Centre for International Governance Innovation

Digital Policy Hub

Digital Policy Hub – Working Paper

Poor Cybersecurity at Frontier AI Labs Could Disincentivize Arms Racing

Wim Howson Creutzberg

Fall 2024 cohort

About the Hub

The Digital Policy Hub at CIGI is a collaborative space for emerging scholars and innovative thinkers from the social, natural and applied sciences. It provides opportunities for undergraduate and graduate students and post-doctoral and visiting fellows to share and develop research on the rapid evolution and governance of transformative technologies. The Hub is founded on transdisciplinary approaches that seek to increase understanding of the socio-economic and technological impacts of digitalization and improve the quality and relevance of related research. Core research areas include data, economy and society; artificial intelligence; outer space; digitalization, security and democracy; and the environment and natural resources.

The Digital Policy Hub working papers are the product of research related to the Hub's identified themes prepared by participants during their fellowship.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

Partners

Thank you to Mitacs for its partnership and support of Digital Policy Hub fellows through the Accelerate program. We would also like to acknowledge the many universities, governments and private sector partners for their involvement allowing CIGI to offer this holistic research environment.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Key Points

- An artificial intelligence (AI) arms race could lead to or exacerbate an arms race in cybertechnology, transforming the cyberwarfare landscape. This transformation could heighten the difficulty of securing frontier AI labs against cyberattacks.
- If states' frontier AI labs are mutually vulnerable to cyberattacks from their adversaries, states have novel incentives to coordinate on AI development under specific circumstances.
- Further incentives for states to coordinate would emerge if their labs were vulnerable to cyberattacks by both state and non-state actors. Such coordination would reduce arms race-driven international security risks from AI, although that risk might still be elevated in the case of bad non-state actors.
- There is a pressing need for further research on how cyberwarfare will shape states' incentives to buy in to international AI coordination regimes. Research in this vein should seek to identify and leverage cybersecurity vulnerabilities that could be used to increase the effectiveness of international coordination efforts.

Definition of Terms

Intellectual property for artificial intelligence systems (AIIP) refers to informational resources relevant to the development of frontier AI, such as model weights or insights about model architecture or algorithms. As such, this term is unrelated to any legal definition of intellectual property (IP).

Frontier AI lab is defined as an operation capable of developing frontier AI systems. For the purposes of this working paper, frontier labs are assets for states competing in an AI arms race. Thus, they include public and private entities but not multilateral international institutions.

To **coordinate** in the context of an AI arms race denotes a state's decision to co-create or abide by multilateral international institutions that regulate the development or use of advanced AI systems.

Bad actor-driven risk refers to the international security threat that derives from an increase in the capacity of bad non-state actors to bring about catastrophic outcomes (for instance, via the deployment of chemical or biological weapons).

Introduction

Background

International coordination on the development of AI systems would allow for better management of the collective risks and benefits that may arise as the technology advances. Such proposals are challenged by the prospect of an AI arms race, namely,

between the United States and China. Western AI governance researchers are split over the implications of this prospect.

For adherents of a broadly realist school of international relations, high stakes are insufficient for inducing coordination; as with all collective action problems, it must be not only collectively beneficial but also individually rational for international coordination to be feasible among the states racing to develop advanced AI.

It follows that even if advanced AI systems entail catastrophic risks, frontier AI companies should accelerate their research and development (R&D) efforts, not halt them. The reasoning on this point echoes the rationale for stockpiling nuclear weapons — if America does not build advanced AI, China will. An AI arms race to develop advanced AI is inevitable, the thinking goes, but losing to China is not. The greater the United States' lead in the race, the greater its affordance to implement appropriate safety measures and the greater its leverage in enforcing international safety regulations (Aschenbrenner 2024).

This view pervades US AI policy to the point of its being taken for granted (Zwetsloot, Toner and Ding 2024). The Biden administration passed bipartisan bills and semiconductor controls aiming to stifle China's AI competitiveness, while the Trump administration has framed the importance of major government appointments in terms of their relevance to helping "win the A.I. arms race with China" (Winter-Levy 2024). And in 2024, the U.S.-China Economic and Security Review Commission's (2024) report to Congress recommended foremost that "Congress establish and fund a Manhattan Project-like program dedicated to racing to and acquiring an Artificial General Intelligence capability."

However, other AI policy experts contest the premises of this "realist" view, such as its assumptions about the infeasibility of international coordination and the feasibility of mitigating catastrophic risks of AI. Since hawkish posturing on AI policy increases the difficulty of international coordination, clarifying the inevitability of an AI arms race is decisive for global AI strategy.

Overview

The state of cyberwarfare is of great relevance to AI development, yet its relationship to AI arms race dynamics has received little consideration in the literature. This working paper takes steps toward filling this gap. First, it argues that extrapolation from current trends suggests an AI arms race would set off or exacerbate a cybertechnology arms race, transforming the nature of cyberwarfare. Second, it considers two ways that this transformation could unfold by outlining a toy model of the strategic environment in which states navigate the AI arms race, showing that in both cases, de-escalating the AI arms race becomes more incentive compatible for both parties.

In the first case, the cyberwarfare landscape is such that frontier AI labs are inherently vulnerable to cyberattack from their direct competitor in the AI race. In the second case, the cyberwarfare landscape is such that frontier AI labs are vulnerable to cyberattacks from both their direct competitor in the AI race and from non-state actors. In both of these scenarios, states making individually rational decisions does not inevitably or indefinitely escalate the AI arms race; the configuration of the cybersecurity landscape

creates possibilities for the coordinated development of AI systems among states to be individually rational. This finding presents a starting point for charting incentive-compatible paths to escaping an AI arms race and raises the possibility that the vulnerability of AI labs to cyberattack could, in some circumstances, decrease overall international security risks from AI.

An AI Arms Race Would Escalate and Transform Cyberwarfare

Escalation

Cyber capabilities are integral to a state's military capacity and national security. Cyberattacks target adversaries' critical infrastructure during wartime and have proven effective in stealing their IP for military technology, despite adversaries' strong incentives to prevent these outcomes. Already, there are concerns that China has obtained AIIP from frontier AI labs (Aschenbrenner 2024). China hawks who view AI as crucial for military and economic dominance see fortifying cybersecurity at American frontier AI labs as an immediate priority (ibid.). Investing in cyber capabilities, then, would be a means to improving states' competitiveness in an AI arms race.

Amid an AI arms race, investing in offensive cyber capabilities would strengthen states' competitiveness in AI. Moreover, using cyberattacks to obtain AIIP from an adversarial state would likely be a cost-effective means of advancing AI capabilities compared to developing AIIP via domestic R&D.

AI R&D is expensive. The current paradigm in AI holds that the most promising way to improve a model's outputs is to significantly scale up its inputs, namely, the model size and training data, which, in turn, requires scaling up the physical infrastructure used to train and run the AI systems. Indeed, the infrastructure used to train these systems has grown by a factor of four to five times per year since 2010 (Sevilla and Roldán 2024). The upfront financial cost of scaling up this physical infrastructure — building and powering large data centres — rises in step. As of 2024, back-of-the-envelope calculations suggest that developing a frontier AI system from the ground up would cost US$100 billion and would reach US$1 trillion by the end of the decade if this scaling up were to continue at the current rate (ibid.; Gordon 2024).

In contrast, cyber R&D costs are not projected to increase at this rate by default and cyberattacks aimed at obtaining AIIP already require a mere fraction of the cost of obtaining the same AIIP through R&D. As of 2023–2024, it is estimated that the cost of developing frontier AI was roughly US$100 million, while frontier AI labs with budgets around US$1 million are vulnerable to cyberattacks. If stealing AIIP via cyberattacks became a top priority for China and the United States in the next five years, this approach would nonetheless cost several orders of magnitude less than developing frontier AI systems. Currently, the highest-priority operations of the few most cyber-capable institutions have budgets of up to US$1 billion — many times smaller than AI infrastructure costs (Nevo et al. 2024). Racing states would therefore seek to scale up investment in cyber offensive and defensive capabilities, which would trigger or exacerbate a cyber arms race.

Transformation

Such a scale up in investment would change the nature of cyberwarfare. As with military conflict in general, scaling up investment in cyberwarfare, including through innovation, would change the offence-defence balance, which is the cost of defensive operations relative to offensive operations (Garfinkel and Dafoe 2021). To date, the cyber offence-defence balance has generally been viewed as offence dominant: it is more cost-effective to operationalize a cyberattack than to mount an equivalent defence (Slayton 2017). It is unclear how the offence-defence balance will scale as cyberwarfare advances (Bonfanti 2022). This uncertainty calls for the consideration of a wide range of possibilities, such as the fact that cyber offence dominance only increases with scale, leaving the United States and China mutually vulnerable to cyberattack.

AI is expected to continue changing the cyberwarfare landscape, which would be exacerbated by an AI arms race, by lowering the barrier to entry for launching cyberattacks of all kinds. For instance, AI could be used to automate cyber capabilities, making cyberattacks that are otherwise labour-intensive newly accessible; it may also introduce new kinds of cyberattacks on par with those of a cybercrime syndicate or an insider threat with privileged access (Brundage et al. 2018). In addition, AI could increase the anonymity of cyberattackers, thus enhancing the difficulty that states face in identifying cyberattackers (ibid.). These changes could disproportionately benefit resource-constrained actors, increasing frontier labs' vulnerability to violent non-state actors (VNAs).

This increased anonymity would further undermine the effectiveness of deterring VNAs that are already difficult to deter (ibid.; Shamir 2021). Moreover, access to AIIP could significantly increase bad actor-driven risk. Advanced AI systems could lower the barriers to entry for creating chemical and biological weapons of mass destruction for actors at every level of capacity (US Department of Homeland Security 2024).

Transformed Cyberwarfare Could De-escalate an AI Arms Race

Mutual Vulnerability to Cyberattack Incentivizes Coordination on AI

A cyberwarfare landscape transformed by an AI arms race could, in return, transform AI arms race dynamics. Suppose the United States and China were locked in an AI arms race that exacerbated a cyber arms race, and both states were maximally scaling up investment in their cyber capabilities, making roughly equal investments. As investment in the cyber arms race scales up, cyber offence dominance only increases. As a result, each state's cyber-offensive capabilities could overwhelm the other's defensive capabilities, leaving both states mutually vulnerable to cyberattacks on their frontier AI labs.

Both the United States and China are deciding whether to escalate the arms race by developing a next-generation frontier AI system or to coordinate with their adversary in developing such a system and splitting the development costs. Assume, for simplicity's

sake, that frontier AI lab cybersecurity cannot be overcome by any actors other than the United States and China. The national security benefit of escalation is, in large part, a matter of how much time it takes the adversary to obtain an equivalent AI system. This period of time decreases as a function of the extent of cyber offence dominance. As the amount of time it takes for the adversary to obtain the AIIP via cyberattack decreases, so does the overall national security benefit of escalation as well as the relative importance of leading in AI development to win the arms race. In this case, leading in AI development is neither necessary nor sufficient for winning the AI arms race. The importance of leading in AI development decreases relative to other variables, such as the speed of diffusion (that is, how quickly a country can integrate AI systems into its military and economy), as well as the speed of "takeoff" (that is, the extent to which AI progress is accelerated by a tight, positive feedback loop). The speed with which the United States or China can reap the national security benefits of frontier AI systems is orthogonal to whether either country chooses to escalate. If takeoff is fast, escalating first poses a greater advantage, but takeoff speed is difficult to predict. The implications of these variables — diffusion and takeoff rates — otherwise lie outside the scope of this paper.

With each escalation, the benefit to national security of a further escalation decreases, while the financial cost of a further escalation increases. Recall that, at the current rate, the total cost of training advanced AI systems increases by a factor of four to five times per year and is expected to be on the order of hundreds of billions of dollars or more by the end of the decade — a projected cost hundreds of times greater than the adversary's cost in stealing the AIIP.[1] In the limit, there is an extent of cyber offence dominance beyond which the dominant strategy for both states is to coordinate, and if either state escalates, their adversary can essentially "free ride" off the other's investment. As such, the choice to escalate the AI arms race becomes unviable as a national security strategy and an inefficient use of state resources, and deteriorates as a bulwark for national security and as a competitive allocation of state resources.

Vulnerability to Non-State Actors Further Incentivizes Coordination

The possibility that under-resourced actors could disproportionately benefit from transformations of the cyber landscape, such as the integration of AI, introduces an additional incentive for coordination in an AI arms race.

Suppose that, in the previous case, the transformed cyberwarfare landscape is characterized by a lower barrier to entry. As such, the cybersecurity of frontier AI labs can be breached by VNAs in addition to the United States and China. For both states, the decision to develop a frontier AI system must take into account the following considerations. If either state develops a next-generation AI system, bad actor-driven risk from AI increases. If the other state obtains an equivalent next-generation AI system by stealing the adversary's AIIP or domestic R&D, the surface area of attack available to VNAs further increases bad actor-driven risk. For it to be individually rational for

1 This is not a one-to-one comparison, as it assumes the cost of launching a high-level cyber operation in this scenario is equal to its current cost. The cyber arms race dynamic on which this scenario is predicated could increase the cost of this kind of operation. Nonetheless, the argument put forward in this section stands so long as the cost of high-priority cyber operations increases at less than four to five times per year.

states to cooperate, the national security threat of the original increase in bad actor-driven risk must exceed the threat of the further increase in bad actor-driven risk combined with the threat of the adversarial state having an equivalent AI system. Such circumstances are possible, though not inevitable, and their emergence would depend on other variables such as the offence-defence balance of the military capabilities of the AI systems in question, the intensity of the arms race and the kinds of offensive military capabilities enabled by the frontier AI systems. Modelling these variables and their implications for states' decisions in an arms race lies outside the scope of this paper. Nonetheless, the national security threat of bad actor-driven risk could make coordinating in an AI arms race incentive compatible.

Conclusion

This paper argues that, in particular cases, the vulnerability of frontier AI labs to cyberattacks could undermine the incentives for states to escalate an AI arms race, leading states to coordinate in developing advanced AI systems. This finding serves as a starting point for identifying how the configuration of the cybersecurity landscape could influence the decisions of states competing in an AI arms race, viewed from a realist perspective. That said, the robustness of this finding is unclear. The analysis does not comprehensively take into account several important variables such as the offence-defence balance of AI, expectations about the takeoff speed of AI or the perceived intentions of adversaries. Future research building off this work could seek to construct formal models of these arguments, or give a more extensive account of the aforementioned variables or forecast how the AI and cyber offence-defence balances will scale. It would also be useful to clarify whether the dynamics identified in this paper apply to methods of stealing AIIP beyond cyberattacks.

Recommendation

- **Establish a federal research program to identify and leverage cyberwarfare advantages to disincentivize AI arms racing**. The advancement of offensive cyber capabilities could de-escalate a US-China AI arms race under some conditions. Research on identifying and leveraging cyberwarfare advantages in order to enforce multilateral AI governance could soon be crucial for mitigating threats to national security. Canada's Communications Security Establishment should collaborate with the Canadian Institute for Advanced Research to establish a research program focused on identifying and forecasting trends in cyberwarfare, including at the intersection of AI. It should also evaluate opportunities for leveraging cyber capabilities among states that opt in to international AI coordination regimes. Such opportunities may include internet protocol undergirding AI developed by non-cooperative states.

Acknowledgements

I would like to thank my supervisor, David Goutor, for his guidance and facilitation of my research, as well as Matthew da Mota and Madison Lee for their feedback and commentary. I would also like to thank Reanne Cayenne and Dianna English for their generosity in supporting my work.

About the Author

Wim Howson Creutzberg is a Digital Policy Hub undergraduate fellow who recently completed a B.A.Sc. at McMaster University. He is interested in governance mechanisms for mitigating collective action problems and AI policy and is researching how international AI policy proposals enforce coordination. Wim has been a Pivotal Research fellow and has also volunteered with the Collective Intelligence Project and the Foresight Institute.

Works Cited

Aschenbrenner, Leopold. 2024. “Situational Awareness: The Decade Ahead.” *Situational Awareness* (blog), June. https://situational-awareness.ai.

Bonfanti, Matteo E. 2022. “Artificial intelligence and the offense–defense balance in cyber security.” In *Cyber Security Politics: Socio-Technological Transformations and Political Fragmentation*, edited by Myriam Dunn Cavelty and Andreas Wenger, 64–79. CSS Studies in Security and International Relations. Abingdon, UK: Routledge.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe et al. 2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv.* February. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

Garfinkel, Ben and Allan Dafoe. 2021. “How does the offense-defense balance scale?” In *Emerging Technologies and International Stability*, edited by Todd S. Sechser, Neil Narang and Caitlin Talmadge, 243–74. Abingdon, UK: Routledge.

Gordon, Cindy. 2024. “Microsoft And OpenAI Partner On $100 Billion U.S. Data Center, Report Says.” *Forbes*, March 31. www.forbes.com/sites/cindygordon/2024/03/31/microsoft-and-openai-partnering-on-stargate-a-100b-us-data-center/.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley and Jeff Alstott. 2024. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models.* RAND Research Report. May 30. Santa Monica, CA: RAND. www.rand.org/pubs/research_reports/RRA2849-1.html.

Sevilla, Jaime and Edu Roldán. 2024. *Training Compute of Frontier AI Models Grows by 4–5x per Year.* Epoch AI Report. May 28. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year.

Shamir, Eitan. 2021. “Deterring Violent Non-state Actors.” In *Netherlands Annual Review of Military Studies 2020: Deterrence in the 21st Century — Insights from Theory and Practice,* edited by Frans Osinga and Tim Sweijs, 263–86. The Hague, the Netherlands: Asser Press.

Slayton, Rebecca. 2017. “What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment.” *International Security* 41 (3): 72–109. https://doi.org/10.1162/ISEC_a_00267.

U.S.-China Economic and Security Review Commission. 2024. *2024 Report to Congress of the U.S.-China Economic and Security Review Commission.* November. Washington, DC: US Government. www.uscc.gov/sites/default/files/2024-11/2024_Annual_Report_to_Congress.pdf.

US Department of Homeland Security. 2024. *Department of Homeland Security Report on Reducing the Risks at the Intersection of Artificial Intelligence and Chemical, Biological, Radiological, and Nuclear Threats.* April 26. www.dhs.gov/publication/fact-sheet-and-report-dhs-advances-efforts-reduce-risks-intersection-artificial.

Winter-Levy, Sam. 2024. “The AI Export Dilemma: Three Competing Visions for U.S. Strategy.” December 13. Washington, DC: Carnegie Endowment for International Peace, https://carnegieendowment.org/research/2024/12/ai-artificial-intelligence-export-united-states.

Zwetsloot, Remco, Helen Toner and Jeffrey Ding. 2018. “Beyond the AI Arms Race: America, China, and the Dangers of Zero-Sum Thinking.” *Foreign Affairs*, November 16. www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race.