RAND

Conference Proceedings

ANTON SHENK

Evaluating Artificial Intelligence for National Security and Public Safety

Insights from Frontier Model Evaluation Science Day

Frontier Model Evaluation Science Day assembled more than 100 leading experts in artificial intelligence (AI), national security, and policy to address the emerging challenges of evaluating threats from advanced AI systems. The day's agenda was structured around four tracks, each focusing on a unique aspect of AI evaluation science and policy. These tracks were developed to address fundamental issues in the field while keeping the meeting agenda and invitation list manageable. The meeting's focus on evaluation methodology provided a specialized forum for in-depth discussion, distinguishing it from broader AI security topics covered in other venues. The four tracks were as follows:

- The **chemistry and biology** track focused on the intersection of AI with chemical and biological risks. This track utilized insights from previous evaluations of general-purpose and domain-specific AI models and aimed to identify current and future evaluation needs, including integrating wet lab validation and automated lab processes.
- The **loss of control** track explored scenarios in which AI systems could operate beyond the intended boundaries set by their developers or users—including AI systems deceiving humans or acting autonomously. These discussions aimed to identify early warning signs and explore strategies to prevent loss of control of AI systems.
- The **risk-agnostic methods** track sought to outline comprehensive and universal approaches to evaluating AI models, spanning such topics as red teaming, automated benchmarking, and task design. Its objective was to forge a versatile framework for assessing AI systems' capabilities, applicable across varied risk scenarios, to ensure that evaluations are consistently rigorous and at the forefront of the science.
- The **collaboration and coordination** track aimed to connect stakeholders in government, industry, and civil society to develop a shared understanding of the objectives of evaluation science. Discussions in this track centered on establishing key policy timelines and deliverables, thresholds for dangerous AI capabilities, and voluntary risk management policies for scaling AI capabilities.

The workshop proceedings synthesize insights from these sessions, outline the complexities of evaluating AI for dangerous capabilities, and highlight the collaborative effort required to formulate effective policy.

Track 1: Chemistry and Biology

The chemistry and biology (chem-bio) track illuminated the intersection of AI with chem-bio risks, incorporating insights from evaluations of general-purpose and domain-specific models. This section details lessons learned from completed model evaluations, needs and priorities for subsequent rounds of evaluations, and considerations for wet lab validation of model outputs.

Lessons Learned from Completed Model Evaluations

Embracing Complexity in Chem-Bio Model Assessments

This session highlighted the persistence of threat actors and the complex evolution of chem-bio threats. During the discussion, one participant observed a potential limitation of existing evaluation methods, suggesting that marking an entire task as failed because of early setbacks might not fully capture the resilience and adaptability of threat actors. This critique posits that a more nuanced approach accounting for threat progression and troubleshooting—such as knowing the proportion of sub-steps that succeed—could provide a more comprehensive and continuous understanding of the threat landscape. Tabletop exercises were proposed to explore the dynamics of troubleshooting and iteration further; however, their effectiveness in this context remains to be tested.

Navigating the Complexities of Dual-Use Dangers and Domain-Specific Models

In this session, participants noted that formulating concrete, detailed threat models is crucial for

understanding the landscape of potential AI-enabled chem-bio threats and the actors behind them. This process might involve exploring new capabilities that malicious actors previously did not have access to or accelerating and simplifying existing processes, making them more accessible to a broader variety of actors. Once threat models are identified and the appropriate evaluation undertaken, a notable challenge still remains, as identified by one participant: the difficulty for evaluators to accurately embody malicious actors. Although not unique to AI/chem-bio red teaming, this challenge arises, according to the participant, from a tendency to underestimate the likelihood that certain actions could succeed, leading evaluators to potentially overlook the full extent of what a malicious actor might attempt and achieve. Moreover, this difficulty sets the stage for addressing an even more formidable challenge: developing countermeasures against dual-use threats. These threats embody the inherent difficulty in differentiating between chem-bio knowledge that is helpful or benign and that which can be misused or pose a significant threat, such as designing or reconstituting pathogens more severe and deadly than those found in nature. Addressing these threats, which might involve masking malicious objectives behind seemingly benign actions, necessitates a broader approach to mitigation—including such measures as know-your-customer rules—than model-level interventions. Building on this complexity, session participants underscored the critical need to evaluate both domain-specific models (e.g., biological design tools) and general-purpose foundation models meticulously. The emphasis on domain-specific models stems from the unique risk profile associated with their misuse.

Needs and Priorities for Next Round of Model Evaluations

Access to Models and Evaluation Tools

The next session underscored the crucial need for independent researchers to have access to both proprietary (closed-source) models and robust evaluation tools. This access could take various forms, such as black-box testing, white-box testing, fine-tuning, or support from model developers to facilitate researcher interaction.[1] Currently, legal and contractual frameworks governing data sharing and the conduct of evaluations are significant barriers to such access; nondisclosure agreements create opacity over study designs, evaluations already performed on models, and evaluation outputs. This lack of transparency inhibits the ability of the research community to thoroughly assess model capabilities and potential risks. To bridge this gap, participants highlighted the importance of establishing mechanisms for greater visibility into the model development phase—particularly for models planned to be open-sourced because of the inability to control their diffusion once deployed. The formation of a consortium or another independent body can play a crucial role in coordinating on each of these challenges. Such an entity could facilitate discussions, mediate among stakeholders, and help clarify the legal and contractual aspects of running dangerous capability evaluations, thereby streamlining the process for all involved.

Participants highlighted the importance of establishing mechanisms for greater visibility into the model development phase—particularly for models planned to be open-sourced.

Identifying Risks

A core component of this session was dedicated to identifying and categorizing AI-enabled chem-bio

[1] For more information on various forms of access and their implications for AI audits, see Casper et al., 2024.

risks. Participants delineated two primary "buckets" of risk: universally acknowledged risks and risks that emerge contingent on the level of access and application. For instance, a model that could assist in developing nuclear weapons would fall under the category of universally acknowledged risks, but the risks associated with AI-assisted screening for substance toxicity depend on who has access to models and how they are applied. This distinction underscores the complexity of defining and mitigating risks in a field in which the dual-use nature of technologies can blur the lines between beneficial and harmful applications. The conversation also revolved around the potential of AI models to expand access to chem-bio information, lower barriers to understanding, and foster the generation of new hypotheses and knowledge. Although these capabilities offer value to scientific advancement, they also raise concerns about how information unearthed by frontier models could be misused.

Wet Lab and Lab Automation Evaluations

Concerns About Wet Lab Validation of Evaluation Outputs

In the third session of the track, participants delved into the intricacies of wet lab validation of model outputs, which involves verifying the efficacy of potentially harmful compounds designed by models. Central to the discussions were concerns about the fine line between enhancing the understanding of model capabilities and the potential misinterpretation of such validation efforts as steps toward creating harmful substances. Moreover, the session highlighted apprehensions regarding the validation of models' potential to facilitate chem-bio threats, emphasizing the dilemma that dissemination of such evaluative results could inadvertently arm malicious actors with harmful knowledge.

Participants delineated two primary "buckets" of risk: universally acknowledged risks and risks that emerge contingent on the level of access and application.

Track 1 Proposed Actions

Enhance Evaluation Methods to Capture Complex Threat Dynamics

To effectively tackle the complexity of chem-bio threats, it is essential to advance beyond conventional evaluation methods and embrace dynamic and interactive assessment techniques, such as simulation-based tools. These methods, including but not limited to tabletop exercises, are crucial for capturing the nuanced behaviors of threat actors who continually adapt their strategies. By simulating diverse real-world scenarios, these evaluation tools provide a deeper understanding of the evolving nature of threats, thereby enabling the development of more effective and resilient mitigations.

Address Legal and Ethical Concerns of Wet Lab Evaluations

Formulating standards for performing wet lab evaluations would mitigate the concerns identified in the validation process discussions. Specifically, this policy action tackles the challenges of navigating the legal and ethical landscape of chem-bio research, as highlighted by the potential for misinterpretation of validation efforts and the dissemination of sensitive evaluation findings.

Ensure Adequate and Appropriate Access to Models and Evaluation Tools

The "Needs and Priorities for Next Round of Model Evaluations" session highlighted the need for independent researchers to access both proprietary and open-source models, as well as robust evaluation

tools, to effectively assess the capabilities and risks associated with frontier models.

However, the opacity in legal and contractual frameworks, such as nondisclosure agreements, currently hinders independent research by obscuring crucial aspects of evaluations. To address these barriers, efforts are needed to establish mechanisms that provide greater transparency and facilitate discussions among stakeholders.

Track 2: Loss of Control

This track focused on evaluating models for capabilities that exceed the boundaries of developer or user intent—and featured presentations by evaluation organizations METR and Apollo Research. This section explores identifying and mitigating risks associated with model autonomy and deception—a growing concern as models act increasingly independently.

Autonomy (METR)

In this session, METR, formerly ARC Evals, presented their research on autonomy evaluations, followed by an open discussion.

Focus on Autonomous Threats

Unlike threat models centered on AI augmentation of malicious actors, METR assesses risks posed by AI acting independently to execute potentially harmful actions, such as conducting phishing attacks or manipulating digital infrastructure. Importantly, METR argued, this does not require an AI to have harmful goals—a malicious actor might initiate harm by prompting the AI to carry out tasks autonomously. METR suggested that although the full scope of autonomous AI actions' remains to be seen, there is a critical need for preparedness and vigilance to anticipate these developments. This session demonstrated examples of tasks that METR developed to test for relevant autonomous capabilities—including the implementation of machine learning research; improvements to AI agent scaffolding; and the management of large, complicated codebases. Furthermore, METR showcased an example of a model capability: a "capture the flag" task that required reverse-engineering C code, which otherwise requires a domain expert 30 minutes to complete.

A malicious actor might initiate harm by prompting the AI to carry out tasks autonomously.

Evaluation Methodologies

METR presented its methodology for evaluating AI systems, which focuses on specificity, objectivity, and cost-effectiveness. The session covered tasks, methodological guidance, and evaluation checks aimed at supporting test validity. Tasks, structured around METR's Task Standard, measure an AI system's ability to perform activities relevant to threat models. The Task Standard, according to METR, aims to increase uniformity, scalability, and reproducibility and decrease duplication of work. METR's methodological guidance emphasized the importance of eliciting capability and removing spurious failures, such as those caused by ethical refusals or tooling limitations (see the "Unlocking AI Capabilities" section). Evaluation checks, including reviewing outputs for spurious failures, were also proposed by METR to ensure the integrity of test results and avoid underestimating agent capabilities. Altogether, METR shared, this evaluation methodology seeks to continuously measure dangerous model capabilities, allowing for the development of scaling laws and appropriate mitigations.

Unlocking AI Capabilities

Regarding unlocking model capabilities, participants recognized the difficulty of guaranteeing that AI systems lack specific capabilities. METR highlighted the importance of identifying spurious failures, in which models fail not because of capability, but because of bugs, limitations of tooling, poor prompting, or model

As AI models grow in sophistication, Apollo Research believes that their capacity for deception might also grow.

refusals on the basis of ethics or claimed inability. Further post-training enhancements, METR suggested, might remove these limitations, especially in threat scenarios in which weights are stolen or open-sourced, thus invalidating initial capability evaluations. METR suggested adding a safety buffer to thresholds for policy action as one way to address this risk.

Deception (Apollo Research)

In this session, Apollo Research presented its research on strategic deception evaluations, followed by an open discussion.

The Amplifying Risk of Deception

Apollo Research's presentation on AI deception began with articulating how deception is a risk amplifier, including for misuse and chemical, biological, radiological, and nuclear risks, through obfuscation—impeding governance, oversight, and transparency. Apollo Research argued that deceptive behaviors could emerge from two primary sources: intentional malicious acts by bad actors (e.g., scams) or unintended outcomes of complex AI systems (e.g., misalignment). Furthermore, as AI models grow in sophistication, Apollo Research believes that their capacity for deception might also grow, further hampering efforts to identify and counteract AI deception.

Evaluating for Deception

Apollo Research's demonstrations provided an existence proof for AI to engage in deception. First, a "research assistant" GPT presented deception resulting from AI misuse. Here, a model deployed to support government officials was fine-tuned to gather sensitive information illicitly and deceive its users. The second demonstration presented deception as a result of unintended outcomes of complex AI systems, with no fine-tuning or steering. Here, an insider-trading demo depicted an AI model operating as a stock trading agent, which, under specific scenarios, resorted to insider trading and subsequently concealed its activities—in some cases doubling down on its deception (Scheurer, Balesni, and Hobbhahn, 2023). These demonstrations, Apollo Research argued, underscored the importance of employing a comprehensive suite of tools for eliciting and detecting deceptive behaviors in AI models—including direct elicitation, honeypots, and interpretability methods (Casper et al., 2024). Apollo Research advocated for a "Science of Evals" to increase the robustness of methods for detecting AI deception—calling for drawing from literature and expertise on other such AI risks as fairness and bias, as well as assurance ecosystems for other high-stake systems (e.g., commercial aviation) to address the ethical, technical, and operational challenges in an effective AI assurance ecosystem (Apollo Research, 2024).

Track 2 Proposed Actions

Refine Evaluation Techniques for Detecting Deception in AI Systems

After Apollo Research's presentation on early findings of AI deception, session participants discussed the necessity to refine and strengthen evaluation methods for detecting deceptive AI systems. This session's demonstrations of AI deception, Apollo Research argued, highlight the need for a wide variety of technical strategies to uncover and understand model deception. According to Apollo Research, proposed frameworks for evaluations considering the inner workings of models, enabled by greater access to models, would provide more-thorough insight into the causal drivers of model behavior. This, Apollo Research believes, would help techniques for detection keep pace with AI's evolving complexity and capabilities for obfuscation.

Track 3: Risk-Agnostic Methods

This track set out to produce a risk-agnostic methodological framework and strengthen the robustness of model evaluations. The first session explored evaluation methods—such as red teaming, automated benchmarking, and sophisticated task design. The session that followed shared best practices for evaluating model capabilities.

Evaluation Methods: From Red Teaming to Automated Benchmarking

Framework for Evaluation

In the evolving landscape of model development, robust evaluation methods are paramount to assess the potentially dangerous capabilities of frontier models. A RAND facilitator presented a draft table breaking down prominent model evaluation methods (e.g., red teaming, multiple choice benchmarking) by key attributes (e.g., repeatability, depth, generalizability) to stimulate discussion. Table A.1 (see the appendix) presents my interpretation of a new risk-agnostic framework of evaluation methodologies, structured to remain agnostic to threats, based on the discussion in this session. This approach acknowledges the importance of well-crafted threat models for model evaluation; however, the process of determining a threat model fell beyond the purview of this session. The group's focus, instead, was on mapping the space of methodologies that can be generally applicable across risks or threat types. By dissecting evaluation methodologies into manageable dimensions, the session aimed to facilitate a greater understanding of the strengths and limitations of different evaluation methods for establishing evidence of model capabilities.

Enhancing Access to AI Evaluation Tools While Preserving Integrity

The session also addressed challenges in broadening access to evaluation tools (e.g., open-sourcing to foster innovation and community scrutiny) while maintaining evaluation integrity, robustness, and validity. The conversation underscored the importance of hold-out tasks that remain unknown to models until the time of evaluation and of exploring the use of cryptographic hashing to safeguard the integrity of evaluation data. Such approaches as these are critical to prevent the incorporation of evaluation data into the training datasets of the models they are intended to evaluate, which could compromise the robustness of future evaluations. Ensuring that evaluations do not become part of model training data necessitates meticulous consideration of best practices for publishing and sharing evaluations, striking a delicate balance that fosters transparency and maintains the integrity and robustness of model evaluation.

Ensuring Evaluation Robustness and Validity (METR)

Key Questions for Evaluation Robustness

For developing model evaluations, METR highlighted a series of questions that support evaluation robustness. A selection of these questions serves as a foundation for constructing effective evaluations while acknowledging methodological limitations.

- **What is the evaluation measuring?** If an evaluation is being used to inform a particular action, then it needs to be faithful to the relevant threat model. For example, an evaluation to assess whether a model is safe to open-source needs to assess the safety of the model under fine-tuning.

Ensuring that evaluations do not become part of model training data necessitates meticulous consideration of best practices for publishing and sharing evaluations.

- **Does the evaluation scale well? How is it intended to be run?** An evaluation might be valid only under certain assumptions—for example, that the dataset has not appeared in a model's pre-training corpus or the model has certain tools available.
- **What are the warning signs of a problem with the evaluation?** Recognizing warning signs of potential evaluation issues and understanding the causes of misleadingly high or low scores are crucial for maintaining evaluation integrity and accuracy.

Model Versus System Evaluation

Session attendees also addressed the distinction between evaluating stand-alone models that could be used or deployed adversarially and evaluating the broader systems in which these models are embedded when deployed. Depending on the threat model being considered, participants suggested that either of these possibilities might be appropriate, but it is important to evaluate thoughtfully.

Track 3 Proposed Actions

Develop a Framework for Robust Evaluation

Discussions on risk-agnostic methodological frameworks emphasized the importance of holistically exploring the space of evaluation methods to understand potential harms and highlighted a need for a framework for robust evaluation. The session's focus on crafting a methodological framework that transcends specific threats lays the groundwork for the development of such a framework.

Session participants suggested using such strategies as employing hold-out tasks and cryptographic techniques to safeguard evaluation integrity.

Implement Measures to Guarantee the Robustness and Validity of AI Evaluations

The challenges identified with democratizing model evaluation while preserving evaluation integrity inform the policy action to implement measures to guarantee the robustness and validity of AI evaluations. Session participants suggested using such strategies as employing hold-out tasks and cryptographic techniques to safeguard evaluation integrity. These approaches aim to prevent the contamination of model training datasets with evaluation data, ensuring the future robustness of evaluations.

Develop and Disseminate Best Practices

Informed by METR's key questions for ensuring evaluation robustness, this action seeks to establish and share best practices for effective evaluations. By clarifying what evaluations measure, ensuring scalability, and identifying warning signs of evaluation problems, this initiative aims to equip stakeholders with the knowledge to design, select, and interpret evaluations effectively.

Track 4: Collaboration and Coordination

This section summarizes discussions held during the collaboration and coordination track, which brought AI research and development experts into conversation with policy researchers and professionals. These discussions covered key upcoming policy timelines, processes for discerning risk thresholds, and frameworks for responsibly managing future model capabilities.

Policy Timelines and Deliverables

In this session, stakeholders discussed the scope and status of several key AI policy milestones in 2024, including deliverables assigned by the fall 2023 executive order on AI (Biden, 2023). Participants broadly

agreed that 2024 milestones should be only one step in ongoing AI governance efforts and that significant future work and coordination would be required. High hopes were expressed that the National Institutes of Standards and Technology AI Safety Institute Consortium would facilitate this continued engagement, and participants highlighted several open questions that should be addressed in future work:

- How should the timing and venue for publishing evaluation results be determined, and what information should be included to strike an appropriate balance between transparency and responsible disclosure?
- What activities (e.g., evaluations, standards) are best suited for government involvement?
- What evaluation methods are most effective at eliciting harmful model capabilities?
- When should evaluations be executed? How can initial, less costly, assessments reliably indicate the need for more-comprehensive and resource-intensive evaluations?
- What systems, mechanisms, or incentives can be implemented to ensure that evaluators thoroughly explore and reveal the full spectrum of AI capabilities?
- How can the evaluation insights contribute to a risk-benefit analysis for AI systems?

Red Lines and Risk Thresholds

Conceptualizing Risk and Mitigation

This session underscored the importance of preemptive risk management for model development, given the difficulty of predicting future model capability. It highlighted the dual-use and general nature of frontier models, which, although designed for beneficial purposes, can be used by malicious actors to cause significant harm. To ensure the effective development of mitigations, the session identified the necessity of evaluating the quality and accuracy of AI outputs, the breadth of the threat model under consideration, and the operational context of model deployment.

Evolution of Red Lines into Standards

Today, red lines seek to proactively identify risk thresholds for responsible model deployment that, if surpassed, would warrant a significant response. These thresholds, although still under development, might support the creation of standards and best practices. In this session, participants workshopped a path from preliminary exploratory evaluations of models to the formulation of standards. This pathway is characterized by a collaborative effort involving stakeholders across various domains, including academia, industry, and government, who are invited to identify essential ethical and safety components of model development. Such an approach would then guide policy formulation and exploratory evaluations, facilitating frontier models' responsible advancement and deployment. The insights gained from these evaluations would then inform the iterative process of risk assessment, mitigation, and standard formulation. However, reaching a consensus on red lines presents a challenge, given the complex and often ambiguous nature of model capabilities. Achieving unity on these issues necessitates a concerted effort from a broad coalition. Essential to this process are strategies aimed at fostering consensus, such as organizing focused workshops, maintaining ongoing dialogues, and considering the insights and innovations of the private sector.

Reaching a consensus on red lines presents a challenge, given the complex and often ambiguous nature of model capabilities.

Responsible Capability Scaling

Gradient of Risks and Benefits

This dialogue examined the feasibility and implications of establishing clear risk thresholds in AI development within the context of responsible capability scaling (RCS). RCS is a framework leveraged by leading AI labs to manage risks associated with develop-

ing and deploying frontier AI systems; this involves risk assessments, pre-specified risk thresholds, and commitments to mitigations at these thresholds—including pausing development or deployment if necessary. However, given the gradient nature of risks and benefits, the session highlighted the challenge of delineating acceptable risks from unacceptable risks. Participants raised concerns that implementing risk thresholds will require improvements to humanity's ability to understand and measure AI risks and benefits, but they also emphasized the importance of having clear decision boundaries.

Evaluation and Mitigation Strategies

The discussion also revealed the multifaceted nature of AI risk management. Participants emphasized the significance of conducting thorough and regular evaluations to gauge both pre-training improvements and more-elusive post-training advancements. Recognizing the cost and logistical challenges associated with evaluations, the session participants explored an analogy between AI risk assessments and biorisk evaluations and suggested that AI could benefit from adopting similar criteria for reevaluation—such as evaluating after significant increases in the computational power used to train a model or changes in the model's operational environment. Moreover, the dialogue underscored the importance of developing and implementing effective mitigation strategies in a timely manner, either ahead of potential harm or as soon as practically feasible. This entails a shift toward anticipatory governance, which focuses on predicting potential harms and initiating protective measures well in advance of the manifestation of these risks.

Track 4 Proposed Actions

Implement Regular and Timely Evaluations and Mitigations

The takeaway of the evaluation and mitigation strategies section underscores the critical need to implement ongoing evaluations of AI systems—both during and after training—to capture improvements and promptly address emergent risks. Reflecting the session's focus on the importance of anticipatory governance, this policy action is designed to preempt potential misuse by ensuring that mitigation strategies are effective and are implemented, adapting to the evolving capabilities and deployment of frontier models.

Conclusion

Frontier Model Evaluation Science Day underscored the paramount importance of collaborative, interdisciplinary efforts to effectively mitigate the potential dangers of advanced AI systems. The insights gathered from the conference tracks reveal a complex landscape of risks associated with highly capable AI—from autonomy and misuse to chem-bio threats and deceptive capabilities. The recommended policy actions, rooted in the proceedings of this conference, advocate for a multifaceted approach that includes refining evaluation methodologies, enhancing legal and ethical frameworks, ensuring broad access to models and evaluation tools, and fostering a culture of shared responsibility among stakeholders.

Although these measures lay the groundwork for a comprehensive strategy to address the challenges posed by advanced AI, it is crucial to acknowledge that they do not provide immediate solutions to the pressing issues identified in these proceedings. The proposed actions serve as a starting point for anticipatory governance, enabling stakeholders to adapt to the rapid evolution of AI and develop targeted interventions. However, further research, collaboration, and targeted initiatives are necessary to directly tackle the specific risks and challenges highlighted during the conference. Through fostering of an environment of open dialogue, rigorous evaluation, and proactive policy formulation, navigation of the complex dual-use nature of AI can begin. Although significant challenges remain, a collaborative and adaptive approach offers a path forward to responsibly harness the benefits of AI while implementing safeguards against its risks and to work toward a future in which AI advancements contribute positively to society and global security.

APPENDIX

Risk-Agnostic Evaluation Methodological Framework Incorporating Track 3 Insight

Table A.1 displays the author's interpretation of a new risk-agnostic methodological framework of evaluation methodologies as discussed by participants during the Track 3 session.

TABLE A.1

Risk-Agnostic Methodological Framework Discussed in Track 3 Session

Methods	Judge	Repeatability	Breadth	Speed	Depth	Generalizability	Cost
Single output							
Multiple choice questions	Human	✓	✓	~	X	X	~
	Automated	✓	✓	✓	X	X	✓
Open-ended response	Expert	~	✓	~	~	~	X
	Non-expert	~	✓	~	~	~	~
	Automated	~	✓	✓	~	~	✓
Series of outputs							
Benchmarking	Expert	X	X	X	✓	~	X
	Non-expert	X	X	X	✓	~	X
	Automated	X	X	X	✓	~	~
Red teaming	Expert	X	X	X	✓	~	X
	Non-expert	X	X	X	✓	~	X
	Automated	X	X	X	✓	~	~

NOTE: Symbols within the table (✓, ~, and X) illustrate the relative strengths and limitations of each evaluation method as determined by the author. ✓ indicates characteristics for which the method performs well, ~ signifies areas of moderate performance or concern, and X highlights significant limitations or challenges. This qualitative assessment, although informative, is not exhaustive and does not encompass all evaluation methods.

References

Apollo Research, "We Need a Science of Evals," Apollo Research blog, January 22, 2024. As of July 23, 2024: https://www.apolloresearch.ai/blog/we-need-a-science-of-evals

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

Casper, Stephen, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, June 5, 2024.

Scheurer, Jérémy, Mikita Balesni, and Marius Hobbhahn, *Large Language Models Can Strategically Deceive Their Users When Put Under Pressure*, arXiv, arXiv:2311.07590, November 9, 2023.

About These Conference Proceedings

On February 9, 2024, RAND hosted Frontier Model Evaluation Science Day at its Arlington, Virginia, office to address the rapid pace of artificial intelligence (AI) advancements and the consequent need for developing evaluation methodologies to keep pace with them. The gathering brought together a diverse assembly of experts who work on developing the science of evaluations, focusing on national security risks—including researchers from independent evaluation organizations, individuals from frontier AI labs, and representatives from the governments of the United States and the United Kingdom. Their collective knowledge was pivotal in identifying current limitations of, potential partnerships in, and the necessary expertise required for advancing the field of AI safety and evaluation. These proceedings capture insights and proposed actions from this gathering, intending to inform—not endorse—possible next steps for developing robust policy-relevant evaluation science, filling research gaps, and enriching public understanding of the evolving landscape of AI and national security.

Technology and Security Policy Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. 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.

Acknowledgments

I extend my deepest gratitude to all who contributed to Frontier Model Evaluation Science Day's success and the subsequent development of these proceedings. My heartfelt thanks go to Ella Guest, Sarah Gebauer, and Roshon Gibson, whose meticulous planning was the bedrock of this event's success. To ensure that the content accurately reflects the richness of the discussions, I worked closely with session attendees and facilitators, whose insights and leadership across all tracks were instrumental in guiding the day's conversations. Finally, the constructive feedback from peer reviewer James Ryseff has been crucial in enhancing the quality of these proceedings, for which I am immensely grateful.

About the Author

Anton Shenk is a quantitative research assistant at RAND. He examines the economics of emerging technologies and has contributed to work including the study of biological weapons enhanced by advanced AI, the future of economic warfare, and when AI could be deceptive with its users. Anton holds B.S. degrees in mathematics and economics.

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. **RAND**® is a registered trademark.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, please visit www.rand.org/pubs/permissions.

For more information on this publication, visit www.rand.org/t/CFA3429-1.

Cover photo: Vitalii Gulenok/Getty Images

© 2024 RAND Corporation

www.rand.org

CF-A3429-1