Advanced Research + Invention Agency ARIA

August 2024

Request for Expressions of Interest

Summary: *This is an early opportunity for expressions of interest from individuals or organisations to be involved in the development of an organisation spearheading the research & engineering that makes up Technical Area 2 of the Safeguarded AI programme.*

***

The Safeguarded AI programme is a £59m-backed R&D effort to build out a general-purpose AI workflow for producing domain-specific AI agents or decision-support tools for managing cyber-physical systems with quantitative guarantees which improve upon both performance and robustness compared to existing operations.

Technical Area 2 (TA2) of the programme will develop the ML elements which harness frontier AI techniques into a general-purpose Safeguarded AI workflow. ARIA is looking for new teams to spearhead the development of this technology in the UK within a single non-profit organisation, which could be a new entity or an entity affiliated to an existing organisation.

In advance of the first funding call going live, we are openly seeking very lightweight initial expressions of interest from:

a) *individuals* who are interested in being involved in a potential founding team, or
b) *existing organisations* (such as established AI labs, critical infrastructure companies, or others with a differentiated value proposition) that may wish to help initialise the new organisation as an affiliated entity.

We encourage you to submit if you are interested in being involved in the TA2 effort in any capacity, be that as an affiliated entity, a partner organisation, an organisational cofounder, or as an individual in a technical, executive or advisory role. As a result of your submission, we might reach out to you or make appropriate introductions.

About how we will fund TA2

Later this year, we will launch an open funding call for a preliminary stage ("Phase 1"). At this stage, we will select multiple groups to support for a short period of time (~4 months) to develop a detailed full proposal outlining the proposed **entity's structure, mission statement, governance mechanisms, security measures, economic model, management model, recruiting plan, external fundraising avenues, etc**. This will be followed by a Phase 2 funding call (to which those proposals will be submitted), where we will select a single awardee to pursue the R&D for TA2.

The allocation of ARIA's funds will be conditional on proposals credibly showcasing their ability to undertake such an R&D effort, while at the same time having highly robust governance mechanisms to align the decision-making of the organisation with the programme's mission to ensure that AI systems are developed and deployed in service of humanity at large. The effort will be supported by the Safeguarded AI programme team, in particular David 'davidad' Dalrymple (Programme Director) and Yoshua Bengio (Scientific Director).

The details of the TA2 organisation are yet to be fully determined as part of the preliminary stage ("Phase 1"), as mentioned above. However, some characteristics that we strongly expect the organisation to have include:

+ Based in the United Kingdom
+ Credible ability to source world-class talent in machine learning research & engineering
+ Funded by one or more other sources before the end of the programme
+ Robust governance mechanisms, amongst others, a diverse board with the sole mission of ensuring that decisions concerning the development, deployment and release of its AI technologies — including algorithms, models, code, products or API access — are made in service of humanity and society at large
+ World-class cybersecurity

+ Flexibility to pursue globally focused collaborations, multilateral information-sharing and strategic partnerships with international counterpart entities—if and only if determined to align with the mission

Technical objectives of TA2

The technical objectives of the organisation will be along these lines (also see Technical Area 2 in the Programme Thesis for more detail):

Objective (a): World-modelling ML. We want to increasingly assist and automate the representation of human knowledge in a formalised way that admits explicit reasoning, including accounting for various forms of uncertainty.

Objective (b): Coherent reasoning ML. In order to practically rely upon a world model to guarantee safety in a complex environment, we need efficient ways to reason and derive correct conclusions from the world model.

Objective (c): Safety verification ML. An important use of the world model (a) and the reasoning machinery (b) is for verifying that a given action or plan is safe according to a given safety specification.

Objective (d): Guaranteed RL Policy training. An agent policy should be trained to achieve finite-horizon safety guarantees, taking advantage of the capabilities developed in objectives (a,b,c).

Technical work on TA2 is to be done in a secure environment, with serious measures in place to avoid leaks of model weights (or even leaks of most concrete algorithmic ideas). Patents may be filed without a patent non-aggression pledge if the TA2 entity sees fit, but most patentable inventions in TA2 should more likely be protected as trade secrets.

International and Inter-Institutional Collaboration

In line with the mission to ensure that AI systems are developed and deployed in service of humanity at large, we envision and seek to foster multilateral technical collaboration, both

nationally and internationally, in order to drive progress, ensure interoperability and the sharing of safety-critical information, enable global deployment of Safeguarded AI workflows and avoid undue incentives to race ahead at the expense of safety. We hope, through this programme, to nucleate opportunities to develop such international collaborations, pioneered by the UK.

During the programme itself, we are exploring the following specific forms of collaboration, particularly with interested ATAS-exempt countries:

+ **Joint workshops**: Researchers funded by this programme attending workshops hosted by international counterpart organisations, and vice versa.
+ **Extended visits** to facilitate research collaborations between our programme and research funded by related programmes of counterpart agencies.
+ **Joint working groups**: Researchers in TA1 with common interests across international efforts meeting regularly online and exchanging notes.
+ **Information-sharing arrangements**: In TA2, where all research will be conducted in a single UK-based entity with strong security and oversight procedures, much research will not be public by default. However, we will encourage the TA2 organisation to form partnerships with international counterpart entities as they arise, particularly for sharing information that is important for the interoperability, compatibility, and joint safety of their systems.

We see our efforts towards the development of safe and beneficial AI as deeply complementary to the mission and work spearheaded by the UK's AI Safety Institute (AISI). While AISI is typically focused on technical work that very directly informs governance and policy, ARIA's focus is on fundamental research that has the potential to unlock novel safety and governance avenues in 3-10 years. We also plan to participate in informing AISI's forward-looking analysis of "safety cases" for Safeguarded AI, and other proposed Guaranteed Safe AI methods. We believe all of these efforts are jointly needed to safely navigate the development and use of advanced AI capabilities throughout the next decade and beyond.