Call for research ideas: AI assurance for general-purpose systems in open-ended domains

Foundational Research Grants (FRG) is calling for research ideas on assuring general-purpose AI systems that operate in open-ended domains. FRG, a new grant program based within the Center for Security and Emerging Technology (CSET) at Georgetown University, funds work on technical topics that bear on the potential national security implications of AI over the long term.

At a glance

- **Topic:** The feasibility of assuring increasingly general-purpose AI systems in increasingly open-ended domains
- **Award size and duration:** Up to $750,000 per project, to be expended out over 6-24 months
- **Selection process:** 1-2 page expression of interest due **August 1**; more detail on full selection process below

Research scope

We are interested in supporting research that **sheds light on how feasible it will be to assure deep learning systems that are increasingly general-purpose and operating in increasingly open-ended domains** (more on this below).

We're using **"assurance"** here in a broad sense, meaning roughly **"the generation of evidence that an ML system is sufficiently safe for its intended use."**[1]

Background and motivation

Machine learning (ML) systems—especially deep learning systems—are rapidly becoming larger, more complex, more capable, and more general-purpose. It is becoming more common to deploy AI systems in open-ended domains, including as chatbots, coding assistants, personal assistants, and tutors. At the same time, these AI systems still show major flaws and vulnerabilities.

A wide range of assurance techniques exist that can be applied to systems with automated or autonomous capabilities.[2] But these techniques are generally either designed to assure traditional software (which is structured very differently from ML models) or autonomous systems carrying out specific tasks in specific environments (which can be thoroughly tested under specific assumptions).

[1] Paraphrasing Ashmore, Calinescu and Paterson 2019.

[2] See e.g. Maksimov 2018 or Gleirscher and Kugele 2019 for overviews of approaches to assuring safety-critical systems in general, or Ashmore, Calinescu and Paterson 2019 on assuring ML systems.

There do not appear to exist established approaches to assurance that are well suited to the kinds of large-scale deep learning models that are currently being developed and deployed in open-ended settings. The goal of this call is to clarify to what extent and how rapidly we might expect to develop assurance approaches that are suitable for such models and applications.

CSET's mission is to inform policymakers about the national security implications of emerging technologies. As such, **the underlying goal of this call for research ideas is to help policymakers and decisionmakers understand the assurance prospects of increasingly general-purpose systems in open-ended domains.** How assurable we should expect such systems to be is a key question in anticipating and shaping their real-world implications: is there a clear path towards being able to use such systems in high-stakes and safety-critical settings, or should we assume that this will be inadvisable for the foreseeable future?

Topics of interest

By using a broad definition of AI assurance, **we seek to include a range of possible research directions, with a focus on the practical implications of the work rather than on a specific disciplinary lineage.** Proposals of interest could come from a range of computer science subfields—including assured autonomy, ML robustness, ML interpretability, and others—as well as fields outside of computer science that focus on assurance-related problems.

We are looking for research that explicitly grapples with the challenge of developing assurance techniques and approaches that will continue to be useful as ML systems become more general-purpose and are used in more open-ended domains. This could include work that:

- Makes the case that a certain assurance approach will (or won't) continue to be useful as ML systems increase in scale, complexity, and generality, and are used in increasingly open-ended settings[3]
- Analyzes how assurance, reliability, trustworthiness, etc. for general-purpose ML systems in open-ended domains in the future (e.g. 10 or 20 years from now) will compare to our current standards for safety-critical systems, perhaps presenting a *best-guess* and *best-case* scenario[4]
- Analyzes whether specific challenges facing AI assurance efforts (e.g. adversarial examples, hallucinations, etc.) will continue to be a major challenge, i.e. whether they are more likely to be the kind of "hiccups" regularly seen in immature technologies that will be ironed out over time, vs. fundamental weaknesses of this kind of system[5]

[3] For instance, a submission could analyze the prospects of mechanistic interpretability for assuring general-purpose systems in open-ended settings.
[4] This could build, for instance, on Lohn 2020.
[5] See e.g. Shamir et al. 2019 for an analysis of why adversarial examples are widespread.

- Assesses how established testing and evaluation practices will (or won’t) need to change in order to account for more general-purpose AI systems operating in more open-ended environments[6]
- Assesses the extent to which existing approaches to systems safety can ameliorate concerns associated with large-scale deep learning models[7]
- Assesses novel assurance challenges posed by—or develops novel assurance approaches for—multi-modal models or large language models, especially when used in increasingly open-ended settings

This call for research ideas is motivated by our sense there is a gap between the assurability of small, specialized ML-based components trained to perform a specific function in a specific context, vs. the assurability of large, cutting-edge deep learning systems being used in open-ended contexts for a wide range of purposes (including, but not limited to, large language models or ”foundation models”). In addition to the examples in the list above, we would consider submissions that make the case that this gap exists and will persist; that the gap exists but is likely to be closed by relatively straightforward extension of existing techniques to newer models; that the gap exists and could be closed by new approaches; or that the gap doesn’t exist in the first place.

We are less likely to support:

- Research aiming to improve state of the art on an existing assurance/safety/trustworthiness benchmark, unless that is explicitly connected to an argument with implications for assurance of general-purpose systems in open-ended domains
- Research on assuring specific categories of ML models, such as particular architectures or narrow use cases, unless applicability to or lessons for assuring larger, more capable, more general-purpose models is explicitly considered
- Surveys of existing assurance approaches, unless surveyed explicitly through the lens of how different approaches do and don’t apply to challenges posed by general-purpose systems in open-ended domains

[6] See e.g. Khlaaf 2023, especially section 4.

[7] See e.g. Jatho et al. 2023

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
- **Award size:** up to $750,000 (including no more than 10% indirect costs)
- **Duration:** 6–24 months (ending by December 2025)

Process for submissions

Stage 1: Short expression of interest (due August 1, 2023). 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 research project, but should explain roughly what you would be interested 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 2023 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 October 24, 2023). 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 August 1, 2023
2. Discussion and development: August 1 through September 12, 2023
3. Proposal: Due October 24, 2023
4. Final selection: Selections made by December 5, 2023

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.