FRAMEWORK FOR MITIGATING EXTREME AI RISKS

SENATORS ROMNEY, REED, MORAN, AND KING

THE PROBLEM: AI'S EXTREME RISKS

Artificial intelligence (AI) has the potential to dramatically improve and transform our way of life, but also presents a broad spectrum of risks that could be harmful to the American public. Extremely powerful frontier AI could be misused by foreign adversaries, terrorists, and less sophisticated bad actors to cause widespread harm and threaten U.S. national security. Experts from the U.S. government, industry, and academia believe that advanced AI could one day enable or assist in the development of biological, chemical, cyber, or nuclear weapons.

While Congress considers how to approach new technology developments, we must prioritize AI's potential national security implications. New laws or regulations should protect America's competitive edge and avoid discouraging innovation and discovery.

OVERVIEW

Our framework establishes federal oversight of frontier AI hardware, development, and deployment to mitigate AI-enabled extreme risks—requiring the most advanced model developers to guard against biological, chemical, cyber, or nuclear risks.

An agency or federal coordinating body would oversee implementation of new safeguards, which would apply to only the very largest and most advanced models. Such safeguards would be reevaluated on a recurring basis to anticipate evolving threat landscapes and technology.

COVERED FRONTIER AI MODELS

The framework would only apply to frontier models—the most advanced AI models developed in the future— that are both: (1) trained on an enormous amount of computing power (initially set at greater than 10^{26} operations) and (2) either broadly-capable; general purpose and able to complete a variety of downstream tasks; or are intended to be used for bioengineering, chemical engineering, cybersecurity, or nuclear development.

OVERSIGHT OF FRONTIER MODELS

I. HARDWARE

Training a frontier model would require tremendous computing resources. Entities that sell or rent the use of a large amount of computing hardware, potentially set at the level specified by E.O. 14110, for AI development would report large acquisitions or usage of such computing resources to the oversight entity and exercise due diligence to ensure that customers are known and vetted, particularly with respect to foreign persons.

II. DEVELOPMENT OF FRONTIER MODELS

Developers would notify the oversight entity when developing a frontier model and prior to initiating training runs. Developers would be required to incorporate safeguards against the four extreme risks identified above, and adhere to cybersecurity standards to ensure models are not leaked prematurely or stolen.

Frontier model developers could be required to report to the oversight entity on steps taken to mitigate the four identified risks and implement cybersecurity standards.

III. DEPLOYMENT OF FRONTIER MODELS

Frontier model developers would undergo evaluation and obtain a license from the oversight entity prior to release. This evaluation would only consider whether the frontier model has incorporated sufficient safeguards against the four identified risks.

A tiered licensing structure would determine how widely the frontier model could be shared. For instance, frontier models with low risk could be licensed for open-source deployment, whereas models with higher risks could be licensed for deployment with vetted customers or limited public use.

OVERSIGHT ENTITY

Congress could give these oversight authorities to a new interagency coordinating body, a preexisting federal agency, or a new agency. Four potential options for this oversight entity:

A. Interagency Coordinating Body. A new, interagency body to facilitate cross-agency regulatory oversight, modeled on the Committee on Foreign Investment in the United States (CFIUS). It would be organized in a way to leverage domain-specific subject matter expertise while ensuring coordination and communication among key federal stakeholders.

B. Department of Commerce. Commerce could leverage the National Institute for Standards and Technology (NIST) and the Bureau of Industry and Security to carry out these responsibilities.

C. Department of Energy (DoE). DoE has expertise in high-performance computing and oversees the U.S. National Laboratories. Additionally, DoE has deep experience in handling restricted data, classified information, and national security issues.

D. New Agency. Since frontier models pose novel risks that do not fit neatly within existing agency jurisdictions, Congress could task a new agency with these responsibilities.

Regardless of where these authorities reside, the oversight entity should be comprised of: (1) subject matter experts, who could be detailed from relevant federal entities, and (2) skilled AI scientists and engineers. The oversight entity would also study and report to Congress on unforeseen challenges and new risks to ensure that this framework remains appropriate as technology advances.