Summary of "Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems"

AI capabilities are evolving quickly and pose novel—and likely significant—risks. The recent advent of more powerful and general-purpose AI systems such as large language models (LLMs) has increased expectations that they will have significant societal impacts. **Several foreseeable extensions of existing LLMs have the potential to significantly expand these systems' risk profiles.** These include multimodality, tool use, deeper reasoning and planning, memory, and interaction with other systems.

AI developers are currently attempting to evaluate systems for specific dangerous capabilities. 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.

In July 2023, the CSET and Google DeepMind hosted a virtual roundtable, as part of a series of roundtables organized by DeepMind to gather different views and perspectives on AI developments. This roundtable sought 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 Roundtable Report summarizes some of the key themes and conclusions of the roundtable discussion and aims to help policymakers "skate to where the puck is going to be." These themes and recommendations do not necessarily reflect the organizational views of either Google DeepMind or CSET.

Recommendation 1: 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 concentration of expertise in the private sector.
- Technical AI and domain-specific expertise—such as that associated with specific national security concerns—needs to be brought together to ensure that

the most pressing risks are adequately captured in testing and evaluation methodologies.

Recommendation 2: 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.
- 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.

For more information:

- Download the report: https://cset.georgetown.edu/publication/skating-to-where-the-puck-is-going
- Contact Us: Helen Toner | Helen.Toner@georgetown.edu