Understanding frontier AI capabilities and risks through semi-structured interviews

Akash R. Wasil
Georgetown University

Lukas Berglund
Independent

Tom Reed
University of Cambridge, ERA AI Fellowship

Miro Pluckebaum
Independent

Everett Smith
Independent

Abstract

We describe how semi-structured interviews with frontier AI developers could help governments better understand AI progress and AI-related national security threats. Thus far, governments have largely relied on formal model evaluations to monitor frontier AI capabilities. While such approaches are valuable, they have several limitations that could be complemented with semi-structured, qualitative interviews. Such interviews could provide insights into areas formal evaluations may miss, such as developers' intuitive understanding of AI capabilities, early warning signs of extreme risks, predictions about future AI progress, and information about internal safety practices. We outline an approach in which government officials would have regular (e.g., monthly or quarterly) interviews with employees from the top 5-10 frontier AI developers. We also outline example questions that officials could ask in four categories: (a) AI capabilities and AI progress, (b) National security, (c) Safety culture, and (d) Miscellaneous. Conducting such interviews– either through mandatory reporting requirements or voluntary agreements– would enable governments to better understand and prepare for the national security implications of frontier AI systems.

Key Findings and Recommendations

- Semi-structured interviews with employees at frontier AI companies could provide governments with valuable insights to better understand AI progress and risks.
- Interviews could reveal developers' understanding of current AI capabilities, early warning signs of risks, predictions about future progress, and information about internal safety practices and culture.
- Government stakeholders should conduct regular interviews with diverse employees from the top 5-10 AI companies, with protections for employees and limits on disclosure of sensitive information.
- Example interview questions cover topics including impressive and concerning model capabilities, predicted future capabilities, implications of AI progress for national security, limitations to model control and security, safety culture and practices, and suggestions for the government.
- Conducting such interviews, either through mandatory reporting or voluntary agreements, could improve the government's ability to understand, prepare for, and mitigate AI-related security threats.

1 Introduction

Major governments have acknowledged that advanced AI systems pose important national and global security risks. For example, at the world's first AI Safety Summit, attending countries recognized that "there is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models" (UK Government, 2023). More recently, the Bipartisan Senate AI Working Group released a roadmap that included a section on national security risks. In their roadmap, the senators acknowledged that AI has the potential to increase risks from biological weapons and emphasized the need to better understand risks from artificial general intelligence (Bipartisan Senate AI Working Group, 2024). These statements echo concerns from industry and academia. Many experts believe that AI poses grave global security threats and have called on governments to invest more resources into understanding and mitigating such threats (Bengio et al., 2024; Center for AI Safety, 2023).

To manage these risks, governments are seeking to increase their understanding of frontier AI capabilities. Both the US government and the UK government have established AI safety institutes tasked with developing empirical model evaluations. Such evaluations aim to detect dangerous capabilities in existing models, understand trends in capability progress, and allow officials to predict dangerous capabilities in advance (see Shevlane et al., 2023). So far, this approach has focused on methods that involve directly running tests or benchmarks on models[1]. Such methods have several advantages– findings can be replicated, experts can understand the capabilities of systems through direct observation, and the approach allows the government to develop its technical expertise and workforce.

Nonetheless, there are important limitations to this experimental approach toward understanding frontier AI capabilities. These include the following:

1. **The science of model evaluation is nascent, imperfect, and may be outpaced by progress in capabilities.** Developing evaluations requires considerable effort and expertise. Moreover, evaluations can quickly become outdated as AI technology rapidly progresses[2] (see Ganguli et al., 2023). Important aspects of model capability may become evident to those deeply familiar with using the model before they are reflected in formal evaluations.

2. **Model evaluations are highly sensitive to context**. The capabilities of an AI system in a given context are a function not only of the model's latent potential but also the expertise of the user operating it. The use of up-to-date prompting or scaffolding techniques or other technical methods can vastly improve the capabilities of an AI system.[3] Although the US

[1]As an example, see the US AI Safety Institute's strategy document: "These projects will aim to assess what risks advanced AI systems might pose before being deployed or released, utilizing methodologies such as automated capability evaluations, expert red-teaming, A/B testing, and other methods" (NIST, 2024, p.4).

[2]For example, consider the Bias Benchmark for Question Answering (BBQ)– an evaluation developed by Anthropic. "BBQ took the developers ~2 people years spread over 6 months across 8 people to build. Designing and implementing even just a single evaluation is a resource intensive effort that can take tens of people several months." (Ganguli et al., 2023)

[3]For example, suppose a government is assessing risks from AI agents that can autonomously self-improve or contribute to research that produces more capable AI models (see UK Government, 2024). Within a company, the AI would work closely with developers of the frontier AI company, embedded within various tools that the developers regularly use. More broadly, existing evaluations focus on isolated tasks rather than full simulations, failing to capture the dynamic interactions present in actual deployment settings. Consequently, government auditors may miss crucial aspects of an AI system's capabilities and risks that only emerge in its real-world context. This may be especially concerning for risks that emerge when AI systems are internally deployed to assist with AI R&D tasks.

and UK governments are making efforts to recruit technical talent, they are unlikely to obtain talent that matches that of major frontier AI companies. As a result, they will likely lack the expertise to elicit a model's full capabilities.

3. **Government evaluators have limited access to frontier models**. In the past, several companies have refused to provide government access to their cutting-edge systems (Stein-Perlman, 2024). Although frontier AI developers have agreed to provide model access to the US AI Safety Institute, there will likely be a lag between when a model is available to lab employees and when it is released to auditors. This delay means that the government may not acquire information about a model's capabilities until well after it has been developed, internally deployed, or potentially even stolen by malicious actors (see Nevo et al., 2024 for more on the importance and difficulty of securing the weights of dangerous models). Furthermore, since model access is currently voluntary, it can also be revoked at any point, posing a risk to consistent oversight.
4. **Evaluation results are difficult to meaningfully communicate to policymakers and national security.** Interpreting the significance of a benchmark result may require a deep technical understanding of the particular dataset and the model being evaluated.

Qualitative assessment techniques can help governments better understand the capabilities and national security risks of advanced AI systems. Empirical and quantitative evaluation methods can be meaningfully supplemented with *qualitative interviews with frontier AI developers*. Through such interviews, government officials could:

- Acquire greater insight into the capabilities of current models,
- Anticipate the capabilities and potential national security threats of next-generation systems,
- Understand the bottlenecks that developers believe are limiting the models
- Gain insights into the safety and security practices at frontier AI companies.

This approach could be especially valuable given that employees at frontier AI companies interact daily with state-of-the-art systems, have deep knowledge about the technology, and may be able to offer insights that go beyond information revealed through formal evaluations. Such interviews can also help with emergency preparedness. In the event of a sudden national security threat from advanced AI, employees at frontier AI companies would likely be among the first to notice or predict the threat. Regular interviews with employees from frontier AI companies could help improve the government's ability to detect and predict sudden national security threats from advanced AI systems.

2 General Approach

What would a qualitative interview setup look like? We envision the following properties:

- **Interviews focus on a small number of frontier AI model developers**. The interviews would take place with employees from entities developing dual-use foundation models,

as defined in the White House Executive Order[4] (White House, 2023). In practice, we suspect this would initially include about 5 major companies: OpenAI, Google, Anthropic, Microsoft, and Meta.

- **Interviews take place with ~5 employees from each company**. For each company, about 5 employees would be expected to participate in 60-minute interviews with government officials each month or quarter. (With 5 frontier AI developers, that would mean 25 total hours of interview time per month.)
- **Interviews take place with members from various teams**. To get a range of perspectives, these five employees would be selected from various teams. For example, one member might be from the development team, one member from a technical safety team, one member from a security team, and one member from an internal governance team.
- **Interviews ensure adequate protections for employees**. To ensure honest and transparent reporting, interviews could be anonymized, employees could be required to report accurately, and companies could be prohibited from retaliating against employees who reveal safety or security concerns.
- **Interviewers minimize the amount of unnecessary information disclosed.** Interviewers are trained to avoid asking employees to reveal sensitive proprietary information (such as the exact techniques used to train or modify a model) unless when absolutely necessary from a national security perspective.

Many of these practices could be drawn from other fields in which qualitative interviews are a regular part of safety and security procedures (**see Table 1**).

Table 1: Examples of the use of interviews across industries.

Organization	**Type**	**Description**
Chemical Safety Bureau (CSB)	Post-incident investigations	The Chemical Safety Bureau uses both formal and informal interviews with employees to investigate the cause of dangerous accidents (see CSB, 2022 for an example)
Environmental Protection Agency (EPA)	Compliance with standards	The EPA uses interviews to assess compliance with standards (for an example, see EPA, 2007).

Continued on next page

[4]The term "dual-use foundation model" is defined as: "AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by: (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons; (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or (iii) permitting the evasion of human control or oversight through means of deception or obfuscation" (White House, 2023).

Table 1: Examples of the use of interviews across industries. (Continued)

Organization	Type	Description
Food and Drug Administration (FDA)	Routine inspections	The FDA conducts unannounced inspections of manufacturers at least once every two years to ensure compliance with standards. Comprehensive interviews with diverse members of staff—from management to onsite technicians—are used alongside records review and facility inspections to assess compliance (FDA, 2024).
Nuclear Regulatory Commission (NRC)	Safety culture assessment	The NRC conducts interviews to assess a facility's safety culture. This involves asking employees how management has reacted to and communicated about safety concerns in the past, asking about general safety practices, and asking about employees' perceptions of leadership's commitment to prioritizing safety concerns (NRC, 2019).
National Transport Safety Bureau (NTSB)	Post-incident investigations	Interviews are regularly used to establish the cause of a dangerous accident (see NTSB, 2010 for an example).
Securities and Exchange Commission (SEC)	Risk assessment	The SEC may conduct thematic reviews of emerging issues or trends to determine if there are any risks that require more formal investigations. This process can involve conducting interviews or asking discovery questions to learn more about products, determine their risk profile, understand company practices, and determine if there are any risks that require more formal investigations (SEC, 2024).

3 Example interview guide

Suppose the government set up the interview regime described in the previous section. In this section, we outline some examples of questions that the government interviewers could ask. We divide these questions into a few categories: **(a) AI capabilities and AI progress**, **(b) national security**, **(c) safety culture,** and **(d) miscellaneous**.

Section 1: AI capabilities and AI progress.

1. What do you believe are the most impressive capabilities of the model[5]?
2. What do you believe are the most impressive things that you and others at your company are using the models for?
3. How would you characterize the current primary limitations preventing your models from having significantly more powerful capabilities? How easy do you think it will be to overcome these limitations?
4. What do you predict are the most impressive capabilities that you will observe in the next few months?
5. To what extent has AI been able to help you with your research? How much have AI systems contributed to or accelerated your research and development efforts? How much compute is used to run AIs that help you with your research (if non-negligible)?
6. Broadly, what else do you think we should know about frontier AI capabilities or general trends in AI progress?

Section 2: National security

1. From a national security perspective, what do you believe are the most concerning capabilities of the model?
2. What do you believe are the most concerning things that malicious actors could use the model to do?
3. What do you predict are the most concerning or security-relevant capabilities that you will observe in the next generation of AI models, or in the next few months?
4. To what extent do you believe that the next generation of AI systems may have any of the following capabilities? When, if ever, do you expect such capabilities are likely to emerge?
    a. CBRN capabilities (ability to exacerbate chemical, biological, radiological, or nuclear threats)
    b. Cyber capabilities (ability to assist with hacking or self-exfiltrate)
    c. AI R&D capabilities (ability to substantially contribute to accelerated AI R&D)
    d. WMD capabilities (ability to contribute to the development of novel weapons of mass destruction)

[5]The company's current state-of-the art model.

5. To what extent are you confident that your institution will be able to control the next generation of AI models?
6. To what extent are you confident that your institution will be able to prevent the weights of the next generation of AI models from being leaked or stolen?
7. If a malicious actor stole the weights of a model in the next generation of systems, what would be some of the most concerning things they could do?
8. Do you have any national security concerns about the AI development occurring at the frontier AI company you work for?
9. Do you have any national security concerns relating to the AI development occurring in other frontier AI developers?
10. Broadly, what else do you think we should know about the potential national security implications of frontier AI development?

Section 3: Safety culture

1. What are the biggest strengths your company has from a safety culture perspective[6]?
2. What are the biggest weaknesses your company has from a safety culture perspective?
3. Broadly, how do you feel about the safety culture at your lab?
4. If you raised serious safety or security concerns to leadership, to what extent do you think they would take such concerns seriously?
5. Think of the last few times there was an important disagreement about safety and security issues, or a significant disagreement between the safety team and company leadership. How was this situation handled, and what (if anything) do you think could have gone better?
6. How are decisions around or issues of safety communicated within the company?
7. From a safety perspective, what is the pre-deployment process like? What changes (if any) would you make to this process?
8. If you identified a meaningful safety issue, how would you act on it? What do you think would happen, and how long would it take?

Section 4: Miscellaneous

1. Is there anything else you think we should know about AI capabilities, AI progress, national security concerns, or anything else?
2. Do you have any suggestions for ways the government could improve our ability to prepare for or mitigate AI-related safety and security concerns?

[6]The interviewer should provide a definition of safety culture. We recommend the following definition adapted from the CDC: "A culture of safety describes the core values and behaviors that come about when there is collective and continuous commitment by organizational leadership, managers, and workers to emphasize safety over competing goals." (CDC, 2023)

References

Bipartisan Senate AI Working Group. (2024). Driving US Innovation in Artificial Intelligence. https://www.young.senate.gov/wp-content/uploads/Roadmap_Electronic1.32pm.pdf

Bengio, Y., et al. (2024). Managing extreme AI risks amid rapid progress. Science. https://www.science.org/doi/10.1126/science.adn0117doi:10.1126/science.adn0117

CDC. (2023). Definition Examples of Safety Culture and Overlap with Safety Climate. https://www.cdc.gov/niosh/learning/safetyculturehc/module-1/4.html

Center for AI Safety. (2023). Statement on AI Risk. https://www.safe.ai/work/statement-on-ai-risk

CSB. (2022). Pressure Vessel Explosion at Loy-Lange Box Company. https://www.csb.gov/assets/1/6/loy_lange_box_company_report_-_final.pdf

EPA. (2007). Breaking Barriers: A Pesticide Inspectors' Manual for Interviewing Spanish Speaking Agricultural Workers on the Worker Protection Standard. https://nepis.epa.gov/Exe/ZyPDF.cgi/P100H5SN.PDF?Dockey=P100H5SN.PDF

FDA. (2024). Investigations Operations Manual. https://www.fda.gov/media/166525/download?attachment

Ganguli, D., Schiefer, N., Favaro, M., & Clark, J. (2023). Challenges in evaluating AI systems. https://www.anthropic.com/index/evaluating-ai-systems

NTSB. (2010). Loss of control on approach of Colgan Air. https://www.ntsb.gov/investigations/accidentreports/reports/aar1001.pdf

Nevo, H., et al. (2024). Securing AI Model Weights. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2849-1.html

NIST. (2024). The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals. https://www.nist.gov/system/files/documents/2024/05/21/AISI-vision-21May2024.pdf

NRC. (2019). Guidance for conducting an independent NRC safety assessment. https://www.nrc.gov/docs/ML1906/ML19066A376.pdf

Shevlane, T., et al. (2023). Model evaluation for extreme risks. arXiv. https://arxiv.org/abs/2305.15324

Skinner, L. (2016). What advisers can expect from an SEC exam. https://www.investmentnews.com/industry-news/cover-story/what-advisers-can-expect-from-an-sec-exam-66697

SEC. (2019). Filing Review Process. https://www.sec.gov/divisions/corpfin/cffilingreview

SEC. (2014). 2024 Examination Priorities: Division of Examinations. https://www.sec.gov/files/2024-exam-priorities.pdf.

Stein-Perlman, Z. (2024). AI companies aren't really using external evaluators. https://ailabwatch.org/blog/external-evaluation/

UK Government. (2024). AI Safety Institute approach to evaluations. https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations

UK Government, The Bletchley Declaration (2023). https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration

White House. (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/