AI Safety Fund

Brief | AI Safety Fund Solicits Initial Research Applications to Safeguard Frontier AI, Welcomes Participation for Next Round

About the AI Safety Fund

The AI Safety Fund (AISF) is a $10 million+ initiative, born from a collaborative vision of leading AI developers and philanthropic partners. Funders include the Frontier Model Forum's founding members Anthropic, Google, Microsoft, and OpenAI, with support from philanthropic partners, including the Patrick J. McGovern Foundation, the David and Lucile Packard Foundation, Schmidt Sciences, and Jaan Tallinn.

Administered independently by Meridian Prime, the AISF awards research grants to independent researchers to address some of the most critical safety risks associated with the proliferated use of frontier AI systems.

The purpose of the fund is to support and expand the field of AI safety research to promote the responsible development of frontier models, minimize risks, and enable independent, standardized evaluations of capabilities and safety. We seek to attract and support the brightest minds across the AI ecosystem to build frontier models aligned with human values.

The AI Safety Fund supports research on state-of-the-art, general purpose AI models. Funding will only be awarded to projects researching deployed versions of those models.

Funding Opportunities and Research Priorities

The AISF will make an initial round of grants to a diverse group of researchers investigating Frontier AI safety, which will be announced in July 2024. These solicited grants will include researching methods for evaluating the capabilities and risks of frontier models, including evaluations, red-teaming, and benchmarking. This initial round will further the AISF's objective of supporting and expanding AI safety research to enable independent, standardized evaluations of frontier AI capabilities and risks.

Open call grant rounds will be announced in the coming months and will prioritize technical research in three core areas related to frontier AI models:

- Identifying safety-critical risks posed by frontier models
- Evaluating and assessing strategies for addressing such risks
- Implementing mitigations to prevent these risks from occurring.

AI has tremendous potential to benefit the common good. Ensuring we can harness that good requires appropriate testing, evaluation, and best practices to mitigate risks. From evaluations of dangerous capacities to ensuring alignment with human values, independent research is a critical element to ensuring frontier AI is developed and deployed safely. Research outcomes will be publicly available on the website and additional opportunities to share work funded by the AISF will be considered with guidance from an Advisory Committee.

AI Safety Fund

Tell us about your research

The AISF is committed to supporting research from diverse voices across the AI safety field and is eager for input on its intended research agenda. Researchers working on frontier AI safety from across technical disciplines who would like to share insights or inquire about how to participate should complete the research interest form here.

For more information, please visit AISFund.org.