Call for Research Ideas: Risks From Internal Deployment of Frontier AI Models

Foundational Research Grants (FRG) is calling for research ideas that would clarify and manage risks from deployment of frontier models within AI companies. FRG, a grant program based within the Center for Security and Emerging Technology (CSET) at Georgetown University, funds work on technical topics related to the national security implications of AI.

At a glance

- **Topic:** Risks from internal deployment of frontier AI models
- **Award size and duration:** Up to $1,000,000 per project, to be expended over 3-24 months
- **Selection process:** 1-2 page expression of interest due **June 30, 2025**; more details below

Research scope

We are interested in supporting research into **understanding and managing risks** from **internal deployment** of **frontier AI models**.

By **"internal deployment,"** we mean availability of an AI system for use exclusively within the company that developed it (Stix et al. 2025). By **"frontier models,"** we mean general-purpose AI models with cutting-edge capabilities.

Background and motivation

For most technologies, including many types of AI, significant risks emerge only after external deployment. For example, narrow machine learning systems for facial recognition only violate privacy when they are integrated into surveillance systems. In a small number of areas, such as lab work involving dangerous chemicals or contagious biomaterials, risks can arise even during the development phase of a new technology.

Frontier AI appears to fall into the latter category. Advanced general-purpose AI systems are often used internally at AI companies for months before public release, for tasks including generating code, running experiments, providing strategic advice, drafting internal documents, and monitoring other AI systems. Internally deployed models may have different safeguards and affordances than externally deployed models (Stix et al. 2025). By default, models undergo third-party safety testing only before external deployment and there is little oversight from outside the company before then. As a result, risks that arise during internal deployment materialize when models' capabilities and propensities are least understood.

This creates a distinct risk landscape with several potential pathways to severe harm. Internal models that exceed publicly available capabilities become valuable targets for theft or sabotage by adversaries. If misalignment risk increases with capability, problems could first emerge inside companies rather than after public release (Acharya and Delaney, forthcoming). Since AI companies use internal models for research and development, any flaws in these systems—whether from sabotage, misalignment, or reliability issues—may propagate into future generations of frontier AI (Acharya and Delaney, forthcoming, Clymer et al. 2025, Benton et al. 2024). Yet another risk pathway is that people with privileged access to internal frontier models could exploit that access in ways that benefit them and harm or disempower others (Stix et al. 2025, Davidson et al. 2025).

We hope that research examining internal risks will help develop practical threat models, monitoring approaches, and risk mitigations. This work could help companies and governments to create clearer protocols for managing AI systems before they reach the public, potentially preventing problems before they occur.

Topics of interest

The goal of this call for research ideas is to create and/or improve tools and frameworks for understanding and managing these risks. Some categories of work that we could be interested in supporting include:

- Threat modeling to **identify and assess** risks from internal deployment.
    - What are pathways for internally deployed models to cause harm?
        - How could a compromised (e.g. sabotaged or misaligned) model used for R&D affect the next generation of frontier models? For example, what are plausible ways that a model could falsify the results of safety experiments (Clymer et al. 2025), fail to report on other models' harmful behaviors (Clymer et al. 2025), or generate adversarial synthetic training data (Vassilev et al. 2024)?
        - What tools, permissions, capabilities, etc. would an internally deployed model need access to in order to cause harm? For example, what are plausible ways that a model could self-exfiltrate (with or without human assistance) or create unauthorized internal deployments (Shlegeris 2024)?
        - How could internally deployed models be *misused* to cause harm (e.g. humans seizing power via differential access to more powerful models (Stix et al. 2025))?
        - How could internally deployed models be *sabotaged* (e.g. backdoored) by adversaries?
        - How could internally deployed models *accidentally* cause harm (e.g. via AI company employees over-relying on them)?

    - How likely are these pathways to harm? How do they intersect with each other?
    - How might these risks change over time? What factors are involved? For example, can we expect pre-release periods for frontier models to lengthen or shorten (where Safe Superintelligence is an extreme example of a long pre-release period)?
- Approaches for **monitoring** risks from internal deployment.
    - How can AI developers evaluate whether models are capable of executing on various risk pathways?
    - How can risk *factors* for internal deployment (such as heavy reliance on internal systems for R&D) be operationalized into risk *indicators* that can be monitored?
    - How should companies monitor these indicators?
    - How should third parties be involved in monitoring these indicators?
    - To the extent that AI models are used for monitoring, how can the overall monitoring system account for the possibility of the AI monitors themselves being compromised?
- **Mitigations** for risks from internal deployment.
    - How can we detect and control *scheming* in the context of internal deployment?
    - How can we detect and remove *backdoors* or other forms of sabotage in the context of internal deployment?
    - How can systems security approaches, control approaches, and related ideas mitigate internal deployment risks?
    - What *tradeoffs* are there between addressing internal deployment risks and other valuable goals, and how can these tradeoffs be reduced? For example, how should (potentially costly) mitigation measures scale proportionally with risk, a la responsible scaling policies?

Within these bounds, we are open to a wide range of possible research approaches, topics, and formats.

Individual projects can request up to $1,000,000 for relatively resource-intensive projects, but we are also open to providing much smaller sums for simpler or shorter projects, e.g. writing position papers addressing the kinds of questions covered above. Since this is a fast-moving field, we are also open to research ideas that may evolve significantly over time. Our key consideration will be that projects remain relevant to the topic of this call for research ideas: understanding and managing risks from internal deployment of frontier AI models.

Submission guidelines and process

Award eligibility, size, and duration

- **Eligibility:** Principal investigators should be based at an institution of higher education or a nonprofit research organization. International institutions are welcome to apply, though they may face additional administrative burdens if awarded grants.
- **Award size:** up to $1,000,000 (including no more than 10% indirect costs)
- **Duration:** 6–24 months

Process for submissions

Stage 1: Short expression of interest (due June 30, 2025). Initial submissions should briefly (in 1-2 pages) describe a potential research project relevant to the scope described above. This submission does not need to lock in final details of the project, but should explain roughly what you intend to work on. Be sure to clearly include:

- Table of key figures, including: key personnel, host institution, rough cost estimate:

Principal Investigator	
Other Key Personnel	
Host Institution	
Rough Cost Estimate	
Anticipated Period of Performance (duration)	

- Goals of research: ≤200 words
- Research approach: ≤200 words
- Relevance of the work to this call for research ideas: ≤300 words
    - This section should draw explicit connections to the “topics of interest” section of this document.

Send expressions of interest to <csetfoundations@georgetown.edu> with subject line: “Response to 2025 call for research ideas”.

Stage 2: Discussion and development. Selected applicants will be invited to discuss and develop their research idea further with the FRG leadership team over the course of 1–3 informal calls.

Stage 3: Proposal (due September 15, 2025). Selected applicants will be invited to submit a short proposal (around 4-6 pages) including:

- Project scope and goals
- Milestones and timeline
- Budget and brief budget narrative

Stage 4: Final selection. Reviewers will evaluate proposals and select projects to be funded based on feasibility and relevance.

Timeline for submissions

1. Expression of interest: due June 30, 2025
2. Discussion and development: June 30 through August 4, 2025
3. Proposal: due September 15, 2025
4. Final selection: by October 14, 2025

Criteria for selection

Submissions will primarily be evaluated in terms of (a) their relevance to the research scope described in this document and (b) technical feasibility and merit.