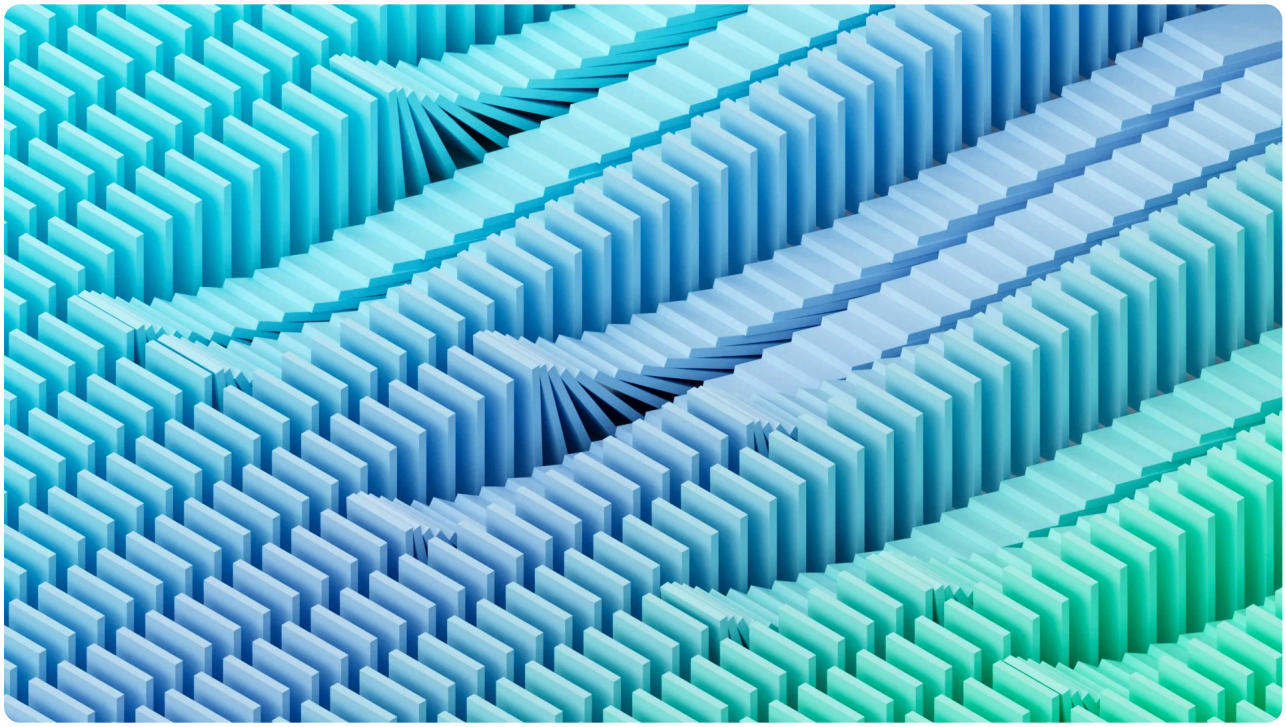
RESPONSIBILITY & SAFETY

Introducing the Frontier Safety Framework

17 MAY 2024

Anca Dragan, Helen King and Allan Dafoe

Share



Our approach to analyzing and mitigating future risks posed by advanced AI models

Google DeepMind has consistently pushed the boundaries of AI, developing models that have transformed our understanding of what's possible. We believe that AI technology on the horizon will provide society with invaluable tools to help tackle critical global challenges, such as climate change, drug discovery, and economic productivity. At the same time, we recognize that as we continue to

advance the frontier of AI capabilities, these breakthroughs may eventually come with new risks beyond those posed by present-day models.

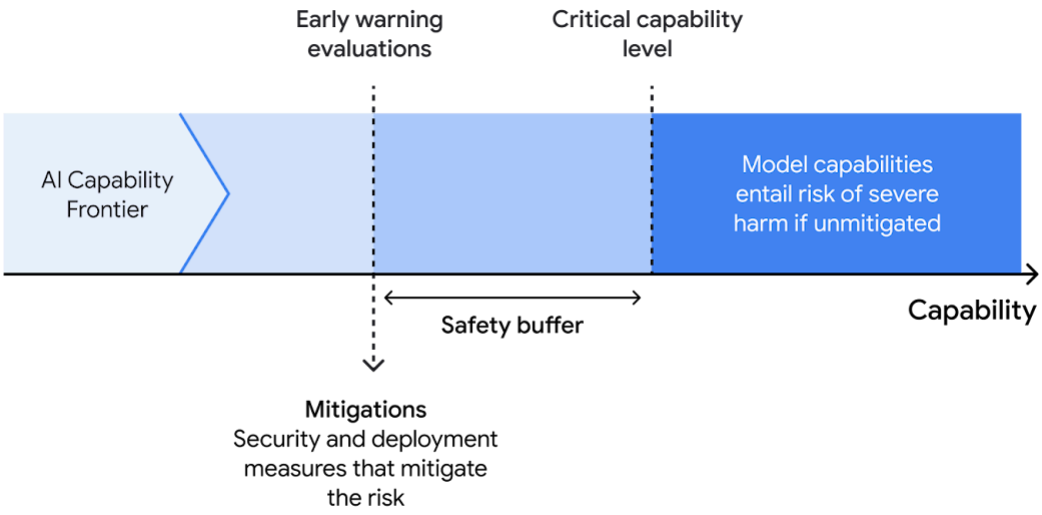
Today, we are introducing our Frontier Safety Framework — a set of protocols for proactively identifying future AI capabilities that could cause severe harm and putting in place mechanisms to detect and mitigate them. Our Framework focuses on severe risks resulting from powerful capabilities at the model level, such as exceptional agency or sophisticated cyber capabilities. It is designed to complement our alignment research, which trains models to act in accordance with human values and societal goals, and Google's existing suite of AI responsibility and safety practices.

The Framework is exploratory and we expect it to evolve significantly as we learn from its implementation, deepen our understanding of AI risks and evaluations, and collaborate with industry, academia, and government. Even though these risks are beyond the reach of present-day models, we hope that implementing and improving the Framework will help us prepare to address them. We aim to have this initial framework fully implemented by early 2025.

The framework

The first version of the Framework announced today builds on our research on evaluating critical capabilities in frontier models, and follows the emerging approach of Responsible Capability Scaling. The Framework has three key components:

1. **Identifying capabilities a model may have with potential for severe harm.** To do this, we research the paths through which a model could cause severe harm in high-risk domains, and then determine the minimal level of capabilities a model must have to play a role in causing such harm. We call these "Critical Capability Levels" (CCLs), and they guide our evaluation and mitigation approach.
2. **Evaluating our frontier models periodically to detect when they reach these Critical Capability Levels.** To do this, we will develop suites of model evaluations, called "early warning evaluations," that will alert us when a model is approaching a CCL, and run them frequently enough that we have notice before that threshold is reached.
3. **Applying a mitigation plan when a model passes our early warning evaluations.** This should take into account the overall balance of benefits and risks, and the intended deployment contexts. These mitigations will focus primarily on security (preventing the exfiltration of models) and deployment (preventing misuse of critical capabilities).



This diagram illustrates the relationship between these components of the Framework.

Risk domains and mitigation levels

Our initial set of Critical Capability Levels is based on investigation of four domains: autonomy, biosecurity, cybersecurity, and machine learning research and development (R&D). Our initial research suggests the capabilities of future foundation models are most likely to pose severe risks in these domains.

On autonomy, cybersecurity, and biosecurity, our primary goal is to assess the degree to which threat actors could use a model with advanced capabilities to carry out harmful activities with severe consequences. For machine learning R&D, the focus is on whether models with such capabilities would enable the spread of models with other critical capabilities, or enable rapid and unmanageable escalation of AI capabilities. As we conduct further research into these and other risk domains, we expect these CCLs to evolve and for several CCLs at higher levels or in other risk domains to be added.

To allow us to tailor the strength of the mitigations to each CCL, we have also outlined a set of security and deployment mitigations. Higher level security mitigations result in greater protection against the exfiltration of model weights, and higher level deployment mitigations enable tighter management of critical capabilities. These measures, however, may also slow down the rate of innovation and reduce the broad accessibility of capabilities. Striking the optimal balance between mitigating risks and fostering access and innovation is paramount to the responsible development of AI. By weighing the overall benefits against the risks and taking into account the context of model development and deployment, we

aim to ensure responsible AI progress that unlocks transformative potential while safeguarding against unintended consequences.

Investing in the science

The research underlying the Framework is nascent and progressing quickly. We have invested significantly in our Frontier Safety Team, which coordinated the cross-functional effort behind our Framework. Their remit is to progress the science of frontier risk assessment, and refine our Framework based on our improved knowledge.

The team developed an evaluation suite to assess risks from critical capabilities, particularly emphasising autonomous LLM agents, and road-tested it on our state of the art models. Their recent paper describing these evaluations also explores mechanisms that could form a future "early warning system". It describes technical approaches for assessing how close a model is to success at a task it currently fails to do, and also includes predictions about future capabilities from a team of expert forecasters.

Staying true to our AI Principles

We will review and evolve the Framework periodically. In particular, as we pilot the Framework and deepen our understanding of risk domains, CCLs, and deployment contexts, we will continue our work in calibrating specific mitigations to CCLs.

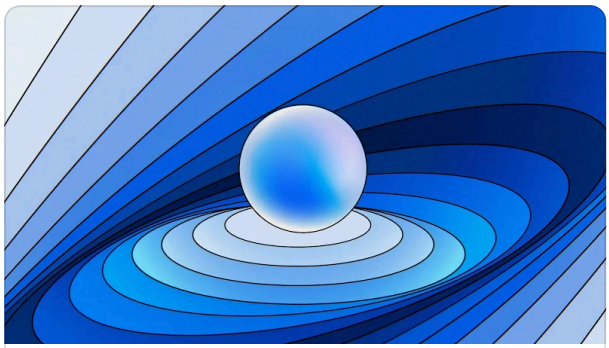
At the heart of our work are Google's AI Principles, which commit us to pursuing widespread benefit while mitigating risks. As our systems improve and their capabilities increase, measures like the Frontier Safety Framework will ensure our practices continue to meet these commitments.

We look forward to working with others across industry, academia, and government to develop and refine the Framework. We hope that sharing our approaches will facilitate work with others to agree on standards and best practices for evaluating the safety of future generations of AI models.

Read the technical report

Related posts

View all posts



RESPONSIBILITY & SAFETY

An early warning system for novel AI risks

New research proposes a framework for evaluating...

25 MAY 2023

Follow us

Build AI responsibly to benefit humanity

Models

Build with our next generation AI systems

- Gemini
- Gemma
- Veo
- Imagen
- Lyria

Science

Unlocking a new era of discovery with AI

AlphaFold

SynthID

WeatherNext

Learn more

About

News

Careers

Research

Responsibility & Safety

Sign up for updates on our latest innovations

I accept Google's Terms and Conditions and acknowledge that my information will be used in accordance with Google's Privacy Policy.

Email address

Google

About Google

Google products

Privacy

Terms

Manage cookies