RAND

Perspective

KARSON ELMGREN, GAURAV SETT, EVERETT SMITH

Considerations on AI Model Red-Teaming and Standards

Insights to Inform a Request for Information (RFI) Related to the National Institute of Standards and Technology (NIST)'s Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Perspective

Considerations on AI Model Red-Teaming and Standards

Insights to Inform a Request for Information (RFI) Related to the National Institute of Standards and Technology (NIST)'s Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

KARSON ELMGREN, GAURAV SETT, EVERETT SMITH

February 2024

RAND

Published by the RAND Corporation, Santa Monica, Calif.

About This Paper

RAND's Technology and Security Policy Center (TASP) conducts research on dual-use technologies, such as artificial intelligence (AI), and their relevance to the national security of the United States. As part of this work, TASP has investigated evaluations and red-teaming of AI systems, including conducting evaluations of AI systems' ability to enable the execution of biological weapon attacks, and organized a workshop on red-teaming as a method to identify capabilities and risks of AI models. This paper provides insights from RAND experts in response to the National Institute of Standards and Technology (NIST)'s December 2023 Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence.[1] This paper is undergoing further peer review as a part of RAND's research quality assurance process; an expanded version will be published to www.rand.org in the near future.

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.

Acknowledgments

The authors thank Liz Nguyen, Aris Richardson, Ella Guest and Ashwin Acharya for their technical insights and assistance with this paper.

[1] NIST, U.S. Department of Commerce, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)," *Federal Register*, Vol. 88, No. 244, December 21, 2023b.

Contents

Considerations on AI Model Red-Teaming and Standards

In this paper, the authors provide a series of expert insights in response to the National Institute of Standards and Technology (NIST)'s recent Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order (EO) Concerning Artificial Intelligence.[2] Our comments are organized in response to some of the topics requested under the assignments listed in sections 1 and 3 of the RFI. The assignments (quoted from the RFI) are provided in italics leading into each topic discussion throughout the paper. The assignments and topics discussed in this paper are

- developing guidelines, standards, and best practices for ai safety and security
  - current standards and norms
  - the role of different actors
  - tooling for identifying impacts and mitigations
  - risk management documentation
  - appropriate scope of red-teaming
  - degree of access
  - internal review
  - external review
  - limitations of red-teaming
- advance responsible global technical standards for ai development
  - systems where standards are most impactful.

[2] NIST, U.S. Department of Commerce, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)," *Federal Register*, Vol. 88, No. 244, December 21, 2023b.

Developing Guidelines, Standards, and Best Practices for AI Safety and Security

In this section we discuss two assignments and ten selected topics listed in the NIST RFI. Each assignment from the RFI is offered in italics before the topic discussions.

"E.O. 14110 Sections 4.1(a)(i)(A) and (C) direct NIST to establish guidelines and best practices in order to promote consensus industry standards in the development and deployment of safe, secure, and trustworthy AI systems."

Current Standards and Norms

One emerging norm among developers of frontier large language models (LLMs) is the publication of documents describing how the developer will condition their development and deployment decisions on estimations of risk related to models, referred to variously as a "responsible scaling policy,"[3] "risk-informed development policy,"[4] or "preparedness framework."[5] This documentation can be understood as an instantiation of the Risk Management Framework (RMF) Govern function in that it involves determining the necessary level of risk management activities (Govern 1.3).[6] Taking this step would also correspond to Map 3.2 in the RMF in documenting possible risks associated with a system's capabilities, Measure 2.3 and 2.6 in evaluating those capabilities and risks, and Manage 1.3 in clarifying responses to detected risks.[7]

Policies for responsible capability scaling aim to tie safety and security best practices for artificial intelligence (AI) labs to the level of danger associated with their AI models. In some cases, ensuring sufficient safety and security may be impossible absent technical breakthroughs or significant improvements in AI lab security. If so, pausing the scaling of AI models until such breakthroughs occur may be necessary to manage risks. This is consistent with NIST's AI RMF 1.0, which recommends, "In cases where an AI system presents unacceptable negative risk levels—such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present—development and deployment should cease in a safe manner until risks can be sufficiently managed."[8] This approach would be warranted depending on the of the severity of the risks and may not be appropriate for other safety issues.

[3] Anthropic, "Anthropic's Responsible Scaling Policy," September 19, 2023.

[4] OpenAI, "OpenAI's Approach to Frontier Risk," section on "Risk-Informed Development Policy," October 26, 2023d.

[5] OpenAI, "Preparedness Framework (Beta)," December 18, 2023e.

[6] NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0),* U.S. Department of Commerce, NIST AI 100.1, January 2023a.

[7] NIST, 2023a.

[8] NIST, 2023a.

Although existing, publicly documented policies regarding risk-management activities are a valuable step, these policies could be improved by being made more detailed and explicit. For instance, companies could clarify how they identify risks to be assessed and managed, at what points they will engage in which risk-management activities (either procedural, such as a predeployment evaluation, or substantive, such as measures tied to quantitative increases in training compute), the precise process and results of their risk estimation (potentially in the form of explicit probabilistic risk assessments[9]), how they will decide what risk levels are tolerable, and how risks will be continuously managed throughout development and deployment.

The Role of Different Actors

AI developers often have significantly more information relevant to evaluating their systems than external parties, whether deployers, end users, or third-party evaluators. For example, eliciting the greatest extent of capabilities from a model often requires detailed experience with fine-tuning, prompt engineering, and building scaffolding for that particular model, which technical staff internal to the developer possess to the greatest degree. As a result, downstream deployers and users should not be solely responsible for evaluating systems for their applications and use cases; upstream developers must share information related to safety and evaluations with other actors in the ecosystem (such as downstream partners), and they must be primarily accountable for the safety of their systems. Government should also establish processes to learn from industry to develop standards that keep pace with the changing technical landscape. One possible model for doing this comes from bank supervision in the financial industry, which is complex and fast-changing, similar to AI.[10]

Tooling for Identifying Impacts and Mitigations

Identifying and mitigating the impacts of AI systems will require monitoring those systems' behaviors in ways that provide data useful for understanding their effects in the real world. Both identification and mitigation of deleterious effects are greatly facilitated when systems are made available in a controlled manner, such as via an application programming interface (API). When model weights are available to download from the internet, understanding the impacts of the model in any detail becomes much more difficult, as deployments of the model are difficult to trace and records of their interactions unavailable. In many cases, the risks of a given system are low enough that such monitoring is too costly relative to the societal benefit that broader access

[9] Martin E. Hellman, "Probabilistic Risk Assessment," in James Scouras, ed., *On Assessing the Risk of Nuclear War*, Johns Hopkins Applied Physics Laboratory, 2021.

[10] For an overview of bank supervision, see Thomas Eisenbach, Andrew Haughwout, Beverly Hirtle, Anna Kovner, David Lucca, and Matthew Plosser, "Supervising Large, Complex Financial Institutions: What Do Supervisors Do?" Federal Reserve Bank of New York, February, 2017. To understand how supervision was designed to address the complexity and rapid change of risk profiles, see Lisa M. DeFerrari and David E. Palmer, "Supervision of Large Complex Banking Organizations," Federal Reserve Bulletin, February, 2001.

to the system brings. However, in some cases, risks may be high enough that more-controlled access is required. In these cases, developers should establish infrastructure for logging and analysis of system behaviors, along with incident response functions to quickly understand the extent of harm and implement mitigations when incidents are detected. These could include restricting certain (categories of) users' access, restricting frequency of access to the model (such as number of API queries or prompts), deactivating certain capabilities or features (for example, model agent features, such as generating and assigning tasks to copies of itself, or model access to external tools), restricting use cases (especially high-risk use cases, such as in safety-critical domains), or shutting down a system (whether through removal from production, physical disconnection of electricity to relevant data centers, or even deletion of model weights).[11]

Structured access tools based on privacy-enhancing technologies could help enable detection and monitoring of impacts of AI systems while protecting individual privacy and corporate intellectual property. Monitoring more-diffuse and society-wide impacts of AI systems would often benefit from access to aggregated data from many individuals' interactions with AI systems—for instance, assessing the degree to which chatbots could be influencing users' views. However, aggregating this data in unencrypted form would constitute a large-scale invasion of privacy. Privacy-enhancing technologies could be usefully applied to enable civil-society and public-sector actors to study the effects of AI systems in the real world without impinging upon citizens' privacy.[12]

Risk Management Documentation

For a regulator to evaluate whether a company's risk-management processes are sufficient or not, it would be useful to have explicit documentation of how companies are identifying, evaluating, and monitoring risks, as well as their risk tolerance and how they have attempted to reduce risks to stay within that tolerance. This information can be used in a "safety case"[13] that argues a company has effectively reduced risk to a given threshold and that the chosen threshold reasonably balances societal interests.

Evaluating AI Systems

Evaluations of AI systems would benefit from the creation of research APIs that provide greater access than is currently available via commercial API services from major AI developers. In particular, research APIs could provide features related to sampling (basic sampling, logits and logprobs, and selecting/modifying sampling algorithms), fine-tuning (supervised fine-tuning, custom loss functions, and reinforcement learning), inspecting and modifying model internals

[11] Joe O'Brien, Shaun Ee, and Zoe Williams, *Deployment Corrections: An Incident Response Framework for Frontier AI Models,* Institute for AI Policy and Strategy, September 30, 2023, pp. 11–13.

[12] For example, see OpenMined, "Privacy-Preserving Data Science, Explained," blog post, May 19, 2020.

[13] David J. Rinehart, John C. Knight, and Jonathan Rowanhill, *Understanding What It Means for Assurance Cases to "Work,"* National Aeronautics and Space Administration, CR-2017-219582, April 2017.

(including parameters, activations and attention, gradients, embeddings and residual streams, and custom function insertion), and providing access to data about models such as training data, model snapshots from the training process, information about the model (such as architecture and size) and to model families.[14] Of these features, few are relatively standard in the industry today.

Evaluations can be made more useful by ensuring they are able to measure a wide range of performance on the task. A wide difficulty range ensures that the evaluation is useful now or soon will be with even relatively weak models, remains relevant as models become more capable, and is able to give significant advance warning of a potentially dangerous threshold.[15] Methods for evaluating misuse (such as cyber actions or chemical, biological, radiological and nuclear defense), in particular, can share some common aspects where different kinds of misuse risks share similar steps to harm, such as jailbreaking to get around safety guardrails and various stages in the weaponization chain for chemical or biological attacks.

Furthermore, as model capabilities increase, evaluations will need to expand correspondingly. Systems acting in the world in increasingly autonomous and complex ways will pose risks beyond misuse, including structural risks of multiagent interaction and misalignment risks. Evaluations will also need to become robust to active deception as models of increasing capability may learn to game evaluations, performing safely during testing or training but preserving unsafe behavior in deployment.[16]

"E.O. 14110 Section 4.1(a)(ii) directs NIST to establish guidelines, . . . including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems."

It should be emphasized that a specific method—such as red-teaming, in the sense of adversarial activities carried out on a system with the support of the system's host organization or owner—should not be the exclusive focus of risk-management activities. Although a crucial tool, red-teaming is not a replacement for a comprehensive risk identification, assessment, and management process. Furthermore, in the field of AI, the term *red-teaming* is often used to refer to a variety of activities related to testing and evaluation of AI systems or their components for

[14] See Appendix A of Benjamin S. Bucknall and Robert F. Trager, *Structured Access for Third Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements*, white paper, Oxford Martin AI Governance Initiative, October 2023.

[15] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Moan Kolt, et al., "Model Evaluation for Extreme Risks," *arXiv*, September 22, 2023, p. 11.

[16] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al., "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training," *arXiv*, 2024.

potential risks.[17] Our comments here focus especially on red-teaming in the narrow sense but also touch on broader processes of evaluation and review within which red-teaming activities must be embedded to adequately constitute a risk-management system.

Use Cases for Red-Teaming

Red-teaming is most useful as a tool to assess a risk that has already been identified as potentially present in an AI system. For instance, companies building LLMs have identified enabling offense cyber operations or chemical or biological weapon development as potential risks of LLMs that red-teaming. Red-teaming studies, such as a study conducted by RAND on biological weapon attack planning, can help to assess these risks.[18] Red-teaming, in the form of attempted jailbreaking, can also be helpful for measuring the effectiveness of risk mitigation measures, such as training to refuse potentially dangerous requests. On the other hand, red-teaming is not necessarily the best tool to identify previously unknown risks—although new potential risks may be uncovered during some forms of exploratory testing, the red-teaming approach is conventionally focused on exploring all possible means to achieve specific harmful effects, rather than identifying all possible harmful effects that might result from a system. Risks related to how a system interacts with society as a whole, such as structural risks,[19] are also outside the scope of what can be effectively illuminated by red-teaming, which focuses on interactions between a single actor (either an individual or a coordinated group) and an AI system.

Degree of Access

Evaluations of AI models, including red-teaming, depend on, at the very least, the ability to repeatedly sample from a model. However, fulsome evaluations of models may also depend on significantly more access than is currently provided by model developers.[20] Some key forms include access to information about the model (for example, which of a family of models is being sampled from or information about pretraining datasets, model size, and fine-tuning processes), access to a full family of models, including continued support for specific models in a family (that is, version stability and back-compatibility), access to output logits, and access to the means to choose and modify sampling algorithms and to fine-tune models (and information

[17] For more discussion of the strengths and limitations of red-teaming, see Marie-Laure Hicks, Ella Guest, Jess Whittlestone, Jacob Ohrvik-Stott, Sana Zakaria, Cecilia Ang, Chryssa Politi, Imogen Wade, and Salil Gunashekar, *Exploring Red Teaming to Identify New and Emerging Risks from AI Foundation Models*, RAND Corporation, CF-A3031-1, 2023, pp. 10–11, 16.

[18] Christopher A. Mouton, Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks*, RAND Corporation, RR-A2977-1, 2023.

[19] Remco Zwetsloot and Allan Dafoe, "Thinking About Risks From AI: Accidents, Misuse and Structure," *Lawfare,* February 11, 2019.

[20] Bucknall and Trager, 2023.

about the fine-tuning process). Current offerings of proprietary models via APIs are generally limited to the ability to sample from a model, and sometimes the ability to choose the sampling algorithm and some forms of fine-tuning access. Limited information is typically provided about the model being sampled—and, sometimes, the model offered under a particular name is changed, or access is removed in ways that disrupt the ability for external parties to conduct research on a specific version of a model.

Internal Review

Internal review throughout the model development and deployment life cycle within AI companies is a requisite part of a risk management ecosystem because the unique affordances of internal teams allow deeper kinds of review that external teams are unable to provide. Some observed practices for effective internal review at AI companies include the following:

- Red-teaming and review activities should be proportional to the expected general level of risk that a system poses. Before risk assessment activities are conducted, staff should document their expectations about likely risks that a model may pose based on information regarding other models in the same family, the quantity or quality of data and compute used for training, or the nature of expected downstream application scenarios (for example, in high-risk domains or potentially affecting very large numbers of people) to determine allocation of resources to red-teaming and review.
- Risk identification, assessment, and mitigation activities should be integrated as early as possible in the AI life cycle. For example, training datasets should be subject to examination for data poisoning and for data that should be excluded from pretraining (for example, biological data relevant for bioweapons).[21] In some cases, implementing mitigations at later stages of the cycle may be much more costly or altogether unfeasible.
- Personnel with a variety of expertise across technical, sociotechnical, and nontechnical fields—such as machine learning; responsible AI; systems safety; social sciences; and risk-specific subject matter, such as biosecurity and cybersecurity—should be involved in red-teaming and review processes. Red teams should include demographically diverse participants with a broad set of ethnic, geographical, and linguistic backgrounds. To most accurately simulate ultimate deployment conditions, red teams should include expertise in the threat actors and adversaries envisioned according to the threat models identified.
- Companies should document who was involved in internal red-teaming, their expertise, and details of the red-teaming activities (such as prompts and corresponding outputs) during a particular stage of development.
- Any evidence discovered during red-teaming that is deemed an early warning sign of certain risks or otherwise relates to the company's risk-management policies (such as observation of model behavior that triggers further review according to the company's responsible scaling policy) should be documented in detail.
- Companies should collect detailed information regarding results from safety measures (such as safety fine-tuning or reinforcement learning from human feedback (RLHF) to facilitate understanding of the effectiveness of safety methods.

[21] Hicks et al., 2023, p. 10.

- The purpose of red-teaming and the decisions that its outputs will inform should be made clear and explicit. Taking this step will help ensure that red-teaming processes are designed to deliver suitable outputs. This step also provides transparency and accountability to act on findings.
- Risks from disclosures or leaks of red-teaming information, which could highlight vulnerabilities of a system, should be taken into account when handling the results. By simulating malicious actors, red-teamers are "doing their homework for them" and should avoid disclosure of maliciously exploitable information—especially where the system owner is not able to fully mitigate the vulnerability. Accordingly, personnel involved in red-teaming must be worthy of trust to handle such information.

External Review

External scrutiny is a crucial tool that can be applied across the development, predeployment and postdeployment phases of the AI life cycle. In a recent workshop facilitated by RAND Europe and the Centre for Long-Term Resilience on the role of red-teaming in identifying risks from AI, participants concluded that external red-teaming is ultimately more important than internal red-teaming, but that external red-teaming only postdeployment (when the system is already available to a significant number of external parties) was not sufficient.[22]

Many of the same best practices as for internal red-teaming also apply to external red-teaming. However, third parties also have some unique requirements to be effective, as listed here:

- During development, third parties can help to forecast risk-relevant features of a system to help inform decisions about whether and how to develop the system. The development phase is a particularly key node in risk-reduction activities because some issues will become more difficult and/or costly to resolve at later phases Information-security practices must also be in place from the inception of a project to be fully effective, so external review can help ensure that these are adequate, considering the expected nature of the system. External review for signs of risks could also be applied during the training process to guide decisions on whether to continue to modify the development process.
- In the predeployment phase, external red-teaming should explore how a model could perform in the real world by attempting to break safety guardrails and elicit dangerous behavior. External parties can also help identify information to be shared with downstream stakeholders to understand and mitigate its risks in their application.[23]
- Postdeployment, external parties should be involved in understanding a system's behavior and impacts in the real world. Assessment of the risks that a system poses will evolve over time as users elicit behavior that was not discovered during testing and as more information becomes available about the system's interactions with the world. This is especially the case if the system's capabilities are extended, such as via integration with other software tools or scaffolding that allows the model to behave as an agent.

[22] Hicks et al., 2023, p. 10.

[23] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al., "Rethink Reporting of Evaluation Results in AI," *Science*, Vol. 308, No. 6641, April 13, 2023.

- To thoroughly understand a system's risks, third parties should search actively and creatively to discover model capabilities, including red-teaming to break safety guardrails and exploration to elicit latent capabilities, rather than merely implementing a set checklist. Benchmarks, while useful, are still in development for many important capabilities, are not fully informative about model behavior, can relatively easily be gamed, and may fail to reflect real-world deployment conditions.
- External parties must have appropriate independence from developers to avoid incentives to produce findings that are favorable from the standpoint of the company. In particular, the developer must not have full influence over the selection and compensation of third-party reviewers, what access they are given, the scope and methods of their review, and what happens after review.
- Third parties must have sufficient resources in the form of time, money, and compute to provide a high-quality review. Adequate time to conduct thorough red-teaming and review is essential. GPT-4 received six months of predeployment evaluations, considered a relatively long period for review of an AI system at that time.[24] For context, in other domains, such as pharmaceutical and aviation, novel products sometimes undergo multiple years of evaluation before being made available on the market. Review should be made as efficient as possible, but thorough review of novel, complex technological products will sometimes require significant time. Third parties must also have sufficient financial resources to attract requisite expertise in relevant domains, including machine learning and risk-specific subjects, and access to enough compute to conduct a large number of assessments, including fine-tuning, at pace.
- As with internal review, external review should be proportional to the expected scope and magnitude of risks posed by a system.
- As with internal review, to ensure holistic red-teaming coverage, external review should include participation from a variety of experts in technical, sociotechnical, and nontechnical fields, such as machine learning, social sciences, domain-specific subject matter, and relevant threat actors. An external red team should also seek to include participation from as diverse a set of demographic perspectives as possible.

One acute requirement for third parties to provide effective external review and red-teaming is access to systems and information across the AI life cycle, detailed as follows:[25]

- Third parties must have access to base forms of a model, without such safety features as safety fine-tuning or safety classifiers, to understand all possible risks of the model, including if malicious actors were to exfiltrate a form of the model without safeguards.
- The ability to fine-tune models is required for third parties to understand the full space of possible risks from a model, given that a certain foundation model will be frequently fine-tuned by both the original developer and downstream deployers or users for various applications, which could introduce or exacerbate risks.
- Levels of access for review (from black box access only to query the model to white box access to model internals) should be proportional to the level of risk. Where privileged

[24] OpenAI, *GPT-4 System Card,* March 23, 2023b.

[25] For more details, see Bucknall and Trager, 2023.

access is required, appropriate security measures should be in place to ensure that only trusted third-party reviewers are able to access potentially risky systems under review.

- Accessible user interfaces should be provided to third parties so that experts from nontechnical backgrounds can effectively participate in assessments.[26]

Addressing Limitations of Red-Teaming

Red-teaming and capabilities elicitation can only ever prove that a model does produce some behavior under at least some circumstances; these approaches are fundamentally unable to prove that a model will not produce that behavior under other circumstances. To ensure the absence of risks of particularly undesirable nature and magnitude, models will need to be evaluated in ways that rely on fundamental understanding of the model rather than on exploratory testing.

One approach that would enable such assurance is mechanistic interpretability, which seeks to explain a model's behavior using a rigorous mechanistic understanding of processes internal to the model. This approach would be broadly useful for a variety of purposes related to managing risks of societal harm but is currently limited by technical challenges. While the field is still nascent, it has recently produced technical results that show significant promise for applicability to even the largest, most complex models.[27]

Another approach to this issue would be to develop a better science of generalization, or how and when model capabilities generalize between contexts. A better understanding of when properties do and do not generalize would help evaluate whether a model's desirable properties generalize to all cases and whether potentially dangerous properties generalize in ways that pose risks. This could be a useful target for publicly funded research and development.

Especially while our understanding of AI systems is in an early stage of development, the unpredictability of emerging capabilities and risks also entails a need for developers to explain their reasoning and to estimate their confidence in their judgments of model safety. In other industries where developers have a far stronger understanding of their systems than external reviewers, it has become increasingly common for developers to produce a safety case.[28] On the basis of evidence collected from testing and evaluation, developers create structured arguments demonstrating the satisfaction of predetermined risk thresholds, such as a maximum probability of failure or principles such as "as low as reasonably achievable" (ALARA). Such a practice can address limitations of evaluations by providing context necessary for making decisions informed by their results.

[26] As Irene Solaiman points out, "Effective design and user interface must be optimized for experts outside of computer science," (Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," *arXiv,* February 5, 2023, p. 11).

[27] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al., *Towards Monosemanticity: Decomposing Language Models with Dictionary Learning*, Transformer Circuits Thread, October 4, 2023.

[28] Rinehart, Knight, and Rowanhill, 2017. Several companies in the autonomous vehicle industry have also already developed safety cases for their systems.

Advance Responsible Global Technical Standards for AI Development

In this section, we discuss one assignment and one selected topic listed in the NIST RFI. The assignment from the RFI is provided in italics.[29]

> *"E.O. 14110 Section 11(b) directs the Secretary of Commerce, within 270 days and in coordination with the Secretary of State and the heads of other relevant agencies, to establish a plan for global engagement on promoting and developing AI consensus standards, cooperation, and coordination, ensuring that such efforts are guided by principles set out in the NIST AI Risk Management Framework and the U.S. Government National Standards Strategy for Critical and Emerging Technology"*[30]

Systems Where Standards Are Most Impactful

Specifying the class of cutting-edge, broadly capable models that have defined recent AI progress has proven challenging. In the EO, a simple heuristic of floating point operations (FLOP) expended during training has provided guidance. However, specific attention and better safety standards will be required for certain nascent features of these systems, such as the following:[31]

- **Multimodality.** Multimodal AI systems are capable of taking in multiple types of input such as text, images, audio, or video. A current example is GPT-4V.[32] Multimodality expands the possible training data and the deployment contexts for AI systems. The intersection of modalities may result in outcomes different from each individual modality.
- **Tool use.** AI systems can often leverage external tools, allowing them to interact with the internet or software applications. The extensions features of ChatGPT, for example, enables the language model to search Bing and run code.[33] Tools can significantly extend the capabilities of AI systems. While the language model underlying ChatGPT often struggles with simple math, plugins to other software services allow the model to input problems into a specialized math engine and receive reliable answers.
- **Interaction.** The interaction between AI systems may lead to emergent behavior. AI systems working together may be able to accomplish far more than individual systems could by themselves, or they may accomplish tasks in different ways. Research has

[29] NIST, 2023b.

[30] NIST, "AI Risk Management Framework," webpage, undated; White House, *United States Government National Standards Strategy for Critical and Emerging Technology*, May 2023; NIST, 2023b.

[31] Shevlane et al., 2023; Helen Toner, Jessica Ji, John Bansemer, Lucy Lim, Chris Painter, Courtney Corley, Jess Whittlestone, Matt Botvinick, Mikel Rodrigues, and Ram Shankar Siva Kumar, *Skating to Where the Puck Is Going: Anticipating and Managing Risks from Frontier AI Systems*, Center for Security and Emerging Technology, October 2023.

[32] OpenAI, *GPT-4V(ision) System Card*, September 25, 2023c.

[33] OpenAI, "ChatGPT Plugins," blog post, March 23, 2023a.

demonstrated the possibility of interactive simulations between AI agents leading to complex social behavior.[34]

- **Hazardous knowledge.** AI systems that have been specifically trained or otherwise given access to hazardous knowledge may pose far greater risks than other foundation models. Such types of knowledge may include chemical, biological, radiological, nuclear, or cyber weapon production.
- **Social engineering.** An AI system that has the capacity to deceive or persuade users (or other humans with whom it interacts) would pose novel risks. The skill to construct believable lies, predict effects of lies, impersonate humans, shape beliefs, or promote narratives can redefine the systems of human control that developers may have designed. Indeed, it is often the case in cybersecurity that humans are the "weakest link," such as in instances of phishing.
- **Long-term planning.** AI systems that can make sequential plans over long time horizons, typically leveraging long-term memory, may pose difficulties for evaluation and assessment efforts that are constrained in the capacity to observe behavior.
- **Situational awareness.** AI systems that determine whether they are being trained, evaluated, or deployed pose challenges to our understanding. These systems could adapt their behavior accordingly.
- **Self-improvement.** AI systems could become capable of conducting research and development tasks to augment their capabilities. This hinders static analysis of these systems.
- **Self-replication.** AI systems that can copy themselves or produce sub-agents outside the local environment pose unique challenges. This may increase the effectiveness of malicious use, such as in a computer worm, or illustrate that capacity for escaping human containment.

These features all pose challenges to the ability to assess and secure AI systems. They are not necessarily distinct features. Many will interact with one another as progress is made towards generally intelligent agents.

[34] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," *arXiv,* August 6, 2023.

Summary

This paper comprises insights from experts on red-teaming, evaluations, and standards for AI systems. As an ecosystem emerges for managing risks from AI systems, our studies suggest that care should be taken to establish proper institutions for identifying, assessing, and mitigating risks. Such institutions would include norms around transparency and proper involvement of external parties in risk-management activities. In particular, publication of organizational policies regarding risk management and provision of adequate access to external researchers could improve the current ecosystem. Red-teaming has an important role to play as one element in a risk-management process, but must be appropriately resourced to be effective and appropriately scoped to avoid overreliance on a single method. On both the domestic and international fronts, standards for ensuring the safety and security of AI systems should prioritize the likely highest-impact categories of systems, and seek to anticipate future technical evolutions that could magnify those systems' impact.

Abbreviations

AI	artificial intelligence
API	application programming interface
EO	Executive Order
LLM	large language model
NIST	National Institute of Standards and Technology
RMF	Risk Management Framework

References

Anthropic, “Anthropic’s Responsible Scaling Policy,” September 19, 2023. As of February 1, 2024:
https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al., *Towards Monosemanticity: Decomposing Language Models with Dictionary Learning,* Transformer Circuits Thread, October 4, 2023.

Bucknall, Benjamin S., and Robert F. Trager, *Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers’’ Model Access Requirements,* white paper, Oxford Martin AI Governance Initiative, October 2023.

Burnell, Ryan, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al., “Rethink Reporting of Evaluation Results in AI,” *Science,* Vol. 308, No. 6641, April 13, 2023.

DeFerrari, Lisa M., and David E. Palmer, “Supervision of Large Complex Banking Organizations,” Federal Reserve Bulletin, February, 2001.

Eisenbach, Thomas, Andrew Haughwout, Beverly Hirtle, Anna Kovner, David Lucca, and Matthew Plosser, “Supervising Large, Complex Financial Institutions: What Do Supervisors Do?” Federal Reserve Bank of New York, February, 2017.

Executive Order [EO] 14110, “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” Executive Office of the President, October 30, 2023.

Hellman, Martin E., “Probabilistic Risk Assessment,” in James Scouras, ed., *On Assessing the Risk of Nuclear War*, Johns Hopkins Applied Physics Laboratory, 2021. As of February 1, 2023:
https://www.jhuapl.edu/sites/default/files/2022-12/Ch4_Hellman.pdf

Hicks, Marie-Laure, Ella Guest, Jess Whittlestone, Jacob Ohrvik-Stott, Sana Zakaria, Cecilia Ang, Chryssa Politi, Imogen Wade, and Salil Gunashekar, *Exploring Red Teaming to Identify New and Emerging Risks from AI Foundation Models*, RAND Corporation, CF-A3031-1, 2023. As of January 25, 2024:
https://www.rand.org/pubs/conf_proceedings/CFA3031-1.html

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al., "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training," *arXiv,* 2024.

Mouton, Christopher A., Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach*, RAND Corporation, RR-A2977-1, 2023. As of February 1, 2024:
https://www.rand.org/pubs/research_reports/RRA2977-1.html

National Institute of Standards and Technology, "AI Risk Management Framework," webpage, undated. As of February 1, 2024:
https://www.nist.gov/itl/ai-risk-management-framework

National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0),* U.S. Department of Commerce, NIST AI 100.1, January 2023a.

National Institute of Standards and Technology, U.S. Department of Commerce, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)," *Federal Register*, Vol. 88, No. 244, December 21, 2023b. As of February 1, 2024:
https://www.govinfo.gov/content/pkg/FR-2023-12-21/pdf/2023-28232.pdf

NIST—See National Institute of Standards and Technology.

O'Brien, Joe, Shaun Ee, and Zoe Williams, *Deployment Corrections: An Incident Response Framework for Frontier AI Models,* Institute for AI Policy and Strategy, September 30, 2023.

OpenAI, "ChatGPT Plugins," blog post, March 23, 2023a. As of February 1, 2024:
https://openai.com/blog/chatgpt-plugins

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

OpenAI, *GPT-4V(ision) System Card,* September 25, 2023c.

OpenAI, "OpenAI's Approach to Frontier Risk," section on "Risk-Informed Development Policy," October 26, 2023d. As of February 1, 2024:
https://openai.com/global-affairs/our-approach-to-frontier-risk#risk-informed-development-policy

OpenAI, "Preparedness Framework (Beta)," December 18, 2023e. As of February 1, 2023:
https://cdn.openai.com/openai-preparedness-framework-beta.pdf

OpenMined, "Privacy-Preserving Data Science, Explained," blog post, May 19, 2020. As of February 1, 2024:
https://blog.openmined.org/private-machine-learning-explained/

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," *arXiv,* August 6, 2023.

Rinehart, David J., John C. Knight, and Jonathan Rowanhill, *Understanding What It Means for Assurance Cases to "Work,"* National Aeronautics and Space Administration, CR-2017-219582, April 2017.

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Moan Kolt, et al., "Model Evaluation for Extreme Risks," *arXiv,* September 22, 2023.

Solaiman, Irene, "The Gradient of Generative AI Release: Methods and Considerations," *arXiv,* February 5, 2023.

Toner, Helen, Jessica Ji, John Bansemer, Lucy Lim, Chris Painter, Courtney Corley, Jess Whittlestone, Matt Botvinick, Mikel Rodrigues, and Ram Shankar Siva Kumar, *Skating to Where the Puck Is Going: Anticipating and Managing Risks from Frontier AI Systems,* Center for Security and Emerging Technology, October 2023.

White House, *United States Government National Standards Strategy for Critical and Emerging Technology*, May 2023. As of February 1, 2024: https://www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf

Zwetsloot, Remco, and Allan Dafoe, "Thinking About Risks From AI: Accidents, Misuse and Structure," *Lawfare,* February 11, 2019.