RAND

Conference Proceedings

JON MENASTER

Foundation Model Convening

Conference Summary

The rapid evolution of artificial intelligence (AI) technology offers immense opportunity to advance human welfare.[1] However, this evolution also poses novel threats. Foundation models (FMs), unlike other AI, are trained on large datasets that show competence across a wide variety of domains and tasks, such as generating text, audio, video, and images. The generalized competence of FMs is the root of their potential. In terms of potential positive outcomes, FMs could provide benefits across a wide variety of sectors, such as education. For example, classroom lessons could be made more personalized and interactive.[2] In terms of negative outcomes, FMs could enable the creation of chemical and biological weapons or amplify disinformation campaigns that undermine democratic elections. For example, biosecurity experts have partnered with a leading AI company to demonstrate that current models can produce sophisticated, accurate, useful, and detailed information about how to design and acquire a biological weapon.[3] Additionally, AI could exacerbate the volume, velocity, variety, and virality of disinformation, leading to automation of disinformation campaigns.[4]

Reflecting these concerns, RAND and the Carnegie Endowment for International Peace hosted a convening in November 2023 with AI industry, academic, and think-tank leaders to discuss a variety of topics relating to FM safety and security.[5] We held a preconvening discussion with a subset of participants to exchange views on advances in FMs, potential risks, and desired governance solutions. This meeting led to a principal-level convening of approximately 30 people whose backgrounds were split among industry, think tanks, and academia. 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. The convening involved a wide variety of sessions, including a fireside chat, a historical presentation, small-group workshops, brief-outs from the workshops to a larger all-convening setting, panel discussions, and remarks from a special guest.

These conference proceedings capture the industry, academic, and think-tank perspectives emerging from the workshops we held. These workshops occurred during the broader convening to inform policymakers and the broader public discussion about AI safety and security.[6] 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.[7] Views that participants expressed were their own and did not reflect their formal organizational affiliations.

Short-Term Policy Actions

Participants identified two areas of focus for policy actions and changes that could be taken now to meaningfully increase the safety and security of FM development.

Use Government-Industry Partnerships to Develop an Action Plan for Threat Response

Participants noted the importance of developing and testing plans in advance of serious crises. One participant suggested creating a strong default of automatic shutdown in the event of a crisis without affirmative human control (i.e., a deadman solution). For example, multiple keys could be distributed to people around the world, and an automatic shutdown of an AI model could occur if the keys were not used within a specified time. Another participant suggested on-chip governance mechanisms that could be triggered to shut down computer hardware if a certain event did or did not happen.

Participants noted that, regardless of the specific plan ultimately used, it is important to reduce the cost of implementing the plan by making it politically and socially viable. Steps to improve plan viability may include ensuring that the economy and critical infrastructure are not so dependent on vulnerable AI models that plan implementation is delayed.

Participants discussed the need to prepare for a wide variety of potential scenarios involving FM safety and security and how tabletop exercises or wargames might be useful ideation and training activities.

Conduct Regular Tabletop Exercises and Wargames to Prepare Responses to a Variety of Scenarios

Participants discussed the need to prepare for a wide variety of potential scenarios involving FM safety and security and how tabletop exercises or wargames might be useful ideation and training activities. These mechanisms could be used to test the potential emergency powers that a government could exercise in the case of loss of control of an AI system, such as the ability to shut down a data center or cut off internet access. This testing might also help highlight the types of additional resources government actors may need to properly protect the public from safety and security concerns that might arise during FM development.

Workshop Insights

The workshops held during this convening advanced the group's collective understanding of AI safety and security and identified insights for policymakers to consider. Additionally, we asked each session to outline potential next steps to address any concerns identified. Each header in this section represents a workshop held during the convening. Participants could decide which workshops they wished to attend. The workshops were moderated and did not incorporate insights from the larger group sessions.

Be Prepared for a Wide Variety of Loss-of-Control Scenarios from Agentic Systems

First, participants disagreed about what constituted an AI *agent* or AI *agentic system*. One definition focused on agents as falling within a spectrum, with an agentic level centered around the degree to which a system can adaptively achieve goals in complex environments without human supervision. Another

set of definitions attempted to simply describe whether a system was an agent or not, potentially by labeling it either as fully autonomous or as requiring a human copilot.

From there, participants discussed loss of control and agreed on several concepts. In particular, attempts by an AI agent to spend or accrue money should be reviewed carefully and not allowed to occur independently. This could entail setting up processes such that an AI agent would be required to receive human approval if it wants to spend more than a certain amount of money. Another suggestion participants discussed was limiting the planning window of agents during the training process, so developers never reward the ability to plan many steps ahead. There was also discussion about the danger of agents colluding with one another, with potential solutions for collusion revolving around limiting agents from spawning new agents and/or being constrained in how they communicate with one another.

In terms of next steps, participants discussed how loss-of-control concerns would be hard to counter from a policymaking perspective unless policymakers are given realistic demonstrations of how such issues could occur—therefore, such demonstrations should be scheduled more often.

Strong Cybersecurity and Information Sharing Are Required to Protect Model Weights

The conversation highlighted the importance of using defensive cybersecurity to successfully protect model weights. These defensive efforts would have the dual effect of restricting human-driven proliferation and hampering models that could potentially self-exfiltrate, e.g., escape model controls. Additionally, participants raised concerns about the need for important debates on the relative consequences of moving cybersecurity toward a national security direction. For example, AI developers may be unwilling to increase their cybersecurity posture to one that resembles a Sensitive Compartmented Information Facility, in which certain types of classified information are processed.

The conversation highlighted the importance of using defensive cybersecurity to successfully protect model weights.

Participants agreed that other industries use frameworks that could increase FM cybersecurity. One example is the Department of Energy's Cybersecurity Risk Information Sharing Program, which is used in the U.S. electricity industry. This public-private partnership delivers relevant and actionable cybersecurity information to electricity industry participants. Other, analogous cybersecurity approaches from such industries as life sciences and nuclear energy and from such organizations as the Cybersecurity and Infrastructure Security Agency can help organize information-sharing and analysis centers.

Participants noted that a key question remains about the best way to make hardware safe against model-weight exfiltration, with tamperproof self-limiting chips being one option that could decrease risk and provide more regulatory flexibility in ways to govern the possession and use of AI hardware.[8] Another promising approach that remains theoretical is the development of machine learning–specific variants of hardware security modules.[9] A forthcoming RAND report on FM cybersecurity will offer recommendations on securing model weights against advanced attackers, which will be useful to both policymakers and AI developers. An interim report was released in October 2023.[10] Another is for all involved entities (academia, nonprofits, industry, governments, think tanks) to continue working together to ensure a strong government-industry relationship to help maintain the strongest possible security posture against emerging cybersecurity threats.

Specific capabilities likely to emerge from new models are deemed critical trade secrets by AI developers, presenting regulators with a dilemma about how to be informed enough to properly regulate the capabilities.

Research and Collaboration Are Needed to Address Dangerous Capabilities

The conversation involved (1) questions about what dangerous capabilities require regulatory intervention and (2) whether systems being built to manage AI risk are robust against unexpected risks or major increases in threats. High-level categories of dangerous capabilities include recursive self-improvement, autonomy, and self-goal setting. Some participants felt AI developers were already close to realizing some of these capabilities. Specific capabilities likely to emerge from new models are deemed critical trade secrets by AI developers, presenting regulators with a dilemma about how to be informed enough to properly regulate the capabilities.

To help resolve these concerns, participants generally agreed that government and AI developers need to collaborate more closely to prepare for potentially disruptive capabilities. A useful next step to achieve this goal would be to create mechanisms to game out how the development of dangerous AI capabilities could be communicated and determine the costs and benefits of the various research and collaboration options. Another important future direction of research would be to identify specific capabilities that would act as a threshold that society would not want to cross and that would result in the end of that line of work or its movement into the classified space.

Developer Ethical Norms Should Emerge from the Technical Community

Participants agreed that AI is a very young field that does not have broad professional norms like other engineering disciplines. Professional norms are an important part of any field, and it will be important for norms to emerge from within the technical community. While participants did not discuss this in more depth, this could happen either organically or through other methods, such as the creation of professional associations. For example, running benchmarks on models is a norm that quasi-organically developed from within the technical community. As these norms emerge, there will need to be consequences or other cost impositions on norm violators. For example, working for companies that violate norms should carry negative professional implications.

Participants discussed the following potential norms:

- building systems that aim to augment humans instead of replace them
- conceiving of AI models as tools, not a new species of conscious beings
- not deceiving people about the extent to which AI is used, which could include watermarking AI-generated content[11]
- rolling back deployed AI systems that are shown to have negative or dangerous effects.

International Governance Must Occur Across Multiple Independent Tracks

There need to be multiple tracks of international governance that span the sharing of scientific information and coordination on regulation.[12] To successfully implement these tracks, it will be paramount

to understand the relevant current AI safety actors and those who are likely to become the next generation of emergent actors. Furthermore, governance success will be difficult because of the conflicting geopolitical interests and national security concerns of key countries. For example, participants worried about one country's newly established stance on AI regulation, which seems nationalistic and opposed to important constraints on open-source model release.

Participants agreed on three generally useful next steps. First, the wide variety of existing diplomacy and dialogue taking place about international governance should be shaped to proceed in a positive direction. This could include better understanding which efforts are the most promising and where there are gaps in the dialogue that need to be filled. Second, national AI safety institutes (such as the newly announced United Kingdom and U.S. AI safety institutes) need to be adequately scoped and promoted. Third, international governance could be achieved by using existing organizations (such as the United Nations), creating new multilateral institutions, or using some combination of these, depending on practical considerations. Some examples of potential new multilateral institutions include a group akin to the International Atomic Energy Agency; the European Organization for Nuclear Research; or, potentially, a smaller entity composed primarily of countries that maintain treaty alliances with one another (e.g., the "Five Eyes": Australia, Canada, New Zealand, the United Kingdom, and the United States).

Developers Face Substantial Legal Risk in the Absence of Clear Laws or Legal Precedents

The convening held two separate sessions discussing legal risks: The first was a panel featuring distinguished legal academics discussing liability issues related to AI, and the second was a specific dialogue track for discussing liability for AI-caused harms.

The legal scholar panel discussed (1) how current liability rules affect AI development and how such rules describe liability distribution in the event of harms, (2) what success looks like with regard to liability and AI, and (3) the extent to which insurers will act as a key check on safe AI development. How current liability rules apply to AI is unclear; thus, there is significant risk for developers if large-scale harms occur and are adjudicated in court. Some participants suggested that liability rules could be clarified and that liability shields could be created that require developers to follow certain safety precautions in exchange for greater liability protections. With regard to insurance, the panel explained that insurers will have the opportunity to play a major role in incentivizing safe AI development because they can refuse to underwrite policies to protect companies that are not engaging in safe AI development and, hence, increase companies' risk of having to pay large damage claims.

How current liability rules apply to AI is unclear; thus, there is significant risk for developers if large-scale harms occur and are adjudicated in court.

The participants framed the discussion by observing that software is not a good analogy for AI because the AI harms that most concerned participants are those to third parties (such as members of the public who are affected by AI use), while much of the law around software liability focuses on end-user harm. There were several areas of broad agreement in the liability conversation, including that liability will be used as a means of setting precedent, absent more-comprehensive regulation. However, participants did not see any clear guidance from past doctrine on how courts might apply liability rules to novel issues raised by AI, and AI liability cases are just beginning

to be filed. Some participants also noted that social feelings about AI (e.g., does the risk outweigh the benefit) will help determine how much liability AI developers are likely to face. Creating industry standards and norms will help clarify liability risks for developers by offering them a clearer path to follow, which might reduce their liability exposure. In the interim, AI developers currently face substantial legal risk as they deploy AI more broadly.

Participants agreed on two significant next steps. First, legal scholars should develop analyses of potential precedents in the legal system that could apply to AI, then use these analyses to help determine what new laws, rules, or other forms of change might be necessary to address legal gaps. Second, liability law, as created by the courts through the accumulation of precedent, and common law, which emerges from judicial decisions, should be written to encourage the development of industry norms and customs of safe AI development and to clearly punish developers who ignore such customs and cause harm.

Funding Research and Developing Guidelines Are Key Next Steps for Red-Teaming

The discussion focused on the current state, challenges, and future directions of red-teaming for threat assessment and model evaluation. The conversation highlighted the importance of balancing scientific rigor against practical application needs. For example, successful red-teaming will require a scientifically rigorous approach that enables the development of high-quality and reproducible methodologies. In particular, participants agreed that red-teaming requires clear definitions and standardized methodologies to be effective across different domains and risk scenarios. However, because of the varying domains, the role and effectiveness of red-teaming remains context dependent, requiring tailored approaches for specific model evaluations. Given the difficulty of solving the these problems, there was disagreement about the usefulness of red-teaming. Specifically, because of the current lack of scientific consensus on what constitutes good red-teaming, it would be premature to heavily weight any particular red-teaming result. Therefore, it could be better to move from a heavy emphasis on red-teaming to balancing red-teaming with an increase in funding for interpretability research.[13]

There was agreement on several key next steps, including a need to promote and fund research in the development of advanced red-teaming methodologies to establish a more scientific standing for the field. Furthermore, it will be important to develop comprehensive guidelines and standards for red-teaming, including scientific protocols for different types of risks and models.

The government has a legitimate national security interest in collecting information from private actors and should use reporting requirements to fulfill that interest.

Reporting and Information-Sharing Will Shape the Government Response to AI Development

Participants agreed on several key areas of a successful reporting and information-sharing regime. First, the government has a legitimate national security interest in collecting information from private actors and should use reporting requirements to fulfill that interest. Second, the government will need to make significant policy decisions throughout all phases of AI development, from pretraining to deployment, and should collect information from private entities to make the best possible policy choices. Third, participants recommended that the government review policies for voluntary collection of sensitive informa-

tion, such as trade secrets, and consider collecting only the information necessary for policy decisions, such as export controls or security and safety requirements. Fourth, it is desirable to establish a high-trust information-sharing regime between government and AI developers, which could encourage sharing of relevant information while allowing AI developers to protect their most sensitive intellectual property. This could involve carefully posing questions for AI developers that strike a balance between being specific enough to elicit useful information while not being too specific, such that a developer is required to divulge trade secrets.

Participants differed regarding the tone of the relationship the government should strike when collecting information using mandatory instruments, such as the reporting requirements authorized in section 4.2 of the Biden administration's executive order on AI.[14] Participants agreed that a high-trust information-sharing regime is best for all parties, but some expressed skepticism that private actors would be willing to share useful information. Because of this view, some participants suggested that the government require greater information-sharing as soon as possible. This might involve making more-forceful requests for information without waiting for voluntary compliance.

Structured Access to Foundation Models Will Be Key for Research and Evaluation Efforts

AI developers granting structured access to their models will be a key component of successful interpretability and alignment research, as well as various types of evaluation and auditing.[15] There was general agreement that application programming interfaces will be a key mechanism for enabling structured access, but it remained unclear who should be responsible for building the interfaces. Developers have more expertise but less independence, while third parties who might use the application programming interface (such as researchers and evaluators) are more independent but are not experts in any particular model. Neither option is perfect.

Participants mostly agreed that, while compute is not the perfect tripwire solution, it is the best currently available.

Participants did agree that different groups should have varied levels of access to different types of model information (similar to the national security concept of *need to know*). This differing level of access will help alleviate developer concerns about intellectual property theft and related privacy issues. However, because of the tension between what AI developers want to protect and what external researchers and evaluators want to access, opinions differ on exactly what level of access is necessary and appropriate.

Several next steps were discussed, including the creation of a safe harbor setup that would alleviate developer legal concerns around granting model access to third-party evaluators.[16] More work is also needed on how model weight inspection might work—this could result in something similar to a Sensitive Compartmented Information Facility, but participants were unclear on how best to operationalize this concept. Participants also thought a valuable next step would be to develop a stronger empirical sense of specifically which levels of access would provide what research or evaluation benefits.

Measuring Compute Is the Best Currently Available Tripwire Solution

Participants mostly agreed that, while compute is not the perfect tripwire solution, it is the best currently available.[17] Participants agreed that compute is clearly better than parameter count, which is not a robust tripwire.[18] Further, when governments use existing authorities to request information from AI

developers, it will be important to recognize what they will willingly give. For example, the training data would be considered important intellectual property that developers will likely not want to report. There was slight disagreement over whether smaller training runs (less than 10^{26} FLOP) with science corpuses in training data should be reported and whether the reporting threshold should be lowered over time. Political tractability will play an important role in determining any changes to reporting thresholds. Over time, greater algorithmic efficiency or the development of powerful narrow models may render compute thresholds irrelevant.[19] In addition to quantitative measures, participants discussed qualitative measurements, which could be assessed with reference to the developer's own marketing materials or internal communications. There, qualitative measurements could be used independently or in conjunction with quantitative thresholds.

In terms of next steps, one way to operationalize tripwires could be to use a ledger (or some other internal accounting system) to mandate reporting about training runs and compute spent and about when a threshold is hit, triggering a reporting requirement. Doing so may be difficult because it is difficult to determine who and for what purposes compute is being used and because there are obvious incentives to provide misleading information. These incentives could lead to falsification of records or evading information-sharing by combining smaller models. To combat these perverse incentives and ensure that the executive order and Defense Production Act provisions are implemented smoothly, stakeholders should establish clear standards and tools for measuring training compute, and the government can consider deploying Bureau of Industry and Security investigators to ensure that companies are not falsifying information.[20]

Conclusion

The speed of the evolution of AI technology has illuminated an urgent need for decisive policy action to ensure that the technology's benefits are enhanced while its potential risks are mitigated.[21] Workshop participants regularly disagreed on a variety of issues, but there was consensus that communication and collaboration on AI policy issues 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 and tabletop exercises and wargames. Such efforts could help create a foundation for more-extensive governance of AI to ensure that the technology's potential risks are appropriately managed.

Notes

[1] Smith et al., *Industry and Government Collaboration on Security Guardrails for AI Systems.*

[2] Harris, *Artificial Intelligence.*

[3] Anthropic, "Frontier Threats Red Teaming for AI Safety."

[4] Sedova et al., *AI and the Future of Disinformation Campaigns.*

[5] This report uses the term *we* to signify RAND and the Carnegie Endowment's collaboration across the preconvening work and the convening itself

[6] Participants have not reviewed and approved this document. This represents a RAND synthesis of workshop participants' statements.

[7] When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), or that of any other participant, may be revealed. See Chatham House, "Chatham House Rule," for more information.

[8] If hardware-enabled governance mechanisms on AI hardware can be secured against tampering, it might enable selective performance limitation of advanced chips needed to develop dangerous FMs. The selective nature of the process would enable the chips to still be used for other commercial or consumer applications. See Kulp et al., "Hardware-Enabled Governance Mechanisms." Model weight "exfiltration attacks allow attackers to steal details about a model such as its architecture or weights." See Google Security Blog, "Increasing Transparency in AI Security."

[9] According to a RAND interim report, hardware security modules could be used to aggressively isolate model weight storage. However, more research and development is needed to achieve these goals. See Nevo et al., "Securing Artificial Intelligence Model Weights."

[10] See Nevo et al., "Securing Artificial Intelligence Model Weights."

[11] The term *watermarking* means the act of embedding information that is typically difficult to remove into outputs created by AI—including into such outputs as photos, videos, audio clips, and text—for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance (Biden, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence).

[12] Multitrack diplomacy involves diplomatic exchanges made across three tracks. *Track 1 diplomacy* refers to official diplomacy, where communication is directly between or among governments. *Track 1.5 diplomacy* occurs when government representatives and nongovernmental experts engage in dialogue or meetings. *Track 2 diplomacy* denotes a wholly unofficial channel for dialogue among nongovernmental experts, without government involvement. See Sokol, "Multi-Track Diplomacy Explained."

[13] There is no concrete definition for *interpretability*. However, it can be thought of as the ability to explain or present in understandable terms to a human the output of AI (Linardatos, Papastefanopoulos, and Kotsiantis, "Explainable AI").

[14] Biden, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."

[15] As with *interpretability*, there is no concrete definition for *alignment*. *Alignment research* is generally defined as attempting to find ways to ensure that AI is aligned with human values and follows human intent. See Leike et al., "Our Approach to Alignment Research," for more on this topic.

[16] According to Cornell Law School's Legal Information Institute, *safe harbor* is a provision granting protection from liability or penalty if certain conditions are met. A safe harbor provision may be included in statutes or regulations to give peace of mind to good-faith actors who might otherwise violate the law on technicalities beyond their reasonable control. See Legal Information Institute, "Safe Harbor."

[17] *Compute* can be thought of as the number of computations needed to perform a particular task, such as training an AI model. The amount of compute used is measured in floating point operations (FLOP). A FLOP is a mathematical operation that enables the representation of extremely large numbers with greater precision. Compute performance is measured in FLOP per second, or how many computations a given resource can carry out in a second. See Vipra and Myers West, *Computational Power and AI*, for more information. *Tripwires* are mechanisms for determining when a particular model meets a threshold for elevated oversight.

[18] According to Our World in Data, *parameters* are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output (Our World in Data, "Parameters in Notable Artificial Intelligence Systems"). Google's Chinchilla AI model showed how the combination of fewer parameters and more training data could lead to increased performance. Because of this, parameter counts are not a useful way to compare model performance. See Lohn, "Scaling AI."

[19] One example of a narrow model would be one used for biological design. The Congressional Research Service defines *biological design tools* as "the tools and methods that enable the design and understanding of biological processes (e.g., DNA sequences/synthesis or the design of novel organisms)." See Kuiken, *Artificial Intelligence in the Biological Sciences.*

[20] As defined by the Congressional Research Service,

> [t]he Defense Production Act (DPA) of 1950 (P.L. 81-774, 50 U.S.C. §§4501 et seq.), as amended, confers upon the President a broad set of authorities to influence domestic industry in the interest of national defense. The authorities can be used across the federal government to shape the domestic industrial base so that, when called upon, it is capable of providing essential materials and goods needed for the national defense. (Neenan and Nicastro, *The Defense Production Act of 1950: History, Authorities, and Considerations for Congress*)

[21] Smith et al., *Industry and Government Collaboration on Security Guardrails for AI Systems.*

References

Anthropic, "Frontier Threats Red Teaming for AI Safety," webpage, July 26, 2023, As of February 22, 2024: https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety

Biden, Joseph R., Jr., "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Executive Order 14110, October 30, 2023.

Chatham House, "Chatham House Rule," webpage, undated. As of February 7, 2024: https://www.chathamhouse.org/about-us/chatham-house-rule

Google Security Blog, "Increasing Transparency in AI Security," webpage, October 26, 2023. As of March 6, 2024: https://security.googleblog.com/2023/10/

Harris, Laurie A., *Artificial Intelligence: Overview, Recent Advances, and Considerations for the 118th Congress*, Congressional Research Service, R47644, August 4, 2023.

Kulp, Gabriel, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman, "Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090," RAND Corporation, WR-A3056-1, 2024. As of February 23, 2024: https://www.rand.org/pubs/working_papers/WRA3056-1.html

Kuiken, Todd, *Artificial Intelligence in the Biological Sciences: Uses, Safety, Security and Oversight*, Congressional Research Service, R47849, November 22, 2023.

Legal Information Institute, "Safe Harbor," webpage, Cornell Law School, undated. As of February 22, 2024: https://www.law.cornell.edu/wex/safe_harbor

Leike, Jan, John Schulman, and Jeffrey Wu, "Our Approach to Alignment Research," *Open AI* blog, August 24, 2022. As of February 7, 2024: https://openai.com/blog/our-approach-to-alignment-research

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy,* Vol. 23, No. 1, December 25, 2020.

Lohn, Andrew, "Scaling AI: Cost and Performance of AI at the Leading Edge," Center for Security and Emerging Technology, December 2023.

Neenan, Alexandra G., and Luke A. Nicastro, *The Defense Production Act of 1950: History, Authorities, and Considerations for Congress*, Congressional Research Service, R43767, October 6, 2023.

Nevo, Sella, Dan Lahav, Ajay Karpur, Jeff Alstott, and Jason Matheny, "Securing Artificial Intelligence Model Weights: Interim Report," RAND Corporation, WR-A2849-1, 2023. As of February 22, 2024: https://www.rand.org/pubs/working_papers/WRA2849-1.html

Our World in Data, "Parameters in Notable Artificial Intelligence Systems," webpage, February 12, 2024. As of February 22, 2024: https://ourworldindata.org/grapher/artificial-intelligence-parameter-count

Sedova, Katerina, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan, *AI and the Future of Disinformation Campaigns*: Part 1, *The RICHDATA Framework*, Center for Security and Emerging Technology, December 2021.

Smith, Gregory, Sydney Kessler, Jeff Alstott, and Jim Mitre, *Industry and Government Collaboration on Security Guardrails for AI Systems: Summary of the AI Safety and Security Workshops*, RAND Corporation, CF-A2949-1, 2023. As of January 25, 2024: https://www.rand.org/pubs/conf_proceedings/CFA2949-1.html

Sokol, Lia, "Multi-Track Diplomacy Explained," Nuclear Threat Initiative, April 19, 2022. As of February 22, 2024: https://www.nti.org/atomic-pulse/multi-track-diplomacy-explained/

Vipra, Jai, and Sarah Myers West, *Computational Power and AI,* AI Now Institute, September 27, 2023.

About the Author

Jon Menaster is a technology and security policy fellow at RAND, where he conducts research on broadly capable artificial intelligence (AI) systems and the policy ramifications of their diffusion. Jon previously worked for the U.S. Government Accountability Office and led programmatic audits and evaluations of executive branch agencies focused on financial market and science and technology areas. He also led project management, research, drafting, and external communications efforts for various issues within GAO's non-audit technology assessment portfolio. Jon received a MA in International Economic Relations from American University. He grew up in Los Angeles and currently lives in the Bay Area.

About This Report

RAND and the Carnegie Endowment for International Peace hosted a Foundation Model Convening from November 10–12, 2023. These conference proceedings capture the industry, academic, and think-tank perspectives emerging from the workshops to inform government, civil society, industry, and the broader public discussion about artificial intelligence safety and security.

RAND Global and Emerging Risk Division

Technology and Security Policy Center

RAND Global and Emerging Risks is a division at RAND that develops novel methods and delivers rigorous research on potential catastrophic risks confronting humanity. This work was undertaken by the division's Technology and Security Policy Center, which explores how high-consequence, dual-use technologies change the global competition and threat environment, then develops policy and technology options to advance the security of the United States, its allies and partners, and the world. For more information, contact tasp@rand.org.

Funding

Funding for this work was provided by gifts from RAND supporters.

Acknowledgements

The author would like to thank everyone from RAND and the Carnegie Endowment for International Peace who collaborated to make this publication and the conference that it describes possible.

Special thanks go to the conference participants and to Lori Matsunaga, and Emily Kampe, without whom the conference would not have taken place. The author wishes to thank Jeff Alstott, Jason Matheny, Emma Westerman, and Tino Cuèllar for their help in structuring the conference and in recruiting participants. Thanks also go to our peer reviewer, Mike Vermeer, for their comments and constructive criticism. Any errors that may remain the author's sole responsibility.

www.rand.org