Conference Proceedings

GREGORY SMITH, SYDNEY KESSLER, JEFF ALSTOTT, JIM MITRE

Industry and Government Collaboration on Security Guardrails for AI Systems

Summary of the AI Safety and Security Workshops

For more information on this publication, visit **www.rand.org/t/CFA2949-1**.

Published by the RAND Corporation, Santa Monica, Calif.

About These Conference Proceedings

The RAND Corporation and the Carnegie Endowment for International Peace hosted a series of meetings that culminated with the AI Safety and Security Workshop on July 12, 2023, at RAND's office in Arlington, Virginia. These conference proceedings capture the industry and government perspectives emerging from the workshops to inform policymakers and the broader public discussion about AI safety and security.

The research reported here was completed in September 2023 and underwent security review with the sponsor and the Defense Office of Prepublication and Security Review before public release.

RAND National Security Research Division

This research was conducted within the International Security and Defense Policy Program of the RAND National Security Research Division (NSRD), which operates the RAND National Defense Research Institute (NDRI), a federally funded research and development center (FFRDC) sponsored by the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the Navy, the Marine Corps, the defense agencies, and the defense intelligence enterprise. This research was made possible by NDRI exploratory research funding that was provided through the FFRDC contract and approved by NDRI's primary sponsor.

For more information on the RAND International Security and Defense Policy Program, see www.rand.org/nsrd/isdp or contact the director (contact information is provided on the webpage).

Contents

Chapter 1

Introduction

The rapid evolution of artificial intelligence (AI) technology offers immense opportunity to advance human welfare. However, this evolution also poses novel threats to humanity. Foundation models (FMs) are an AI trained on large datasets that show generalized competence across a wide variety of domains and tasks, such as answering questions, generating images or essays, and writing code. The generalized competence of FMs is the root of their great potential, both positive and negative. With proper training, FMs could be quickly deployed to enable the creation and use of chemical and biological weapons, exacerbate the synthetic drugs crisis, amplify disinformation campaigns that undermine democratic elections, and disrupt the financial system through stock market manipulation. Moreover, it is plausible that FMs could lead to breakthroughs in artificial general intelligence,[1] posing existential risks as AI becomes increasingly capable of performing tasks and making decisions once the sole purview of humans.

Reflecting these concerns, the RAND Corporation and the Carnegie Endowment for International Peace (CEIP) hosted a series of workshops in July 2023 with government and AI industry leaders to discuss developing security guardrails for FMs. We held four working-level meetings to exchange views on advances in FMs, potential risks, and desired governance solutions. These meetings culminated in a principal-level workshop at RAND that included deputy- and under secretary–level representatives from U.S. government, senior executives from major AI developers, and White House staff. Participants identified concerns about AI's impact on national security, potential policies to mitigate such risks, and key questions to inform future research and analysis.

These conference proceedings capture the industry and government perspectives emerging from the workshops to inform policymakers and the broader public discussion about AI safety and security.[2] The workshops were held under the Chatham House Rule to foster open communication and frank feedback by not attributing views to individuals in discussions of the workshops. Views that participants expressed were their own and did not reflect their formal organizational affiliations. Indeed, few organizations had formal positions on most of the issues emerging from the workshop, as their thinking is still in nascent stages of development.

[1] *Artificial general intelligence* refers to a theoretical class of AI capable of performing intellectual tasks on par with human cognition.

[2] Participants have not reviewed and approved this document. This represents a synthesis and analysis from RAND of statements made by workshop participants.

Chapter 2

Workshop Insights

The workshops hosted by RAND and CEIP advanced the group's collective understanding of AI safety and security and identified the following insights for government and industry to consider.

Anticipate Foundation Model Developments to Better Prepare for Future Risks

Both government and industry workshop participants reported feeling that they were "behind the ball" when it came to identifying and understanding the capabilities and risks of current FMs. Developers noted that they were often surprised by the capabilities they discovered in their own products. Participants voiced strong interest in conducting research designed to better predict and prepare for future FM risks and the impacts of FM on society. Participants also agreed that it would be important to predict future risk areas so that mitigations could be developed before FM technology advanced to the point at which those risks would be realized.

Broad agreement on the need to accelerate understanding of AI's capabilities provided an opportunity for participants to probe specific areas in need of development. For example, they identified developing additional knowledge on how to evaluate models and recognize and mitigate threats as high priorities for both government and industry actors.

Develop Better and More-Comprehensive Threat Assessment, Red-Teaming, and Mitigation Capabilities to Enhance Artificial Intelligence Safety for Foundation Models

Both government and industry participants identified a need to develop standardized Red-teaming and threat assessment capabilities to enhance AI safety for FMs. Because capable FMs are a relatively immature industry, the evaluation, assessment, and Red-teaming techniques for detecting undesired AI behavior are underdeveloped. Participants stressed the value of further developing such techniques. In addition, participants urged that independent researchers participate in developing ways to evaluate and control AI to ensure that such practices become widespread. The recent voluntary collaboration between AI developers and the White House provides a potential model for building a government-industry testing and Red-teaming process.[3]

Participants also underscored the need to develop techniques for mitigating threats and preventing models from enabling dangerous behavior. The National Science Foundation's recent announcement

[3] White House, "FACT SHEET."

establishing a series of National Artificial Intelligence Research Institutes could kindle an effort to build government research capability; such efforts could be enhanced by partnering with industry.[4] There was also discussion about how to create open standards for Red-teaming, with some participants suggesting that it may be too inefficient to confine all Red-teaming to private relationships between companies or between companies and the government. These open standards could also enable open-source Red-teaming by encouraging positive norms among researchers and developers of large language models outside formal Red-teaming conducted by corporations or government actors.

Develop a Mechanism for Government-Industry Collaboration on Red-Teaming

Both government and industry participants expressed their interest in building a robust, reliable, and accurate Red-teaming process that could assess a model's potential to disclose dangerous or classified information, such as instructions to construct a bioweapon or nuclear weapon. Participants from both industry and government agreed that this process should occur in a classified environment to provide the level of access needed to assess risks.

Areas of uncertainty remained about execution of the government-industry Red-teaming and threat assessment. First, there were uncertainties about which experts would be involved, particularly when working with classified information. Industry participants highlighted their concerns about whether relevant information, including information potentially derived from classified sources, could be quickly provided to experts. These participants suggested that it would be useful for government agencies to identify and provide experts with knowledge of national security risks who could inform risk evaluations. Participants were also uncertain whether classified Red-teaming would be done by government teams or by cleared individuals within companies, as well as how remediation of issues related to classified information would be handled by developers.

There were also questions about how the results of government-industry Red-teaming and AI assessments would be used. Industry participants were uncertain about the consequences of a model failing a threat assessment. For example, what would happen if a model were shown to disclose classified nuclear weapons information? Industry participants requested clarification on how negative results might affect their ability to deploy models and how a process could be implemented to demonstrate that troublesome issues had been addressed. They noted implications for liability and suggested that some form of indemnification may be necessary to garner industry participation.

Identify Thresholds for Models to Receive Additional Scrutiny

A broad cross-section of workshop participants agreed that the most-severe scrutiny should be focused on the most-capable and most-powerful models. This was a particular concern for government workshop participants, who wished to focus their limited resources on the models with the greatest potential for large-scale national security impacts. Industry participants generally agreed that they would need such thresholds to focus their own capabilities.

[4] National Science Foundation, "NSF Announces 7 New National Artificial Intelligence Research Institutes."

There was no consensus about the preferred threshold for additional scrutiny—but not because participants strongly disagreed on proposed thresholds. Rather, given the speed with which AI technology is evolving, they were not sure whether ideas suggested in workshop sessions would capture all the particularly powerful or dangerous models that may arise in the near future of this rapidly changing landscape. Noting the strong relationship between the compute capacity used to train a model and a model's capabilities, several participants suggested that investigative resources should focus on models above a certain compute threshold. Alternatives were discussed as well, such as a capabilities-based threshold that would target models demonstrating certain abilities, but some participants were uncertain whether such a threshold was feasible, considering the uncertainty around AI development at the time of these workshops.

Consider International Implications of Any U.S. Government Artificial Intelligence Guardrails

Participants highlighted that U.S. government officials should anticipate that any FM governance requirements they impose may be mimicked by other countries, who would demand similar access to models as that provided to the U.S. government. This was of high concern for several industry participants and was echoed by government officials as an issue they were tracking.

Prioritize Research and Analysis That Aim to Address Risks of Open-Source Models

Participants shared concerns about the risks posed by open-source models, noting ways in which safety guardrails could be removed from such models after they have proliferated widely. Although these risks were identified, participants reached no consensus about how to best mitigate these potential threats or whether there were feasible options to do so. Participants did suggest that it would be more tractable to exercise oversight beforehand rather than try to monitor open-source models after they are created. Some participants believed that, for the foreseeable future, creating FMs would require very large amounts of computation available to just a few companies. Use of that computation for creating models above a certain size or capability could be subject to safety and security requirements. Both sides recommended further efforts to mitigate open-source risks, as well as to communicate with the open-source community to identify potential paths forward.

Chapter 3

Short-Term Policy Actions

Participants at the principal-level meeting identified a series of short-term policies and actions that could be taken now to make meaningful progress toward addressing national security threats from AI. These quick wins could be executed in three to six months and would help both government and AI developers grow the skills needed to manage AI in the future:

- Establish a mechanism for Red-teaming and threat assessments that allows the government to test models in a classified environment and share certain results with AI developers.
- Further develop model-assessment procedures to test FMs for national security threats.
- Consider declaring FMs and high-performance computing as critical infrastructure.

Establish a Government-Industry Red-Teaming and Threat Assessment Mechanism

Participants largely supported creating a mechanism that enabled government-industry Red-teaming and threat assessments in a classified environment for the most highly capable AI. These assessments would focus on identifying and measuring capabilities relevant to U.S. national security, such as the ability to enhance bioweapon production.

A classified threat assessment component would allow the government to test for particularly sensitive capabilities when such testing would be inappropriate in an unsecured environment. Such a component would also enable testing for the presence of classified information within the model's training data and assessing a model's capabilities to recall, find, or compose classified information. For example, it would be possible to test a model's ability to provide nuclear weapon information, which is *born secret* under current law, or to conduct a more thorough assessment of dangerous capabilities for which the results would be automatically protected.[5]

The Red-teaming and threat assessment process would require a clear definition of the roles and responsibility for both industry and government, as well as well-defined controls on information entering or leaving the program and a shared understanding about the consequences if dangerous capabilities were detected. Workshop participants identified the U.S. government's cyber Vulnerabilities Equities Process as a potential model for developing a Red-teaming process.[6]

Workshop participants also identified several steps that AI developers and government could take to create a Red-teaming partnership. On the industry side, developers must find ways to provide

[5] In the United States, documentation on nuclear weapons is automatically classified at the time of creation; see DeVolpi et al., *Born Secret*.

[6] White House, "Vulnerabilities Equities Policy and Process for the United States Government."

model access to the government that affords appropriate security to government testers. This could involve creating specialized application programming interfaces for government users that would be visible only to cleared individuals within labs or that otherwise protect the contents of government testing from surveillance by corporate employees.

On the government side, questions remained about how to pass sensitive data to AI developers, especially classified information. Most AI developers do not have cleared personnel, and they have foreign nationals working directly in the AI development process. Government leaders will have to identify ways to share potentially sensitive information with AI developer teams and build partnerships with industry that respect the United States' needs to secure information—while ensuring that Red-teaming results regarding dangerous threats are appropriately communicated to the developers who can mitigate them. A process for clearing certain engineers at AI developers might be necessary to ensure a sufficient pool of industry-side individuals to support shared Red-teaming and threat mitigation work around classified information.

Government leaders will also need to clearly specify the implications for developers if vulnerabilities or concerning capabilities are identified in a model. Industry workshop participants noted that this would be a major area of concern for them in any government-industry AI Red-teaming partnership. This issue remains unresolved, but a solution could potentially involve providing some form of liability shield to a company if it is alerted to a vulnerability and undertakes good-faith efforts to resolve it.

Develop and Spread Threat Assessment Techniques

The need to get ahead of FM developments and understand how to mitigate potential threats may require government-industry collaboration, especially to identify potentially dangerous capabilities and to share methods for controlling FMs and preventing them from enabling dangerous behavior. An early example of such a technique that could be encouraged across FM developers by government-industry collaboration is the GPT-4 system card created by OpenAI that was informed by a preliminary model evaluation by the Alignment Research Center, which discussed specific tests performed on GPT-4 before its release to identify dangerous capabilities that were tuned out of the model.[7] Recent commitments by major U.S. AI developers also include a pledge for "sharing information across the industry and with governments, civil society, and academia on managing AI risks," including sharing best practices for AI safety.[8] These initiatives are consistent with the consensus of workshop participants about the importance of developing and spreading techniques to assess and mitigate threats. Government and industry could consider how best to support research into these practices through additional funding to ensure that AI safety efforts keep pace with the development of AI capabilities.

Much of the work in this space will also involve developing specific policies and procedures for sharing useful information. Policymakers can build on these early initiatives to begin standardizing certain evaluation processes—for example, by encouraging the publication of system cards based on

[7] OpenAI, *GPT-4 System Card*.

[8] White House, "FACT SHEET."

third-party evaluations that assess new models before their release. Industry and government could explore partnerships for further popularizing specific evaluation and safety techniques and creating forums for sharing this valuable information. These efforts could elicit new techniques from specific AI developers and spread them more widely and quickly through the industry.

Consider Declaring Foundation Models and High-Performance Computing as Critical Infrastructure

The information technology (IT) sector is currently one of the U.S. Department of Homeland Security's critical infrastructure sectors, with planning and support for the IT sector overseen by the Cybersecurity and Infrastructure Security Agency.[9] However, the IT sector is far broader than just AI, encompassing other types of infrastructure, such as internet service.[10] Declaring FMs and high-performance computing as a specific critical subsector would improve efficiency by focusing on these technologies rather than on the entire IT sector.

This declaration could serve as a foundation for additional policymaking by the Department of Homeland Security regarding FMs and high-performance computing. However, such a declaration could act as an access point that would help AI companies participate in infrastructure-related bodies and coordination groups that work to improve critical infrastructure functions and resilience. For instance, within the IT sector and the forums of the Information Technology Sector Coordinating Council and Government Coordinating Councils, industry players in this space directly liaise with each other and the federal government to accomplish such tasks as streamlining information sharing, establishing best practices for risk management, and prioritizing cybersecurity efforts.[11] Establishing AI as a separate critical infrastructure sector can help to build relationships between government and AI labs by creating a single point of contact for the AI industry and to focus government officials' attention on the sector. This approach will also provide an opportunity for the government to begin developing plans for managing FMs as they are integrated into the country's infrastructure and determining how best to ensure their functionality in potential disasters.

As FMs begin to be deployed throughout the United States, they will become increasingly important to society's functioning. It is possible that FMs may even directly support or administer critical infrastructure systems. A critical infrastructure declaration could jump-start thinking about how to support FMs' role in infrastructure and how to best mitigate risks as the technology becomes more widespread. Such a declaration could also be a low-cost way to begin laying the groundwork for a deeper partnership between AI developers and government in safeguarding future AI deployment and ensuring responsible widespread usage.

---

[9] Cybersecurity and Infrastructure Security Agency, "Information Technology Sector."

[10] Cybersecurity and Infrastructure Security Agency, "Information Technology Sector."

[11] U.S. Department of Homeland Security, *Information Technology Sector-Specific Plan*.

Chapter 4

Conclusion

The speed at which AI technology in general and FMs in particular are evolving has kindled an urgent need for government-industry collaboration to ensure that the technology achieves its greatest potential while also guarding against potential risks. Workshop participants regularly disagreed on a variety of issues, but there was consensus that communication and collaboration on AI were essential moving forward. This spirit of collaboration can be augmented by trust-building exercises and quick wins for AI governance through government-industry partnerships. Such efforts may lay the groundwork for future, more-extensive governance of AI and ensure that the technology's national security risks are appropriately managed.

References

Cybersecurity and Infrastructure Security Agency, "Information Technology Sector," webpage, undated. As of August 4, 2023: https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors/information-technology-sector

DeVolpi, A., G. E. Marsh, T. A. Postol, and G. S. Stanford, *Born Secret: The H-Bomb, the Progressive Case and National Security*, Pergamon Press, 1981.

National Science Foundation, "NSF Announces 7 New National Artificial Intelligence Research Institutes," May 4, 2023. As of October 12, 2023: https://new.nsf.gov/news/nsf-announces-7-new-national-artificial

OpenAI, *GPT-4 System Card*, March 23, 2023.

U.S. Department of Homeland Security, *Information Technology Sector-Specific Plan: An Annex to the NIPP 2013*, 2016.

White House, "Vulnerabilities Equities Policy and Process for the United States Government," November 15, 2017. As of October 12, 2023: https://trumpwhitehouse.archives.gov/sites/whitehouse.gov/files/images/External%20-%20Unclassified%20VEP%20Charter%20FINAL.PDF

White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," July 21, 2023. As of October 12, 2023: https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/