connection of multiple devices. Its quantifiability would allow for clear benchmarks to be established for verification.
5. **Compute is concentrated:** Only a small number of companies currently have the computing power, resources, and



proprietary tools to develop frontier AI.[12] Approximately 90% of the world's most advanced AI chips are manufactured by a single company (TSMC). The concentration of technological capability for compute means that a verification system could be put in place, requiring buy-in from a smaller number of players.

These technical arguments in favor of using compute for verification are bolstered by the longstanding practice of private sector verifiers in many other industries. Airline operators are required to verify passenger lists to prevent unauthorized travel; financial institutions are required to verify identities to prevent money laundering; and compute providers are already required to provide information to prevent child sex abuse material online.[13]

CHALLENGES AND CONSIDERATIONS

At the same time, compute-based verification faces challenges. In a setting without cooperative access and trust, such as between geopolitical rivals, it would require improving the security of relevant hardware and software components to make them capable of detecting and responding to tampering attempts. Multilateral auditing of the software and hardware components would be required to build confidence in the absence of secret "backdoors." The security of supply chains and data centers would have to be strengthened to ensure that chips are not smuggled or manipulated before being installed in a secure enclosure.[14] AI developers can resort to renting datacenter capacity in alternative jurisdictions as a workaround to neutralize hardware-based verification that is dependent upon geolocation or tracing shipments of hardware.[15]

Finally, sophisticated solutions for on-chip verification would take a few years to develop and implement on a sufficient fraction of the world's compute. In this context, an incremental approach to compute-based verification would need to begin with other techniques, including cryptographic audit trails and/or self-reporting. The Scientific Advisory Board could consider in more detail other approaches to AI verification in future briefs.

COMPUTE-BASED REPORTING

A promising model for verifying frontier AI currently under discussion by a range of computer scientists would involve an international system for information-sharing about highly capable model development. From a technological standpoint, such an approach could leverage compute providers as key intermediaries in the oversight process. Model developers would notify their host government or a designated international body before training a highly capable model. Compute providers would then verify this notification before granting access to large-scale compute resources. At the outset, this would only apply to large training runs, avoiding undue burden on smaller entities.[17]

Such an approach could enable a licensing system for AI developers, requiring registration before accessing large-scale compute resources. It would provide visibility into where and by whom frontier models are being trained, including potentially adversarial actors. It could create economic incentives for compliance by restricting access to compute to certified AI developers from participating jurisdictions.[18]

From a technological standpoint, such an approach is feasible. There are a range of political challenges that go beyond the scope of this brief, but which would need to be assessed alongside the technical aspects.

HARDWARE-ENABLED VERIFICATION

An alternative or complementary approach would be to integrate verification mechanisms into AI computing hardware. Hardware-enabled mechanisms could verify certain properties of a frontier AI training run or deployment without requiring the developer or operator to share additional (potentially sensitive) information with the verifying parties. These mechanisms also would have the advantage of not relying on the operational capacity of an auditor to oversee deployments or uses of AI which will be increasingly large-scale and numerous. A hardware-enabled approach is particularly attractive in a context where it is hard to identify a third party trusted by all parties to perform the auditing function.

Hardware-enabled verification mechanisms would also overlap significantly with technological developments that are necessary to improve the security of frontier AI models. This is a key challenge, given the necessity of these improvements to prevent sophisticated attacks on frontier models, including physical and cyber-attacks.

Finally, given the context of a frontier competition, there is not yet a consensus on the technological feasibility of hardware-enabled verification mechanisms that would be useful for verification between state-level parties. Expert estimates on implementing such a verification approach range from tens of millions of dollars and two years to production at scale, at the optimistic end, to total impossibility at the pessimistic end.

WHAT COULD BE VERIFIED?

- **Compute capacities:** tracking the total number of AI chips in use, their energy consumption, and/or computational capabilities
- **Protocols:** tracking whether AI systems follow declared protocols
- **Data usage:** verifying data and training practices
- **Location:** tracking where AI capacities are being deployed
- **Application-specific verification:** verifying whether an AI system was being used in a particular industry (e.g. healthcare, finance, military)

CONSIDERATIONS

AI verification could serve important functions for leaders, including:

1. Detecting non-compliance as part of an agreement and/or its enforcement system.
2. Deterring parties from deliberate violations of agreements.
3. Building confidence by permitting parties to demonstrate their compliance in an open but secure manner.[19]

Specifically, an AI verification system could perform the following functions:

- Provide a common basis on which to develop standards, risk thresholds, and steps to verify AI capacities.
- Place AI verification within the technical domain, as something that could be developed across the chips that all major AI systems use.
- Increase transparency about AI capacities and tendencies, potentially reducing the risks of escalation or an "AI arms race."
- Some verification technologies could also be used to prevent violations of agreements.

However, there could be significant resistance to AI verification. In some cases, actors may see verification mechanisms as affecting the security of their systems, or an overly intrusive form of governance of their domestic matters. This so-called "transparency-security tradeoff" is most acute where basic information like location or compute capacities could inadvertently leak other strategically sensitive information. From a technological standpoint, such concerns could be addressed by considering a range of approaches to verification, including:

- **Confidential computing**—a hardware feature in the most recent chips—can enable some forms of verification already. In this case developers and evaluators would agree to run mutually reviewable code on mutually private inputs (a frontier AI model and an evaluation set) and learning nothing about each other's private inputs except the code's final results. However, this confidentiality is not likely to be secure against all state-level capabilities.
- **Retrofitted hardware mechanisms**, such as tamper-detecting power meters and interconnected bandwidth monitors that can remotely attest to the measurements that they are taking (chip-level energy usage and communication patterns over time), and to their uninterrupted integrity. A key factor for this mechanism would be that it cannot attest to internal information about the computation being performed.
- **Next-generation hardware mechanisms**, like integrating a verification mechanism into the data and code pathways in such a way that any computational property can be attested to, or even enforced, without violating confidentiality or integrity properties (perhaps even against state-level physical tampering attempts).
- **Neutral data centers**, jointly built and physically secured by parties, could provide trustworthy environments for high-sensitivity computation, enabling verification without compromising national security secrets.[20]

As such, while there are many political and technical challenges, some combination of approaches holds significant promise at addressing one of the major risk factors associated with frontier AI development.

REFERENCES

1 This brief represents the views of the independent scientists on the Scientific Advisory Board. It does not necessarily reflect the UN's position or those of network institutions. Mention of a commercial company or product in this document does not imply endorsement by the UN or the authors. The use of information from this document for publicity or advertising is not permitted. Trademark names and symbols are used in an editorial fashion with no intention on infringement of trademark or copyright laws. The Board is especially grateful to David Dalrymple and Nora Ammann for reviewing this brief, and to Rajeev Alur, Mauricio Baker, Miles Brundage, Maral Dadvar, David Dalrymple, Kapal Dev, Carina Geldhauser, Benjamin Harack, Sébastien Marcel, Neema Mduma, Rita Orji, Anderson Rocha, Stuart J Russell, Aaron Scher, and Robert Trager for participating in the expert roundtable that informed this work.

2 The definition of "frontier AI" is drawn from: UK DSIT. "Capabilities and risks from frontier AI: A discussion paper on the need for further research into AI risk," Technical report, DSIT UK, October 2023. URL https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf. See also Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., ... & Zilberman, N. (2024). Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation. arXiv preprint arXiv:2403.08501.

3 Center for New American Security, "US-China Competition and Military AI," July 2023, available at: https://www.cnas.org/publications/reports/u-s-china-competition-and-military-ai.

4 See, e.g.,Simmons-Edler, Riley, et al. "AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research." arXiv preprint arXiv:2405.01859, 3 May 2024; Rivera, Juan-Pablo, et al. "Escalation Risks from Language Models in Military and Diplomatic Decision-Making." arXiv preprint arXiv:2401.03408, 7 Jan. 2024; Katzke, Corin, and Gideon Futerman. "The Manhattan Trap: Why a Race to Artificial Superintelligence is Self-Defeating." arXiv preprint arXiv:2501.14749, 22 Dec. 2024; Miotti, Andrea, and Akash Wasil. "An International Treaty to Implement a Global Compute Cap for Advanced Artificial Intelligence." arXiv preprint arXiv:2311.10748, 1 Nov. 2023.

5 Anderljung, M., Smith, E. T., O'Brien, J., Soder, L., Bucknall, B., Bluemke, E., Schuett, J., Trager, R., Strahm, L., and Chowdhury, R. "Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework," Technical report, Centre for the Governance of AI, November 2023b. URL https://www.governance.ai/research-paper/towards-publiclyaccountable- frontier-llms. arXiv:2311.14711 [cs]. See also, Farrell, H. and Newman, A. Underground Empire: How America Weaponized the World Economy. Henry Holt and Co., September 2023. ISBN 978-1-250-84055-4. URL https://us.macmillan.com/books/9781250840554/undergroundempire.

6 Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213. See also, Choi, D., Shavit, Y., & Duvenaud, D. K. (2024). Tools for Verifying Neural Models' Training Data. Advances in Neural Information Processing Systems, 36.

7 For more on regulatory progress, see, Future of Life Institute. EU Artificial Intelligence Act: The Act Texts, 2024. URL https://artificialintelligenceact.eu/the-act/; Jiang, B. and Cao, A. China to create and implement national standard for large language models in move to regulate AI, while using its power to transform industries. South China Morning Post, July 2023: https://www.scmp.com/tech/policy/article/3226942/china-create-and-implementnational-standard-large-language-models-move-regulate-ai-while-using-its; Sheehan, M. "Tracing the Roots of China's AI Regulations," Technical report, Carnegie Endowment for International Peace, February 2024. https://carnegieendowment.org/2024/02/27/tracingroots-of-china-s-ai-regulations-pub-91815; Hacker, P. The European AI liability directives – Critique of a half-hearted approach and lessons for the future. Computer Law & Security Review, 51:105871, November 2023. ISSN 0267-3649. doi: 10.1016/j.clsr.2023.105871. URL https://www.sciencedirect.com/science/article/pii/S026736492300081X; Whittlestone, J., Avin, S., Heim, L., Anderljung, M., and Sastry, G. Response to the UK's Future of Compute Review. Technical report, Centre for the Governance of AI, March 2023. URL https://www.governance. ai/research-paper/response-to-the-uks-future-of-compute-review.

8 See, e.g., https://english.www.gov.cn/news/202307/13/content_WS64aff5b3c6d0868f4e8ddc01.html.

9 See, e.g. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213. Kolt, N., Anderljung, M., Barnhart, J., Brass, A., Esvelt, K., Hadfield, G. K., Heim, L., Rodriguez, M., Sandbrink, J. B., & Woodside, T. (2024). Responsible

Reporting for Frontier AI Development. arXiv preprint arXiv:2404.02675. Retrieved from https://arxiv.org/abs/2404.02675.

10 See, Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., ... & Coyle, D. (2024). Computing Power and the Governance of Artificial Intelligence. arXiv preprint arXiv:2402.08797; Besiroglu, Tamay, Sage Andrus Bergerson, et al. (Jan. 2024). The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny? URL: https://arxiv.org/abs/2401.02452v2 (visited on 01/12/2024); 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.

11 See, Epoch (2022). Parameter, Compute and Data Trends in Machine Learning. URL: https://do cs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit?usp=embed_facebook (visited on 01/13/2024). – (Apr. 2023). Key Trends and Figures in Machine Learning. URL: https://epochai.org/trends (visited on 01/13/2024).

12 See, Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute Trends Across Three Eras of Machine Learning. In 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, July 2022. doi: 10.1109/IJCNN55064.2022.9891914. URL http://arxiv.org/abs/2202.05924.arXiv:2202.05924.

13 Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., ... & Zilberman, N. (2024). Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation. arXiv preprint arXiv:2403.08501; see also, Abbott, K. W., Levi-Faur, D., and Snidal, D. (eds.). Regulatory intermediaries in the age of governance, volume 670 of The Annals of the American Academy of Political and Social Science. SAGE, March 2017a. ISBN978-1-5063-9011-6. URL https://www.jstor.org/stable/i26361533.; Financial Action Task Force. International Standards on Combating Money Laundering and the Financing of Terrorism and Proliferation. Technical report, FATF, Paris, France, November 2023. URL https://www.fatf-gafi.org/content/dam/fatf-gafi/recommendations/FATF%20Recommendations%202012.pdf.coredownload.inline.pdf; European Union. Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA, July 2011. URL https://eur-lex.europa.eu/eli/dir/2011/93/oj.

14 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.

15 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/.

16 For a summary of these issues, see https://blog.heim.xyz/considerations-and-limitations-for-ai-hardware-enabled-mechanisms/.

17 E.g., >10^26 FLOP)

18 NB: this recommendation is being developed by the Oxford Martin AI Initiative. See, https://www.oxfordmartin.ox.ac.uk/ai-governance.

19 These points from Harack, Ben, et al. Verification for International AI Governance. Draft report, 2025.

20 See, Sastry, Gillian Hadfield, et al. "Frontier AI Regulation: Managing Emerging Capabilities." Center for AI Governance, 2024; see also, Harack, Ben, et al. Verification for International AI Governance. Draft report, 2025.