Roundtable Report

Skating to Where the Puck is Going:

Anticipating and Managing Risks from Frontier AI Systems

Authors

Helen Toner*
Jessica Ji*
John Bansemer*
Lucy Lim*
Chris Painter
Courtney Corley
Jess Whittlestone
Matt Botvinick
Mikel Rodriguez
Ram Shankar Siva Kumar

***Roundtable Organizers**

CSET CENTER for SECURITY and EMERGING TECHNOLOGY | Google DeepMind October 2023

Executive Summary

The advent of more powerful AI systems such as large language models (LLMs) with more general-purpose capabilities has raised expectations that they will have significant societal impacts and create new governance challenges for policymakers. The rapid pace of development adds to the difficulty of managing these challenges. Policymakers will have to grapple with a new generation of AI-related risks, including the potential for AI to be used for malicious purposes, to disrupt or disable critical infrastructure, and to create new and unforeseen threats associated with the emergent capabilities of advanced AI.

In July 2023, the Center for Security and Emerging Technology (CSET) at Georgetown University and Google DeepMind hosted a virtual roundtable to assess the current trajectory of AI development and discuss measures that industry and governments should consider to guide these technologies in a positive and beneficial direction. This Workshop Report summarizes some of the key themes and conclusions of the roundtable and aims to help policymakers "skate to where the puck is going to be," in the words of ice hockey great Wayne Gretzky.

The rise of LLMs has demonstrated that AI is becoming more general-purpose. Current systems are already capable of performing a wide range of distinct tasks, including translating text, writing and editing prose, solving math problems, writing software, and much more. However, there was broad consensus across roundtable participants that these systems are only one iteration of what are likely to be even more capable systems within the next few years. AI developers are actively working to make these systems more powerful, which in turn increases safety and security concerns. Five ways in which existing AI systems are currently being augmented are **multimodality, tool use, deeper reasoning and planning,** larger and more capable **memory,** and increased **interaction** between AI systems.

In anticipation of new types of risks that current and upcoming models may pose, AI companies have begun undertaking "model evaluations" of their most advanced general-purpose AI models. Such evaluations attempt to identify dangerous capabilities such as **autonomous replication** (a model's ability to acquire resources, create copies of itself, and adapt to novel challenges); **dangerous knowledge** about sensitive subjects such as chemical, biological, radiological, or nuclear weapon production; capacity to carry out **offensive cyber operations**; the **ability to manipulate, persuade, or deceive** human observers; **advanced cognitive capabilities** such as long-term planning and error correction; and understanding of their own development, testing, and deployment (sometimes called **situational awareness**). These capability

evaluations are a productive first step, but should not be seen as a comprehensive approach to managing risks. At best, they can provide an early indication of some potential risks, which could trigger mandatory reporting requirements and additional safety measures.

The risks of these new systems are evident, but responsibility for managing that risk is less so. Several participants noted that the current status quo may place too much of the responsibility for managing risk on industry. Instead, policymakers should work towards achieving a balance between government and industry. Under this approach, governments must have sufficient technical expertise to support independent auditing functions, including the ability to describe and execute evaluations. At the same time, governments should encourage and incentivize industry to advance the science of evaluations and report their results. Third-party testing and auditing organizations can provide important additional capacity, but do not negate the need for governments to increase their own oversight capabilities.

To manage these risks, potential policy levers can be grouped into three categories:

1. creating visibility and understanding,
2. defining best practices, and;
3. incentivizing and enforcing certain behaviors.

Governments should encourage—or perhaps require—private-sector actors to test and evaluate their systems and to report dangerous capabilities and real-world threat intelligence to oversight bodies in order to increase regulators' visibility into the state of play. Policy interventions could also help standardize the testing and evaluation process within the AI pipeline. Options to incentivize or enforce these behaviors could include leveraging government procurement requirements, establishing industry certifications for frontier AI systems, and incentivizing increased transparency of discovered vulnerabilities and reporting of evaluation results by reducing liability for companies that disclose responsibly.

Table of Contents

Executive Summary 1

Introduction 4

Progress and Trends in AI 7

- Advancements in General-Purpose AI 7
  - Implications 9

Potential Concerns and Indicators 10

Responsibility for Managing Risk 13

- Key Variables for AI Risk Management 14

Levers to Manage Risk 15

Conclusion 18

Authors 20

Acknowledgments 20

Endnotes 21

Introduction

AI is experiencing a moment of profound change, capturing unprecedented public attention and becoming increasingly sophisticated. As AI becomes more powerful, and in some cases more general in its capabilities, it may become capable of posing novel risks in domains such as bioweapons development, cybersecurity, and beyond. Two features of the current AI landscape are especially challenging from a policy perspective: the rapid pace at which research is advancing, and the recent development of more general-purpose AI systems, which—unlike most AI systems, which are narrowly focused on a single task—can be adapted to many different use cases. These two elements add new layers of difficulty to existing AI ethics and safety problems.

In July 2023, Georgetown University's Center for Security and Emerging Technology (CSET) and Google DeepMind hosted a virtual roundtable to discuss the implications and governance of the advancing AI research frontier, particularly with regard to general-purpose AI models. The objective of the roundtable was to help bridge the gap between the state of the current conversation and the reality of AI technology at the research frontier, which has potentially widespread implications for both national security and society at large. Throughout the discussion, participants noted both the challenge of predicting how frontier AI will progress and the importance of anticipating possible futures, likening this situation to legendary ice hockey player Wayne Gretzky's famous saying: "I skate to where the puck is going to be, not where it has been." This Workshop Report aims to help policymakers skate to where the metaphorical AI puck is going.

The roundtable discussion centered on three topics:

1. **Potential Concerns and Indicators.** New risks and challenges are likely to arise from the increasing general-purpose capabilities of AI systems. To effectively prepare for these concerns, participants were asked to identify potential indicators or tripwires, enabling the development of strategies to proactively manage the challenges that may emerge in the security domain. These indicators could pertain to security-focused AI applications and extend to AI-generated geopolitical risks.

2. **Division of Responsibility for Managing Risk.** Interest in AI-based risk spans across government, academia, industry, and civil society, with different departments and agencies within government holding jurisdictional responsibility. The increasing generality of AI systems and their potential for use across many different sectors may necessitate new distribution of responsibility across actors.

3. **Levers to Manage Risk.** The policy community possesses unique levers to manage risks and challenges associated with increasingly general AI capabilities. The distinct challenges posed by general-purpose AI, however, may require the development of new policy and regulatory mechanisms. Participants sought to identify levers that are available and needed to manage the specific types of concerns associated with frontier AI systems. This also includes considering the responsibility that AI companies should assume, whether by actively developing security or defense applications or having technology relevant to the domain.

Box 1: Scope and Definitions

Some discussions of AI focus solely on systems that are already in widespread use today; others revolve around speculative future systems that may (or may not) be developed at some point in the future. This roundtable, by contrast, was scoped to explore risks that could arise from relatively foreseeable extensions of AI technologies that exist today.

The research frontier of what AI systems can do has expanded rapidly over the last decade, and shows few signs of slowing down. It is important for analysts and policymakers to be able to discuss how existing trends might continue and what implications they might bring, while being careful not to become excessively speculative. We use the phrase "AI research frontier" or similar phrasing to refer to AI models that are comparable to or slightly beyond the current cutting edge—recognizing that this is not a precise designation.

Throughout this document, we primarily focus on so-called "general-purpose AI" systems—such as large language models (LLMs)—which can be used for a wide range of different tasks. General-purpose AI systems are different from the "narrow AI" systems that characterized the field until recently, which specialize in a single task such as playing chess or classifying images. General-purpose AI is also a distinct term from "artificial general intelligence" or AGI, a contested term that is typically used to refer to an AI system with cognitive capabilities comparable to a human. The general-purpose AI systems that exist today can perform a broad range of different tasks, but in most cases are not as capable at those tasks as humans.

Progress and Trends in AI

Most debates about the future of AI are anchored in current technologies—such as today's LLM-based chatbots—but lack a clear sense of which tools or capabilities might emerge.

Many of today's AI technologies are relatively new. The transformer-based LLMs that dominate the news in 2023 have only existed since 2017, and did not enter widespread use until a few years later. [1] Before that point, discussions about AI centered on deep-learning-based classifiers, reinforcement learning agents, and models of various complex real-world systems. Deep learning, which has been so critical to AI progress over the past decade, only emerged in earnest in 2012. The pace of development since then has been incredibly swift—bolstered by algorithmic innovations, increasingly accessible high-end computing capabilities, and sustained investment in AI research and development—and shows few signs of slowing. The rapid tempo of change in what AI systems are capable of makes policymakers' work more difficult, as they must design policies that can both handle risks and harms from existing AI systems and also anticipate and adapt to further changes in the future.

The recent emergence of more general-purpose models has further complicated the picture. For decades, most AI systems have been designed to perform a single, narrowly defined task, such as playing chess, recognizing objects in an image, or ranking web content. By contrast, LLMs are capable of performing a wide range of distinct tasks, including translating text, writing and editing prose, solving math problems, writing software, and many more. While narrow AI systems will continue to be common in many areas, general-purpose AI is already entering more widespread use and is sure to spread further. Many existing approaches to AI policy, which conceptualize risks and harms in terms of specific use cases, will need to be adapted in order to handle the new difficulties posed by these more powerful AI models.[2]

Advancements in General-Purpose AI

The most capable general-purpose AI systems currently available to the public are LLMs which are generally accessible in the form of text-in, text-out chatbots, such as OpenAI's ChatGPT, Google's Bard, and Anthropic's Claude. However, there are several ways in which AI developers are actively working to augment these systems. Roundtable participants were asked to discuss how the current frontier of AI research might expand over the next few years to encompass new capabilities, which capabilities are most concerning, and how to evaluate the relationship between a model's capabilities and its potential to have real-world impacts.

Though it is difficult to predict exactly how each of these augmentations will play out, it is clear that they will expand the capabilities of these systems and, correspondingly, expand the safety and security concerns associated with them. Five such areas are as follows:*

1. **Multimodality:** A multimodal AI system is one that is capable of receiving multiple types of input (such as text, images, audio, or video) or generating multiple types of outputs. For example, an increasingly standard approach among AI developers is to feed images into LLMs. Multimodality makes existing AI models more powerful and may have significant privacy and security ramifications when deployed in certain contexts such as surveillance or law enforcement.

2. **Tool use:** Frontier AI systems will soon output not just static language, images, audio, or video, but will also likely have the capability to interact with the open internet or user data and applications. ChatGPT's experimental "plugins" feature is an early example of this capability. For example, AI systems with access to web search capabilities can query websites to synthesize information across multiple sources, while those with user-interface controls can take actions on websites instead of simply generating text. These additional functions are often referred to as "tools," because they allow AI systems to make use of other programs, including web applications, user data, scientific databases, and more. These developments are significant because soon these models will no longer be standalone systems, but rather fully integrated ones capable of interacting with a broader environment outside the AI itself relatively autonomously through a set of tools. Depending on the tool, this could also significantly increase the systems' capabilities.

3. **Deeper reasoning and planning:** A major current area of research focuses on extending LLMs' reasoning capabilities, making it highly plausible that future models will be significantly more powerful in this regard. For example, "chain-of-thought" prompting, in which LLMs generate intermediate reasoning steps when responding to a prompt, can significantly improve models' performance on certain tasks such as arithmetic or word problems.[3] Future models are likely to

* These five points are drawn from material presented by Matthew Botvinick at the roundtable.

incorporate such techniques by default, making them better equipped to handle complex multi-step tasks that involve sequential reasoning or planning.*

4. **Memory:** AI companies are experimenting with increasing LLMs' memory capabilities by either increasing the amount of information in their context window (the total amount of input text and output that an LLM can process while in use) or incorporating offline memory stores in order to improve their episodic memory. While greater memory capabilities could make future AI systems more personalized and easier to continually update, personalization also introduces greater privacy risks. AI systems with longer-term memory are also more likely to change their behavior over time, complicating efforts to evaluate them for safety or security concerns.

5. **Interaction:** Interaction *between* frontier AI systems is highly likely to produce unpredictable emergent behaviors. For example, in one experiment, researchers observed that a virtual world populated by generative agents produced believable individual and emergent social interactions.[4] This may help advance research related to collective intelligence or the modeling of complex systems, but interaction between systems deployed in the real world is likely to create privacy risks and unforeseen coordination problems that will be novel to human observers.

Implications

Each of these extensions of current AI systems would bring many advantages, but each would also have downsides. For instance, deeper reasoning and planning may make models more useful and more transparent (in the sense of being able to explain their reasoning), but would also mean that models could more effectively execute dangerous or undesirable actions, such as deceiving or manipulating human interlocutors. Several participants agreed that of the five points, tool use is the most concerning near-term capability because of the wide array of potential actions it enables, as well as the potentially high stakes and unpredictable outcomes of those actions. Researchers and engineers are already experimenting with setups that allow LLMs to use software interfaces that enable them to take actions in both the digital

* Here, we describe how the technology may develop within the current paradigm. It may be the case that to achieve generalized i.e. "out of domain" causal reasoning, new deep learning architectures or other computing paradigms are required beyond the existing transformer architecture. This in turn may require new testing and evaluation mechanisms.

and physical worlds, transforming them from passive generators to active agents.[5] As these interfaces increase, tool use will introduce new risks because there are significantly more points of vulnerability to attack.[6]

Multimodality, memory, and interaction also present new threats that are equally difficult to predict, such as coordination problems that may arise between AI agents acting independently. Mixing and matching these capabilities may further complicate the risk calculus; for instance, an AI model that combines multimodality with increased memory could pose a significantly greater privacy risk than a model with only one of those capabilities. Furthermore, each of these new capabilities also makes the AI systems themselves vulnerable to new forms of attack. A model that takes multimodal inputs, for example, is vulnerable to being jailbroken or exploited in more ways than a model that only accepts text inputs.[7]

As these augmentations are developed further and rolled out more widely, it will become more important—and more difficult—to distinguish between the capabilities and risks of the AI models in isolation, and the capabilities and risks associated with the environments in which they operate or the ways in which they might be exploited by malicious actors. This distinction can cut both ways. Model capabilities that may seem concerning may in fact be harmless—for instance, if a model produces instructions on how to create a chemical weapon, but the necessary reagents are strictly controlled. On the other hand, ways in which a model may seem too limited to cause harm may be misleading—for instance, if a model's context window is too short to develop and carry out a mass spearphishing attack in one go, but tool use and memory allow the model to call external programming libraries, save files to refer back to later, and so on. Conflating model-only capabilities and real-world risks makes it more difficult to identify and respond to the risks at hand.

Potential Concerns and Indicators

AI systems already cause significant harm.[8] Beyond these existing issues, researchers developing or scrutinizing the latest general-purpose AI models have raised concerns about a range of ways in which models at or just beyond the current AI frontier could potentially cause severe harm at a societal scale. Several organizations have begun to perform so-called "dangerous capability evaluations" of frontier AI models, including some of the AI companies themselves.[9] Dangerous capabilities currently being tested for or actively under consideration include:

- autonomous replication (a model's ability to acquire resources, create copies of itself, and adapt to novel challenges);[10]

- knowledge about chemical, biological, radiological, or nuclear weapon production;
- capacity to carry out offensive cyber operations;
- the ability to manipulate, persuade, or deceive human observers; advanced cognitive capabilities such as long-term planning and error correction; and,
- understanding of their own development, testing, and deployment (sometimes called situational awareness).[11]

While these evaluations are useful for assessing risk, they are only one part of a more comprehensive risk assessment which includes advanced threat actor modeling, red teaming, gathering real-world threat intelligence, and post-deployment monitoring.

This state of affairs—with commercial AI companies developing systems with potentially dangerous capabilities that they themselves can only test for after the systems are built—poses a challenge for policymakers. By necessity, the concerns being raised are speculative, since they relate to the development of novel capabilities that have only been observed in primitive forms. However, waiting to take action until it is definitively proved that AI systems do have the contemplated capabilities would be irresponsible, given the potential severity of the harm that could result.

The approach some companies are taking is to enumerate specific dangerous capabilities, then develop methods to evaluate whether a given system has those capabilities. This approach of identifying and testing for specific capabilities of concern has advantages and disadvantages. Two major advantages are that it makes it possible to identify specific behaviors that could be tested for in the lab, and that it facilitates consensus-building around clearly dangerous possibilities such as those listed above. Limitations of this approach include that testing for an individual capability may overlook potential risks from combinations of capabilities, and that it is impossible to enumerate all possible risks and concerns—there will always be unknown unknowns. In order for capabilities-based evaluations to effectively reduce risks, they will need to be designed with reference to a broad range of threats. They will also need to balance being sufficiently sensitive to provide early indicators of concerning capabilities without overstating risk from rudimentary capabilities.

While there is some tentative agreement about which dangerous capabilities AI systems should be tested for, it is an ongoing challenge to develop specific evaluations that keep pace with the rate of AI development. Furthermore, there is no consensus on evaluation standards or the capabilities that testers should evaluate. Nor are there established reporting mechanisms to share the results of evaluations with other AI developers or policymakers.[12]

Several participants noted that cybersecurity best practices are also relevant in securing AI systems. Such practices include mechanisms to ensure data confidentiality or integrity as well as secure software development practices. Threat emulation techniques or red teaming can also help discover vulnerabilities throughout the AI development pipeline. Some AI companies are already incorporating these techniques in order to increase security, along with other established practices such as real-world threat intelligence collection. Cybersecurity measures such as authentication and access controls are also already being used to protect against unintended effects enabled by plug-ins and greater integration of the models with the internet (i.e. the safety concern of "tool use").[13] Participants noted that such measures need to become standard practice as AI models become more integrated with other systems and user data. Focusing on shoring up the security of today's AI systems will provide a set of actionable steps to address some of the most pressing existing vulnerabilities associated with AI, while also laying the groundwork for addressing future harms and risks.

Lessons from cybersecurity may also help improve the evaluations process. Several participants noted that rather than there being tension between dangerous capability evaluations and cybersecurity best practices, the two areas seem to naturally complement each other and could in fact be used to reinforce each other within the AI development pipeline. For instance, regulators could require secure software development standards for frontier AI systems, or dangerous capability evaluations could serve as a trigger for the adoption of increasingly stringent information security practices to reduce the chance that a dangerous model is leaked or stolen.

Finally, it is important to avoid over-indexing on pre-deployment evaluations as a shortcut to risk reduction and ensure that the evaluations we do conduct provide early warnings of potential threats. While some risks will be evident from the capabilities of the models themselves, many more will result from the way those models interact with their environments and society at large. Just because a model passes a dangerous capability evaluation does not mean that it cannot be harmful or unsafe in other ways, be exploited by a malicious threat actor, or that new risks will not emerge. Distributed harms are also much more difficult to identify than discrete harms, especially when they develop over extended periods of time. Dangerous capability evaluations are unlikely to be able to predict diffuse societal harms, such as worsening inequality or geopolitical instability, that may also be associated with the adoption of advanced AI. It will ultimately be extremely hard to discover many harmful capabilities or detect them as they emerge. The challenge of testing for capabilities that may as yet be unknown

should not be underestimated; unexpected dangerous capabilities may emerge even in instances where effective evaluations for known capabilities have been performed.

Responsibility for Managing Risk

Assuming that a consensus is reached about which concerns are likely to cause the most risk to national security, who is responsible for managing that risk? Participants were divided on the degree to which companies should be liable for risk, whether or not existing practices can serve as a template for AI risk management, and the role governments should play in the process.

Box 2: The Value—and Limitations—of Learning from Analogies

Policymakers faced with a new technology frequently reach for analogies that can help them make sense of the unfamiliar. Recent hearings on AI by the Senate Judiciary Committee, for instance, have leaned heavily on social media as an analogy for AI.

Analogies like this can be very helpful as a way to learn lessons from previous issues, and can also provide templates for policy interventions that can be adapted rather than starting from scratch. For example, roundtable participants discussed the government-led cybersecurity vulnerabilities and equities process (VEP), which provides an incentive to companies to disclose cyber vulnerabilities by removing the risk of liability to those companies. This is an interesting example of a voluntary, but powerful, way to manage risks that could perhaps be adapted for AI.

No analogy is perfect, however. A VEP-like process, for instance, would be difficult to apply directly to AI because AI vulnerabilities are frequently difficult to fix even in principle. Policymakers should seek opportunities to port lessons over from other technologies, but should also be clear-eyed about the ways in which new problems do, sometimes, need new solutions.

Key Variables for AI Risk Management

Participants identified two key variables that should be considered when determining who should manage risk. The first is time, or how much notice relevant parties would have about when a dangerous capability might emerge. Two separate timelines are relevant:

1) when a capability is first detected; and,
2) when it actually causes significant harm.

Companies developing AI systems should work towards being able to detect potentially dangerous capabilities—and notify regulators—as early as possible in order to maximize the time between these two points. At present, governments might not be able to respond rapidly if given short notice of the emergence of a dangerous capability. Government and industry should collaborate to establish emergency response protocols such that each party can effectively respond to concerning developments.

The second relevant variable is the level of expertise and knowledge required to mitigate the harm. This is expected to be highly context-dependent. Participants noted that there are some instances in which much of the relevant expertise and knowledge for risk management is highly concentrated in government; for example, in areas such as cyber offense and biological or nuclear weapons. By contrast, however, government has very little visibility into or expertise in risks from unforeseen autonomous behavior by advanced AI systems. The valuable role of government will be the incentive structures, both positive and negative, that it establishes. This also highlights the importance of distinct roles and responsibilities for industry, government, and third parties, such as independent auditing organizations. In scenarios where these roles and responsibilities collapse, it will be harder to maintain accountability.

One possible short-term response could be advocating for the development of crisis management plans for potential AI accidents or incidents in which an AI system causes severe unintended harm. For instance, the U.K. national risk register assesses various risk scenarios based on impact criteria such as fatalities, economic impact, public perception, and international relations. [14] While the bar for severity for inclusion in the register is quite high, there may be some value in formally recognizing certain AI risks as national risks. Another option is to conduct tabletop exercises with government and industry officials that examine possible high-impact scenarios, delineate roles and responsibilities for actors in a crisis, and recommend potential crisis response actions by relevant actors.

Several participants agreed that the status quo places too much of the responsibility for managing risk on industry. Striving to ensure that government can play an active role in risk management would not only resist the centralization of power in a small number of well-resourced firms, but also encourage the development of knowledge and expertise in government or other oversight bodies. Governments should build the capacity to perform independent safety testing and risk assessments and consider mandating companies to report the results of their internal risk assessments, capabilities, and limitations similar to the recent voluntary measures announced in the United States.[15] Threat intelligence sharing between AI companies and governments could also facilitate active communication and signal the emergence of adversarial exploitation of AI systems. In addition, third-party testing and auditing organizations likely have a role to play here.

Finally, while capability evaluations are a valuable tool for managing risk, they are only valuable to the extent that AI companies are able to articulate how they will respond in the event that evaluations trigger an internal tripwire. Companies should have on-the-shelf policies (sometimes called "responsible scaling policies" or "safe development policies") that describe internal and external reporting requirements and other safety protocols that are activated if a dangerous capability is discovered. Such a policy should clearly state what capabilities will be tested for, when evaluations will be run (ideally including during the training process as well as prior to deployment), and the circumstances under which it would no longer be safe to continue advancing the state of the art of an organization's AI system.

Levers to Manage Risk

Policy levers to manage risk from frontier AI models can be grouped into three categories by their respective goals:

1) creating visibility and understanding,
2) defining best practices, and;
3) incentivizing and enforcing certain behaviors.

There are different intervention points across the AI lifecycle, and different risks can arise at different points.

Table 1. Categorizing Policy Levers along Various Stages of the AI Development Lifecycle.[16]

Policy Lever	AI Development Lifecycle Stage: Model development	Initial deployment and proliferation	Deployment in narrow contexts	Broader societal impacts
Visibility and under-standing	Pre-training disclosure	Pre-deployment disclosure	Incident sharing	Measuring and forecasting societal impacts
Defining best practices	Risk assessment guidelines, evaluations, and standards for developers	Deployment decisions informed by risk assessments	Sector-specific guidelines, e.g. assurance requirements for AI in high-stakes contexts	Guidelines for ongoing monitoring and risk assessment
Incentives and enforcement	Public funding for safety research Licensing and/or liability for development	Licensing and/or liability for deployment Export controls Open source restrictions	Domain-specific regulation	

Source: Centre for Emerging Technology and Security

Many participants were interested in exploring incentives for private-sector actors to reduce risk, particularly incentives to test and evaluate their frontier AI systems and to report dangerous capabilities to oversight bodies. Some ideas included leveraging government procurement requirements, establishing industry certifications for frontier AI systems, and loosening liability in exchange for transparency.

Another set of potential levers center on standardizing tests and evaluation processes during the AI development pipeline. A hypothetical timeline might involve AI companies adopting voluntary standards to publish information about their testing, evaluation, and forecasting procedures; governments requiring specific kinds of testing once companies coalesce around a set of shared standards; and policymakers establishing a liability framework based on these relationships over time that actually imposes costs on private sector actors who fail to meet the government requirements. For instance, the incorporation of model cards, a type of documentation that provides details about a specific AI model's intended use case, training data, performance and evaluation metrics, and other attributes, has become best practice within the AI development community along with similar documentation formats for various datasets.[17] Incorporating "testing cards" into the development workflow could be the first step toward the widespread adoption of dangerous capability evaluations.

Lack of consensus among relevant stakeholders is a consistent theme that hinders the identification of feasible policy levers. However, this lack of consensus suggests a larger role for government bodies and standard-setting agencies. For instance, participants generally agreed that it is more feasible for governments or oversight bodies to mandate some form of transparency around testing policies and procedures than it is for them to delineate exactly what those testing policies should be.

Conclusion

Overall, roundtable participants strongly agreed on the need to prepare for the capabilities and risks associated with frontier AI models now. Although there is a great deal of uncertainty, there is also a pressing need to build expertise and begin developing a consensus among relevant stakeholders. Key takeaways from the discussion include:

AI capabilities are evolving quickly and pose novel—and likely significant—risks.

- These advancements also pose novel policy and governance challenges, since risks posed by these models cannot be cleanly segmented and managed according to sector or use case.
- Several foreseeable extensions of existing LLMs—including multimodality, tool use, deeper reasoning and planning, memory, and interaction—have the potential to significantly expand the risk profile of these systems. Tool use, in particular, creates a wide range of new potential risks and vulnerabilities.
- Potential capabilities of concern include autonomous replication (a model's ability to acquire resources, create copies of itself, and adapt to novel challenges); knowledge about chemical, biological, radiological, or nuclear weapon production; capacity to carry out offensive cyber operations; the ability to manipulate, persuade, or deceive human observers; advanced cognitive capabilities such as long-term planning and error correction; and situational awareness.
- While enumerating specific capabilities is useful for the development of evaluation methodologies and consensus about risk, it likely may not fully capture the risks of deployed AI systems in practice. When evaluations do detect dangerous capabilities, this could act as a trigger for reporting requirements and additional safety measures, but passing an evaluation should not be seen as a guarantee of safety.
- As AI companies continue to iterate on a set of best practices for identifying novel issues with frontier models, it is also critical that developers also dedicate appropriate resources to identify and mitigate known harms.

Distributing AI-related knowledge and expertise more evenly, especially within government, is important for managing risks associated with frontier AI systems.

- Government's ability to provide meaningful oversight of risks associated with frontier AI development is hindered by the fact that the center of gravity of AI expertise and knowledge currently rests in the private sector.
- In other respects, some types of expertise such as that associated with specific national security concerns will be concentrated in government. Technical and domain-specific expertise needs to be brought together to ensure the most pressing risks are adequately captured in new testing and evaluation methodologies.

There are several concrete policy levers that can help both AI developers and governments prepare for risks associated with frontier AI.

- Transparency and reporting requirements could help create visibility for regulators and the public by facilitating access to information about the capabilities and potential risks of frontier AI models.
- Supporting the development of a third-party ecosystem of organizations that can test and audit dangerous capabilities and other risks could be a valuable way to manage risk, and could have greater flexibility and capacity than if this responsibility rests solely with government.
- Both policymakers and developers should consider the development of crisis management plans for AI accidents. Developers should also have clear protocols for sharing information with empowered authorities and other AI developers if concerning capabilities or threats are discovered.

Authors

Helen Toner is Director of Strategy and Foundational Research Grants at CSET and also serves in an uncompensated capacity on the non-profit board of directors for OpenAI.

Jessica Ji is a Research Analyst on the CyberAI Project at CSET.

John Bansemer is a Senior Fellow and the Director of the CyberAI Project at CSET.

Lucy Lim is a Research Scientist in Strategic Governance and Frontier Safety at Google DeepMind.

Chris Painter is a member of the technical staff at ARC Evals.

Courtney (Court) Corley is chief scientist for artificial intelligence in the AI and Data Analytics Division at Pacific Northwest National Laboratory.

Jess Whittlestone is Head of AI Policy at the Centre for Long-Term Resilience.

Matt Botvinick is Senior Director of Research at Google DeepMind.

Mikel Rodriguez leads the AI Red Team at Google DeepMind.

Ram Shankar Siva Kumar is a data cowboy at Microsoft Security Research and tech policy fellow at the CITRIS Policy Lab and the Goldman School of Public Policy at UC Berkeley.

Acknowledgments

We would like to thank several participants in the roundtable who contributed greatly to the discussion but were unable to participate in the writing process: Chris Meserole, Dorothy Chou, Jade Leung, Jeff Alstott, and Marina Favaro.

From CSET, we would like to thank Igor Mikolic-Torreira for his feedback and Tessa Baker, Shelton Fitch, and Jason Ly for their editorial and design support.

Document Identifier: doi: 10.51593/2023CA004
Document Last Modified: 25 October 2023

Endnotes

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," arXiv preprint arXiv:1706.03762 (2023), https://arxiv.org/abs/1706.03762.

[2] For more on the policy challenges posed by general-purpose models, see: AI Now Institute, Amba Kak, Sarah Myers West, "General Purpose AI Poses Serious Risks, Should Not Be Excluded From the EU's AI Act | Policy Brief," AI Now Institute, April 13, 2023, https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act; and Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv preprint arXiv:2307.03718 (2023), https://arxiv.org/abs/2307.03718.

[3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv preprint arXiv:2201.11903 (2023), https://arxiv.org/abs/2201.11903.

[4] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," arXiv preprint arXiv:2304.03442 (2023), https://arxiv.org/abs/2304.03442.

[5] Will Knight and Khari Johnson, "Now That ChatGPT Is Plugged In, Things Could Get Weird," Wired, March 28, 2023, https://www.wired.com/story/chatgpt-plugins-openai/.

[6] For instance, a technique called indirect prompt injection exploits the fact that web-browsing LLMs do not distinguish between instructions and data to induce undesirable behaviors. See: Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, Mario Fritz, "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," arXiv preprint arXiv:2302.12173 (2023), https://arxiv.org/abs/2302.12173.

[7] Christian Schlarmann and Matthias Hein, "On the Adversarial Robustness of Multi-Modal Foundation Models," arXiv preprint arXiv:2308.10741 (2023), https://arxiv.org/abs/2308.10741.

[8] "Welcome to the Artificial Intelligence Incident Database," AI Incident Database, accessed August 15, 2023, https://incidentdatabase.ai; Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, Andrew D. Selbst, "The Fallacy of AI Functionality," arXiv preprint arXiv:2206.09511 (2022), https://arxiv.org/abs/2206.09511.

[9] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, "Model evaluation for extreme risks," arXiv preprint arXiv:2305.15324 (2023), https://arxiv.org/abs/2305.15324; OpenAI, "GPT-4 System Card," March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf; "Frontier Threats Red

Teaming for AI Safety," Anthropic, July 26, 2023, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety.

[10] Megan Kinniment, Lucas Jun, Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks," Alignment Research Center, July 31, 2023, https://evals.alignment.org/Evaluating_LMAs_Realistic_Tasks.pdf.

[11] For more on "situational awareness," see: Kelsey Piper, "Situational awareness," Planned Obsolescence Blog, June 6, 2023, https://www.planned-obsolescence.org/situational-awareness/.

[12] Since the roundtable was held, Google, Microsoft, OpenAI, and Anthropic announced a new industry body that, among other aims, is intended to facilitate information sharing between industry and government. See: Paul Sawers, "OpenAI, Google, Microsoft and Anthropic form body to oversee safe 'frontier AI' development," TechCrunch, July 26, 2023, https://techcrunch.com/2023/07/26/openai-google-microsoft-and-anthropic-form-body-to-oversee-safe-frontier-ai-development/.

[13] For more discussion of cybersecurity practices for AI, see: John Bansemer and Andrew Lohn, "Securing AI Makes for Safer AI," *Center for Security and Emerging Technology Blog,* July 6, 2023, https://cset.georgetown.edu/article/securing-ai-makes-for-safer-ai/.

[14] "National Risk Register 2020," Gov.uk, December 18, 2020, https://www.gov.uk/government/publications/national-risk-register-2020.

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

[16] Adapted from: Ardi Janjeva, Nikhil Mulani, Rosamund Powell, Jess Whittlestone, Shahar Avin, "Strengthening Resilience to AI Risk," Centre for Emerging Technology and Security, August 2023, https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk.

[17] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, "Model Cards for Model Reporting," arXiv preprint arXiv:1810.03993 (2019), https://arxiv.org/abs/1810.03993.