Technical Requirements for Halting Dangerous AI Activities

Peter Barnett [1] **Aaron Scher** [1] **David Abecassis** [1]

Abstract

The rapid development of AI systems poses unprecedented risks, including loss of control, misuse, geopolitical instability, and concentration of power. To navigate these risks and avoid worst-case outcomes, governments may proactively establish the capability for a coordinated halt on dangerous AI development and deployment. In this paper, we outline key technical interventions that could allow for a coordinated halt on dangerous AI activities. We discuss how these interventions may contribute to restricting various dangerous AI activities, and show how these interventions can form the technical foundation for potential AI governance plans.

1. Introduction

Humanity appears to be on a trajectory toward developing AI systems that substantially outperform human experts across many cognitive domains. While potentially beneficial, the development of advanced AI carries profound risks (Center for AI Safety, 2023; Bengio et al., 2024). These risks are exacerbated if the transition to advanced AI is very fast, as the world would have less time to prepare. A key driver may be AI systems that are capable of automating AI R&D, leading to a feedback loop that rapidly increases AI capabilities.

To avoid these risks and have time to prepare, governments may wish to coordinate to halt or restrict dangerous AI development and deployment (Scholefield et al., 2025). Such a halt would address the following key risks:

- **Loss of Control:** AIs pursuing unintended goals could disempower humanity, potentially leading to extinction.
- **Misuse:** Malicious actors could use powerful AI for catastrophic ends (Zelikow et al., 2024), e.g., designing novel bioweapons (Crawford et al., 2024) or orchestrating large-scale cyberattacks on critical infrastructure.
- **Geopolitical Instability and War:** An AI arms race or the potential emergence of a strategically dominant AI could trigger conflicts.
- **Concentration of Power:** AI could lead to an extreme concentration of power, potentially enabling authoritarian lock-in through unprecedented surveillance and control (Davidson et al., 2025).

Beyond mitigating these risks, the design of a halt must also consider political viability, the preservation of AI's benefits, and reversibility once safety is assured.

2. Halting or Restricting AI Activities

The main goal of a halt is to stop the advancement of AI capabilities globally, but the mechanisms used in a halt will also be useful for verifiably restricting or halting specific, dangerous AI activities. Based on the current trajectory of AI development, we believe the main intervention point for these mechanisms is *AI compute* (Sastry et al., 2024). Plans generally revolve around limiting access to large quantities of advanced AI chips or monitoring the use of these chips, including by shutting them off or monitoring workloads (Scher & Thiergart, 2024). Other than compute, there are various approaches to *monitoring AI activities*, such as through mandatory reporting requirements (Belfield, 2024), auditing, and espionage. Another key function of halt mechanisms is *limiting proliferation* of dangerous AI capabilities. Besides compute, AI progress is largely driven algorithmic advances (Ho et al., 2024). Depending on the state of AI capabilities, it may also be necessary to *restrict algorithmic progress*.

Authorities will require different capacities, depending on the risk landscape. These capacities may include: **restricting training** (e.g., via compute thresholds), **restricting inference** (e.g., preventing misuse or uncontrolled automated AI R&D), and **restricting post-training**. These three capacities are useful for understanding which part of AI development a particular intervention helps address. For instance, if dangerous AI systems have already been trained, governance would need to focus on inference, and training-specific interventions would be less relevant. Capacities help organize the space of interventions, but our primary analysis concerns AI governance plans (Section 4 and which interventions are useful or necessary for those plans. In

[1]Machine Intelligence Research Institute, CA, USA. Correspondence to: Peter Barnett <peter@intelligence.org>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

this paper, we primarily discuss non-emergency technical interventions; see Appendix A for emergency shutdown interventions.

3. Technical Interventions

Halting or restricting AI activities requires a suite of interconnected technical capabilities, likely including some subset of the following interventions.

3.1. Chip location

Tracking the location of AI hardware is essential for many compute governance approaches; authorities must know the location of AI hardware to inspect it or shut it down.

- **Track chip shipments via manufacturers/distributors**: Engage suppliers to track AI chip distribution and deployment locations, covering both historical, ongoing, and future shipments.
- **Hardware-enabled location tracking**: Use secure hardware features (e.g., FlexHEGs (Petrie et al., 2024)) for chips to verify their location and status remotely (Brass & Aarne, 2024; Aarne et al., 2024).
- **Centralize compute in declared datacenters**: Consolidate AI compute in a small number of secure, registered datacenters to simplify tracking and control.
- **Periodic datacenter inspections**: Conduct periodic physical audits of datacenters to verify chip inventories.
- **Ongoing datacenter monitoring**: Establish continuous governmental oversight of datacenters, which may leverage inspectors or tamper-resistant surveillance cameras.

3.2. Chip manufacturing

Interventions targeting semiconductor manufacturing offer leverage at an earlier stage of AI development. Preventing certain actors from gaining access to hardware could effectively block them from doing dangerous AI activities.

- **Monitoring for the construction of new fabs**: Surveil industrial development to detect the construction of new advanced AI chip fabrication plants (Wasil et al., 2024).
- **Restrict/control the equipment and materials**: Restrict access to critical equipment (e.g., EUV machines) and material inputs (e.g., silicon wafers) for advanced chip production (Thadani & Allen, 2023).
- **Surveillance and inspections of fabs**: Conduct surveillance and on-site inspections of semiconductor fabs to verify compliance with agreements or restrictions. For example, only producing a maximum number of chips or only manufacturing authorized chip designs.
- **Verifiably shut down fabs**: Deactivate specific chip fabs, verifiable via satellite imagery, electricity monitoring, etc.
- **Verify that new chips have HEMs to spec**: Mandate and verify that new AI chips have hardware-enabled governance mechanisms (HEMs) (Kulp et al., 2024) (e.g., FlexHEGs) (Petrie et al., 2024).

3.3. Compute monitoring

Beyond tracking hardware, controlling its actual use is paramount. This involves establishing thresholds and implementing mechanisms to monitor and regulate how AI compute resources are employed.

- **Define capabilities thresholds**: Define measurable AI capability levels (e.g., autonomous replication) that trigger specific required responses (e.g., halt development, implement security measures) (Karnofsky, 2024; Koessler et al., 2024; Shevlane et al., 2023; Raman et al., 2025).
- **Define compute thresholds**: Establish limits on compute used for training models (Heim & Koessler, 2024).
- **Determine which datacenters need monitoring**: Establish criteria for datacenters that require oversight based on their compute capacity.
- **Monitor or restrict at the datacenter level**: Implement controls directly at datacenters, such as chip kill-switches, interconnect limits, workload classification based on external measurements (Scher & Thiergart, 2024).
- **Monitor or restrict via hardware-enabled mechanisms**: Leverage specialized hardware features to enforce usage policies or monitor AI activities (Kulp et al., 2024; Aarne et al., 2024; Petrie et al., 2024).
- **Monitor or restrict via software**: Employ software-based tools for monitoring AI tasks, analyzing code executing on compute, or restricting certain operations (Scher & Thiergart, 2024). This may be easier to subvert than other methods.
- **Monitor or restrict consumer compute**: Limit sales of consumer compute (e.g., personal computers, gaming GPUs, etc), and potentially mandate monitoring for certain hardware.
- **Inference-only hardware**: Develop AI hardware architecturally limited to inference (i.e., incapable of efficient model training) to prevent prohibited development.

3.4. Non-compute monitoring

Monitoring AI models and AI development projects, rather than their compute resources. These interventions would provide increased transparency into AI development.

- **Required reporting of capabilities**: Implement mandatory disclosure of model capabilities (Belfield, 2024).
- **Third-party/government evaluations of AI models**: Establish processes for independent AI evaluations by third-parties (Casper et al., 2024).
- **Other inspections for safety/security practices**: Conduct broader inspections of AI development facilities focusing on safety and security practices (Stix et al., 2025).
- **In-house auditors**: Mandate AI developers have embed-

ded audit teams for capability assessment and compliance, similar to "compliance monitors" in the finance industry.
- **Automated auditors**: Use AI systems to monitor AI development. These automated auditors could catch prohibited activities while preserving privacy.
- **Espionage**: Governments could use existing espionage tactics to gather information on AI development.
- **Whistleblowers**: Establish protected channels for insiders to report concerns about dangerous AI development.

3.5. Limiting proliferation

If powerful AI model weights or critical algorithmic insights are made available to malicious actors or the public, preventing their use becomes very challenging (Seger et al., 2023). Authorities could prevent the proliferation of models by preventing them from being trained in the first place.

- **Model weight security**: Implement and verify security to protect AI model weights from unauthorized access or leakage (Nevo et al., 2024).
- **Algorithmic secret security**: Protect algorithmic insights from proliferation.
- **Structured access**: Rather than providing broad access to potentially dangerous AI systems, provide controlled access to a limited set of users (Shevlane, 2022).
- **Limit open model release**: Implement policies to restrict the broad release of AI models with high-risk capabilities.
- **Hardware-specific models**: Develop AI models which can only run on specific hardware (Clifford et al., 2024). If model weights leak, the model cannot easily be run without the required hardware.
- **Non-fine-tunable models**: Create methods for training models that cannot be fine-tuned to unlock new dangerous capabilities (Deng et al., 2024; Tamirisa et al., 2024).

3.6. Research

Unmonitored compute would pose risks if algorithmic progress continues (Ho et al., 2024). Oversight could help prohibit dangerous research while enabling beneficial work.

- **Tracking personnel**: Track activities of key researchers. This could include tracking their location and affiliation, to ensure they are not part of a secret AI development project.
- **Define and prohibit algorithmic progress research**: Identify and ban specific lines of algorithmic research deemed too dangerous or destabilizing.
- **Surveillance of AI research activities**: Monitor the computers and research activities of AI researchers to ensure compliance with research guidelines.
- **Model spec prohibiting algorithmic research**: Include prohibitions on accelerating AI research in AI model specifications (OpenAI, 2025; Scher & Thiergart, 2024).

4. Plans

These technical interventions form the basis for various proposed AI governance plans, summarized in this section. Table 1 shows the technical requirements for each plan.

Last-minute Wake-up World leaders may not take substantial steps to halt AI development until AI capabilities are very advanced and substantial harm has occurred. The Last-minute Wake-up plan attempts to halt AI development in that world, first responding to an acute emergency (Appendix A) and then implementing a global compute monitoring regime. Depending on progress in the development of verification mechanisms and methods for predicting risk, effective compute monitoring regimes differ substantially in their details. For instance, compute monitoring could involve completely shutting off chips or permitting some verified workloads (Scher & Thiergart, 2024). Verification could allow society to safely benefit from AI systems. This regime would primarily prevent the advancement of AI capabilities via strict limits on the amount of compute used for AI training. It may also need to monitor or restrict the post-training and inference of existing AI models, if they pose a risk. Hopefully, consumer computers would be exempt from this compute monitoring regime, unless sufficiently advanced AI models have proliferated or the compute requirements for dangerous AI activities are very low (e.g., due to substantial algorithmic progress). Therefore, this plan also involves slowing the rate of algorithmic progress by banning relevant research and avoiding the proliferation of the most capable models via internal security measures and limitations on model release.

Chip Production Moratorium The compute hardware required to enable dangerous AI may not yet have been built. The amount of AI-relevant compute is expected to increase tenfold in the next two years (Dean, 2025), and hundredfold by 2030 (Pilz et al., 2025; Sevilla et al., 2024). Progress in the near-term is likely to rely on continued scaling of this input, production of which is concentrated in dozens of highly advanced chip fabs (Sastry et al., 2024).

This moratorium aims to globally pause the production of new AI compute. This is made feasible due to the cost, size, and fragility of semiconductor production. It avoids difficulties common to AI verification by requiring only that new cutting-edge chips are not produced, which rival states can independently verify with ease. As a last resort, it may even be implemented unilaterally, though in practice, the possibility of unilateral action may promote coordination between rivals. This plan's effectiveness depends on how soon it can be enacted. Governing existing compute and restricting algorithmic research are helpful extensions, but they may not be necessary to delay dangerous AI by decades.

A Narrow Path Miotti et al. (2024) proposes immediate national actions and an international treaty to prevent the development of artificial superintelligence (ASI) for two decades. It establishes a "defense in depth" strategy using regulatory oversight, including a licensing regime based on compute thresholds, prohibitions on dangerous capabilities like AI self-improvement, and mandatory safety justifications for AI systems.

Keep the Future Human Aguirre (2025) advocates preventing the development of smarter-than-human, autonomous, general-purpose AI—and ASI—by focusing on hardware-enabled compute governance to enforce limits on the compute used to create AI systems. It also manages risk through enhanced liability when AI systems feature autonomy, generality, and intelligence.

Superintelligence Strategy Hendrycks et al. (2025) proposes a national security framework based on deterrence via Mutual Assured AI Malfunction (MAIM). States use threats of sabotage to prevent each other from developing destabilizing AI capabilities. This is combined with nonproliferation efforts against rogue actors, such as hardware export control, algorithmic secret security, and model weight security.

5. Assessment of Plans and Requirements

In Table 1, we provide an overview of the cataloged interventions, and the relevance of each intervention to the key capacities (restricting training, inference, and post-training) and plans.

The plans in Section 4were highlighted because they, to some extent, involve halting dangerous AI activities. The prior plans (Miotti et al. (2024), Aguirre (2025), Hendrycks et al. (2025)) are some of the only plans for a halt described in sufficient detail to allow assessment. The novel plans (Last-minute Wake-up, and Chip Production Moratorium) are intended to substantially reduce catastrophic risk from advanced AI. The plans were also chosen to be diverse, covering different approaches to managing catastrophic AI risks.

We grade the technological readiness of each of the interventions as follows:

- **High**: No technological hurdles to implement this intervention.
- **Medium**: Significant effort needed for real-world implementation.
- **Low**: Lacks even functional prototypes, important aspects have not yet been defined.

For each intervention and each plan, we assess the extent to which the plan depends on the intervention. *Required* means that this intervention is essential to a given plan or capability. *Maybe required* means that this intervention may be needed; for example if there are multiple ways to fulfill a specific function (e.g., methods to oversee AI research) or if there is substantial uncertainty about whether the function will be needed under a plan (e.g., some interventions are only important if AI capabilities have already advanced to dangerous levels, which may or may not happen before halting). *Helpful* means that this intervention is useful for some functions of a plan, but not essential.

These assessments are preliminary best estimates and subject to refinement. For previously published plans, assessments are based on identifying the functions that plans aim to accomplish, whether the plan explicitly mentions the intervention, and whether we think the plan implicitly relies on the intervention. For plans first discussed here, we reach this assessment by identifying the main functions the plans rely on and key interventions to perform these functions.

6. Conclusion

Our analysis of technical requirements for halting dangerous AI activities can improve understanding and prioritization of AI governance work. All of these plans requires substantial control over AI compute through chip tracking, datacenter monitoring, or manufacturing restrictions. For most plans, limiting model access via security measures and restricting open release is essential, as proliferation could make control of models impossible. Most plans also rely on interventions currently at low technological readiness.

These findings highlight urgent priorities. The required infrastructure and technology must be developed before it is needed, such as hardware-enabled mechanisms. International tracking of AI hardware should begin soon, as this is crucial for many plans and will only become more difficult if delayed. Without significant effort now, it will be difficult to halt in the future, even if there is will to do so.

Table 1: Technical interventions for halting or restricting dangerous AI activities, and the capacities and plans they are required for.

Legend:
● = Required
⊙ = Maybe required, depending on the situation
○ = Helpful, but not required

Intervention	Technological Readiness	Capacities: Restrict Training	Capacities: Restrict Inference	Capacities: Restrict Post-training	Plans: Last-minute Wake-up	Plans: Chip Production Moratorium	Plans: A Narrow Path*	Plans: Keep the Future Human*	Plans: Superintelligence Strategy*
Chip location									
Track chip shipments via manufacturers/distributors	High	⊙	○	○	●	⊙	●	●	●
Hardware-enabled location tracking	Low	○	○	○	○	⊙	⊙	○	⊙
Centralize compute in declared datacenters	Medium	○	○	○	⊙	⊙	○	○	○
Periodic datacenter inspections	High	⊙	⊙		●	○	⊙	○	○
Ongoing datacenter monitoring	High	⊙	⊙	○	●	○	○	○	●
Chip manufacturing									
Monitoring for the construction of new fabs	High	⊙			●	●	○	●	○
Restrict/control the equipment and materials	Medium	⊙			○	⊙	○	○	○
Surveillance and inspections of fabs	Medium	○			●	●		○	○
Verifiably shut down fabs	High	⊙				●		○	○
Verify that new chips have HEMs to spec	Low	○	⊙	⊙	○		○	●	○
Compute monitoring									
Define capabilities thresholds	Medium	⊙	●	●	●	⊙	●	●	●
Define compute thresholds	High	⊙		○	●	⊙	●	●	○
Determine which datacenters need monitoring	Medium	●	⊙		●	⊙	●	●	●
Monitor or restrict at the datacenter level	High	⊙	⊙	○	●		●	⊙	○
Monitor or restrict via hardware-enabled mechanisms	Low	⊙	⊙	○	○		○	●	○
Monitor or restrict via software	Low	○	⊙	○	○		○	○	○
Inference-only hardware	Medium	⊙		⊙	○	○		○	○
Monitor or restrict consumer compute	Low	⊙	⊙	⊙	⊙				
Non-compute monitoring									
Required reporting of capabilities	High		●	●	●		●	●	●
Third-party/government evaluations of AI models	High	⊙	⊙	○	⊙		●	●	●
Other inspections for safety/security practices	Medium	○			⊙		●	●	○
In-house auditors	Medium	⊙	⊙	⊙	⊙		○	○	○
Automated auditors	Low	⊙	○	⊙	⊙			○	○
Espionage	High	⊙	○	⊙	●	○			●
Whistleblowers	High	⊙	⊙	⊙	○	○	○	○	○
Limiting proliferation									
Model weight security	Medium	⊙	●	●	●		●	●	●
Algorithmic secret security	Medium	⊙		○	⊙	○	○	○	●
Structured access	High	⊙	⊙	⊙	●		●	○	●
Limit open model release	High	○	●	●	●		●	●	●
Hardware-specific models	Low	○	⊙	○	○			○	○
Non-fine-tunable models	Low			○	○		○	○	○
Research									
Tracking personnel	High	○	○		⊙	○			⊙
Define and prohibit algorithmic progress research	Medium	○	⊙		⊙	⊙	●	⊙	●
Surveillance of AI research activities	Medium	○	⊙	○	⊙	⊙	●		⊙
Model spec prohibiting algorithmic research	High	○	⊙		⊙	○	○	⊙	○

*This appraisal of the technical requirements for these plans is not necessarily endorsed by their original authors.

Impact Statement

We try to advance the understanding of technical requirements for governing potentially dangerous AI systems. The development of capabilities for halting or restricting dangerous AI activities, while intended to mitigate catastrophic risks, could itself have complex societal consequences.

These measures, targeted at reducing catastrophic risk, represent significant expansion of international governmental authority over AI development. Regulating dangerous AI after it is developed and deployed will require much more invasive measures, and may not be possible at all. Therefore, targeted oversight of AI activities and key inputs could serve to minimize future government intrusion while averting large-scale harm.

We acknowledge that restricting certain AI activities would necessarily slow the realization of some technological benefits. However, we believe that many advantages of narrow AI systems remain accessible under these regimes, while avoiding the most destabilizing and dangerous effects of unrestricted development. The optimal approach prioritizes careful progress rather than unrestricted advancement, allowing research to move forward when it is safe.

References

Aarne, O., Fist, T., and Withers, C. Secure, governable chips. *Center for a New American Security. https://www.cnas. org/publications/reports/secure-governable-chips*, 2024.

Aguirre, A. Keep the Future Human: Why and How We Should Close the Gates to AGI and Superintelligence, and What We Should Build Instead. *arXiv preprint arXiv:2311.09452*, 2025. URL https://arxiv.org/abs/2311.09452v4.

Belfield, H. What Information Should Be Shared with Whom "Before and During Training"?, December 2024. URL http://arxiv.org/abs/2501.10379. arXiv:2501.10379 [cs].

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

Brass, A. and Aarne, O. Location verification for ai chips. Technical report, Institute for AI Policy and Strategy, April 2024. URL https://www.iaps.ai/research/location-verification-for-ai-chips. Available at: https://www.iaps.ai/research/location-verification-for-ai-chips.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, June 2024. doi: 10.1145/3630106.3659037. URL http://arxiv.org/abs/2401.14446. arXiv:2401.14446 [cs].

Center for AI Safety. Statement on AI Risk | CAIS, March 2023. URL https://www.safe.ai/work/statement-on-ai-risk.

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., and Anderljung, M. Infrastructure for ai agents. *arXiv preprint arXiv:2501.10114*, 2025.

Clifford, E., Saravanan, A., Langford, H., Zhang, C., Zhao, Y., Mullins, R., Shumailov, I., and Hayes, J. Locking machine learning models into hardware. *arXiv preprint arXiv:2405.20990*, 2024.

Cohen, S., Bitton, R., and Nassi, B. Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications, January 2025. URL http://arxiv.org/abs/2403.02817. arXiv:2403.02817 [cs] version: 2.

Crawford, F. W., Webster, K., Epstein, G. L., Roberts, D., Fair, J., and Nevo, S. *Securing Commercial Nucleic Acid Synthesis*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/RRA3329-1.

Davidson, T., Finnveden, L., and Hadshar, R. AI-Enabled Coups: How a Small Group Could Use AI to Seize Power. 2025. URL https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power. Accessed: 2025-04-17.

Dean, R. Compute forecast. AI 2027, April 2025. URL https://ai-2027.com/research/compute-forecast. Available at: https://ai-2027.com/research/compute-forecast.

Deng, J., Pang, S., Chen, Y., Xia, L., Bai, Y., Weng, H., and Xu, W. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2553–2571. IEEE, 2024.

Heim, L. and Koessler, L. Training Compute Thresholds: Features and Functions in AI Regulation, August 2024. URL http://arxiv.org/abs/2405.10799. arXiv:2405.10799 [cs].

Hendrycks, D., Schmidt, E., and Wang, A. Superintelligence Strategy: Expert Version, March 2025. URL http://arxiv.org/abs/2503.05628. arXiv:2503.05628 [cs].

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. Algorithmic progress in language models, March 2024. URL http://arxiv.org/abs/2403.05812. arXiv:2403.05812 [cs].

Karnofsky, H. A Sketch of Potential Tripwire Capabilities for AI. dec 2024.

Koessler, L., Schuett, J., and Anderljung, M. Risk thresholds for frontier ai. *arXiv preprint arXiv:2406.14713*, 2024.

Kulp, G., Gonzales, D., Smith, E., Heim, L., Puri, P., Vermeer, M. J. D., and Winkelman, Z. *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/WRA3056-1.

METR. The Rogue Replication Threat Model. https://metr.org/blog/2024-11-12-rogue-replication-threat-model/, 11 2024.

Miotti, A., Bilge, T., Kasten, D., and Newport, J. A Narrow Path, December 2024. URL https://www.narrowpath.co/.

Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., and Alstott, J. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. Technical report, RAND Corporation, May 2024. URL https://www.rand.org/pubs/research_reports/RRA2849-1.html.

OpenAI. OpenAI Model Spec, February 2025. URL https://model-spec.openai.com/2025-02-12.html.

Petrie, J., Aarne, O., Ammann, N., and Dalrymple, D. Interim report: Mechanisms for flexible hardware-enabled guarantees. Technical report, 8 2024.

Pilz, K. F., Sanders, J., Rahman, R., and Heim, L. Trends in ai supercomputers. *arXiv preprint arXiv:2504.16026*, 2025.

Raman, D., Madkour, N., Murphy, E. R., Jackson, K., and Newman, J. Intolerable risk threshold recommendations for artificial intelligence. *arXiv preprint arXiv:2503.05812*, February 2025.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., and Coyle, D. Computing Power and the Governance of Artificial Intelligence, February 2024. URL http://arxiv.org/abs/2402.08797. arXiv:2402.08797 [cs].

Scher, A. and Thiergart, L. Mechanisms to Verify International Agreements About AI Development, November 2024. URL https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development.

Scholefield, R., Martin, S., and Barten, O. International agreements on ai safety: Review and recommendations for a conditional ai safety treaty. *arXiv preprint arXiv:2503.18956*, 2025.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., hÉigeartaigh, S. Ó., Korinek, A., et al. Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*, 2023.

Sevilla, J., Besiroglu, T., Cottier, B., You, J., Roldán, E., Villalobos, P., and Erdil, E. Can AI Scaling Continue Through 2030?, 2024. URL https://epoch.ai/blog/can-ai-scaling-continue-through-2030. Accessed: 2025-03-29.

Shevlane, T. Structured access: an emerging paradigm for safe ai deployment. *arXiv preprint arXiv:2201.05159*, 2022.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks, September 2023. URL http://arxiv.org/abs/2305.15324. arXiv:2305.15324 [cs].

Shlegeris, B. AI catastrophes and rogue deployments. June 2024. URL https://www.alignmentforum.org/posts/ceBpLHJDdCt3xfEok/ai-catastrophes-and-rogue-deployments.

Stix, C., Pistillo, M., Sastry, G., Hobbhahn, M., Ortega, A., Balesni, M., Hallensleben, A., Goldowsky-Dill, N., and Sharkey, L. Ai behind closed doors: a primer on the governance of internal deployment. *arXiv preprint arXiv:2504.12170*, 2025.

Tamirisa, R., Bharathi, B., Phan, L., Zhou, A., Gatti, A., Suresh, T., Lin, M., Wang, J., Wang, R., Arel, R., et al. Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*, 2024.

Thadani, A. and Allen, G. C. Mapping the semiconductor supply chain. *Center for Strategic and International Studies*, 11, May 2023.

Wasil, A. R., Reed, T., Miller, J. W., and Barnett, P. Verification methods for international ai agreements. *arXiv preprint arXiv:2408.16074*, 2024.

Zelikow, P., Cuéllar, M.-F., Schmidt, E., and Matheny, J. Defense against the AI dark arts: Threat assessment and coalition defense. Report, Hoover Institution, Stanford, CA, 12 2024. A Publication of the Hoover Institution.

A. Emergency Shutdown

The body of this short paper has focused on technical interventions to halt or restrict dangerous AI activities in non-emergency situations. However, it may be crucial for authorities to have emergency shutdown capacities, capacities that allow them to rapidly shut down and limit the harm from dangerous AI systems. For example, this could include shutting down a datacenter running a rogue autonomous AI or an AI being used for a large-scale cyberattack by a malicious actor (METR, 2024; Shlegeris, 2024). An emergency may, unfortunately, be necessary for governments to become sufficiently concerned about dangerous AI, such that they consider plans to halt or restrict dangerous AI activities.

Various technical interventions previously mentioned in Section 3 would be useful in an emergency shutdown situation, specifically measures around chip location, compute use, and limiting proliferation. Here we provide an initial list of technical interventions specifically useful for emergency shutdown.

- **Track down rogue AIs**: Find out which physical hardware a rogue AI system is running on, in order to shut it off. An example would be tracking an AI system via its IP address. In an optimistic scenario, a rogue AI system will have only spread to a limited number of datacenters that are easily located and shut off. A more pessimistic scenario could involve many copies of an AI system running on consumer hardware or running on a distributed botnet. Tools like AI watermarking and AI text detection may help identify and locate rogue AI systems. Tracking down rogue AIs could be very difficult or effectively impossible.

- **AI spread reduction**: Prevent rogue AI systems spreading and accumulating resources (Cohen et al., 2025). This could include strong know-your-customer (KYC) or proof-of-humanity verification on services a rogue AI might use, such as cloud compute providers, online financial services, and online work platforms like Upwork (Chan et al., 2025). Other interventions include wide-scale preemptive red-teaming of cyber vulnerabilities or publicity campaigns for humans to not cooperate with rogue AI systems.

- **Last-resort shutdown measures**: If AI systems are contained to datacenters, shutting them down is relatively technically straightforward. It would be much more difficult to shut down an AI system that has spread widely across non-datacenter compute (such as personal computers). It is unclear if shutting down such a system is technically feasible, but approaches for further research might include shutting down parts of the electrical grid, EMP weapons to disable electronics in a wide area, and pervasive cyber weapons.

- **Circuit breakers on societal infrastructure**: AI systems may cause cascading failures in complex systems such as the stock market, power grid, or supply chains. Such systems should have fail-safe mechanisms and other components to mitigate such failures.