Advanced Research + Invention Agency ARIA

Safeguarded AI: TA2 Phase 2 – Machine Learning Organisation

Call for proposals

Date: 25 June 2025

Final

Advanced Research + Invention Agency ARIA

Advanced Research + Invention Agency ARIA

Technical Area 2, Phase 2 - Call for Proposals

Summary of Safeguarded AI, Technical Area 2

Backed by £59m, the Safeguarded AI programme aims to develop provable safety guarantees for transformative AI through a novel approach — we envision a secure production facility where frontier AI capabilities remain contained, while being harnessed to produce valuable outputs: narrower AI applications with mathematical guarantees of safety in their contexts of use. This approach would enable the use of advanced AI in safety- and business-critical domains where it's currently too risky to deploy, focusing particularly on enhancing the resilience and performance of critical infrastructure.

Technical Area 2 (TA2) will develop the machine learning (ML) elements which are needed to harness frontier AI techniques into a general-purpose Safeguarded AI workflow. We plan to make a £18m award to a single entity pursuing this research agenda. Because of the global significance of these capabilities (if developed successfully) and the potential externalities in case of insufficient risk management or security, we require the entity ultimately hosting the TA2 research agenda to also push the frontier on organisational governance and security standards. We welcome applications for new founding teams or existing entities looking to create a new affiliated non-profit entity.

Following our initial Phase 1 call for proposals, we are now opening applications for Safeguarded AI's Technical Area 2 Phase 2. In this funding call, we are looking to award a single **£18m** grant to one entity to host the entire TA2 research agenda.

In Phase 1, we have funded a handful of teams to develop a full proposal. Nevertheless, applications to Phase 2 are open to everyone, including teams who have not received a Phase 1 funding. The application deadline for the Phase 2 call is **01 October 2025**.

In this solicitation, we will focus on information specific to submitting a Phase 2 proposal. If you are a prospective applicant, **we strongly encourage you to consult the Phase 1 Funding Call for information including:**

- An overview of the Safeguarded AI programme (Section 1)
- Information about what and how we are funding in Technical Area 2 of the programme (Section 2)
- Details on Technical Area 2's objectives, including technical and organisational objectives (Section 3)

TA2 Phase 2 — Structure & Management

TA2 Phase 2 will span from the beginning of 2026 to the end of the programme, EOY 2027. Throughout this period, the successful TA2 team will report against a series of technical milestones. An initial version of the milestones will be proposed by applications as part of their application, and then refined and finalised together with the programme team at the start of the project. Additionally, ARIA's standard project management requirements include light touch quarterly reporting on progress and cost information.

Throughout the project period, the programme team will also facilitate interactions with Creators across the rest of the Safeguarded AI programme, including quarterly Creator events, among others.

Approach to Intellectual Property

In TA2 Phase 2, technical work will be conducted in a secure environment, with serious measures in place to avoid leaks of e.g. model weights or concrete algorithmic ideas, such as the measures discussed in this report.[1] We are not putting up any requirements on how IP is handled. Instead, applicants are asked to propose, as part of Phase 2 applications, how to best handle IP in line with TA2's overarching mission.

Eligibility

To deliver the TA2 agenda, we are looking for exceptional and ambitious researchers, organisational leaders or experienced founders who are driven by the idea of developing an alternative R&D pathway toward safe and transformative AI.

[1] Nevo, Sella, et al. "Securing AI model weights." Research reports, RAND (2024).

With respect to the **entity** which will ultimately deliver the TA2 research agenda, **necessary requirements** are:

+ Based in the United Kingdom
+ Credible ability to source world-class talent in machine learning research & engineering
+ Robust governance mechanisms, including (among 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 multilateral information-sharing and strategic partnerships with other private and/or government-sponsored entities— if and only if determined to align with the mission

In addition to ARIA's standard eligibility criteria here, the following types of entities are **not** eligible for funding to deliver TA2[2]:

+ For-profit companies
+ Universities directly hosting TA2

Based on these eligibility criteria, a non-exhaustive list **types of applicants** eligible for a TA2 award includes:

+ New founding teams with a credible skillset and interested in quickly establishing a new UK-based non-profit institution from the ground up;
+ Leading AI companies willing to create a UK-based affiliated[3] non-profit entity to host the TA2 R&D agenda, expanding the market for their AI capabilities into multiple critical infrastructure sectors;
+ Established companies with existing critical-infrastructure businesses willing to create a UK-based affiliated[3] non-profit entity to become a pioneering supplier of guaranteed-safe AI capabilities; or

[2] We exclude these types of applicants because we don't expect them to meet our desiderata for robust organisational governance and security.

[3] An affiliated entity, in these examples, could involve shared branding, shared personal, financial support, licensing agreements; but must have separate boards and legal structures.

+ Established academic institutions willing to create, or partner in creating, a new UK-based affiliated[3] non-profit entity[4], where TA2 R&D can be pursued under conditions of first-of-class information- and cyber-security.

For Non-UK applicants only

The entity that will host the TA2 R&D agenda will be based in the UK. Non-UK citizens are welcome to apply to pursue this work, but need to be prepared to relocate to the UK.

For non-UK citizens, we have provided some additional guidance in our FAQs including available visa options.

Application Submission Guidelines

The Phase 2 application is expected to span 30-50 pages, and include the following sections:

- **1-page executive summary** (included in 50 page limit)
    - Provide a concise, high-level overview of your application with any key information you wish to highlight.
- **Project information** (included in 50 page limit)
    - A detailed discussion of the five **Technical Application Questions** (see below) (up to 10 pages)
    - A detailed discussion of the nine **Organisational Application Questions** (see below) (up to 40 pages)
    - A discussion of what the major obstacles are you foresee, that could prevent successful execution of the TA2 vision, both technically and organizationally
- **Team information** (included in 50 page limit)
    - An overview of the key team members involved in Phase 2, their respective roles, areas expertise and track record
    - Information about any outstanding talent you're planning to hire, and your plan for attracting/hiring that talent (including talent for your technical team). Ideally, this looks like concrete **evidence of conditional employment commitment** by these individuals (e.g. in the form of a letter stating as much).
    - A discussion of why you/your team are motivated to solve this problem

[4] Illustrative examples include the UK's The Francis Crick Institute or the Alan Turing Institute.

- **Evidence of conditional funding commitment** (outside of 50 page limit)
    - Credible evidence of (conditional) funding commitments from non-ARIA sources of a minimum of 20% of ARIA's Phase 2 funding (i.e. at least £3.6m external funding committed conditional on ARIA's Phase 2 funding being awarded).
- **Administrative questions** (outside of 50 page limit)
    - These help us ensure we are responsibly funding R&D. Find an overview of all questions here. These questions will be answered via the application portal and do not count towards the 50 page limit. When completing the cost information, please refer to ARIA's standard eligible cost guidance.

Proposals have to be formatted in accordance with the following guidelines:

- Page count: up to 50 pages of A4 (including diagrams, excluding references)
- Font: Garamond, Computer Modern, or Arial. Colour: black. Size: 11-point font or larger
- Margins: At least 0.5" margins all around
- File Type: PDF

Please refer to the section below for guidance on how to submit your application.

Technical & Organisational Application Questions

Note that these are the same set of questions as discussed in the Phase 1 Funding Call, in Sections 3A and 3B.

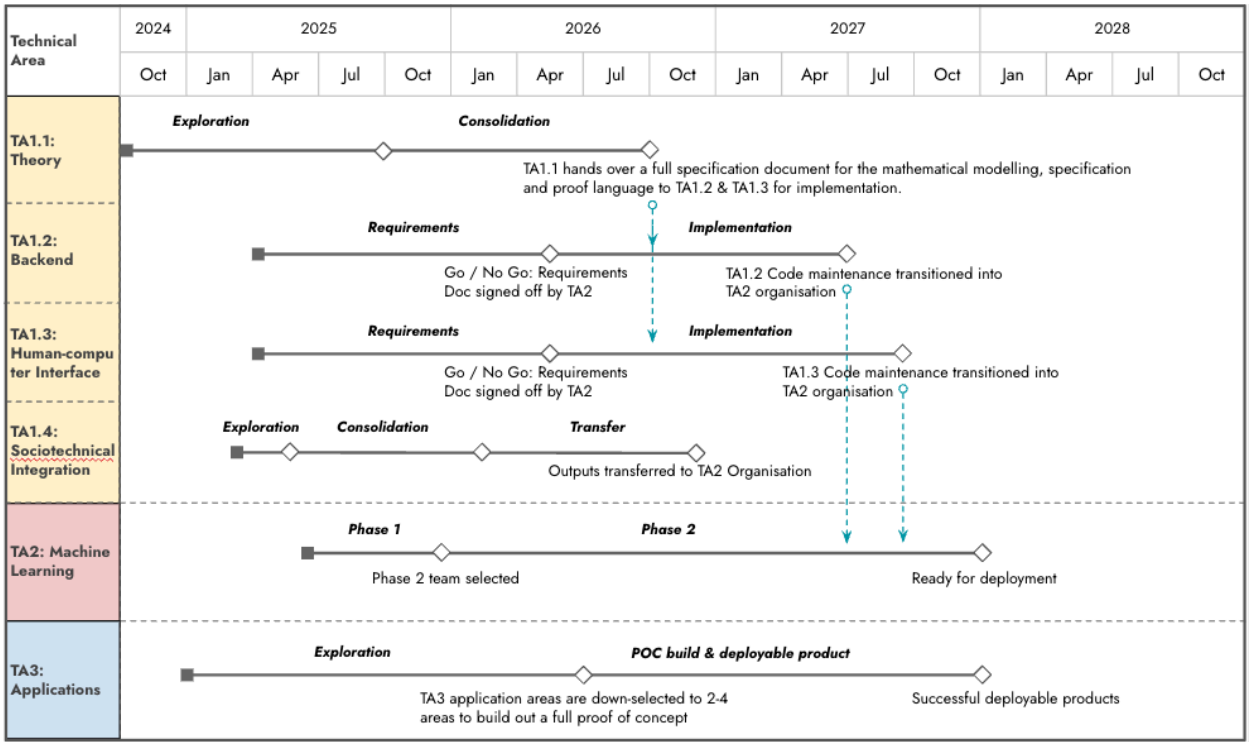
A. Technical Application Questions

1. What are the research directions you plan to try and/or hypotheses you plan to test during Phase 2? Provide arguments for each one about why it is promising, and why it is challenging. Feel encouraged to list many research directions/hypotheses, knowing that you will want to drop most of them as you learn what does/doesn't work.
2. How will you make use of increasingly automated AI R&D capabilities to bootstrap your research agenda even faster?
3. How do you plan to allocate budget, compute and human resources across the four technical objectives (TA2 a-d) throughout the programme duration (over the course of 2026 and 2027)? For interactions with the rest of the programme, see the

following key delivery milestones for other Technical Areas, also visualised in the figure below.

a. Delivery milestones for TA1
    i. May 2026: TA1.2 & TA1.3 have their initial requirement documentation signed off by TA2.
    ii. Sep 2026: TA1.1 hands over a full specification document for the mathematical modelling, specification and proof language to TA1.2 & TA1.3 for implementation.
    iii. Aug/Sep 2027: TA1.2 & TA1.3 complete their implementation work. Code maintenance is handed over to the TA2 hosting entity.
b. Delivery milestones for TA3
    i. Jul 2026: The TA3 application areas are down-selected to 2-4 areas to build out a full, deployable product.
    ii. EOY 2027: TA3 teams deliver the deployable products or real pilot deployments for their respective application domains.

4. What quantitative metrics do you propose for each of the four technical objectives? For each of these metrics, establish a baseline using off-the-shelf tools, to permit evaluating your Phase 2 progress against.
5. How do you predict that these metrics will develop over the course of the 2 years as a result of your work, as well as accounting for generic AI capabilities improvements (assuming no slowdown)?



B. Organisational Application Questions

1. What will be the **mission statement** of the organisation?
2. What will be the **legal and organisational structure** of the organisation? Among others:
    a. What type of legal entity will you choose, and why, to ensure that future profit incentives will not compromise with the organisation's mission? Some possible candidates of UK legal structures include: trust, unincorporated association, community interest company limited by guarantee, charitable foundation, charitable incorporated organisation, etc.
    b. Where certain organisational characteristics are important to the success of the TA2 R&D effort (e.g. operating as a non-profit), how will you ensure that they are, as far as possible, immutable?
    c. What will the board structure be, and what will be the board's powers and responsibilities? How will the board selection and replacement procedure be designed? How will you make sure the incentive structures acting on the board will effectively and reliably enable board members' to fulfil their responsibilities under the charter?

3. How will you secure additional **external funding**? How do you envision the **economic model** of this organisation in the long term? Among others:
    a. What is your plan for securing sustainable, longterm non-ARIA funding and/or investments (both during and after the programme period) to enable the successful pursuit of the R&D agenda and independence from ARIA?
        i. In particular, **for Phase 2 applications, we require evidence of (conditional) funding commitments of a minimum of 20% of ARIA's Phase 2 funding** (i.e. at least £3.6m external funding conditional on ARIA's Phase 2 funding being awarded). This funding would cover, among others, organisational expenses that fall outside of the research effort, as well as to further strengthen the R&D budget available to pursue TA2's research objectives.
    b. What (if any) model for economic returns do you plan to adopt, and why? How will you ensure your economic model fits with your organisation type and structure (and overall non-profit status) and your plans for external funding?
    c. How will you decide when an invention should be protected and how (e.g. by patent versus trade secret, etc.)? How will you decide when to release inventions openly in the public interest? What (if any) licensing scheme will you adopt for TA2 capabilities developed by the organisation? (You can treat ARIA's standard IP terms as the default, and make a case for how and why (if at all) you propose to deviate from these terms.)
    d. How will you make sure that any TA2 IP is and will not, at any point, become alienated from the entity and its original mission, e.g. due to restructurings or through successor entities/spin-offs or by exclusively licensing it to a different entity?
4. How do you envision the needs for and provision of appropriate **security**, including cyber- and information security? What aspects of this organisation's future work do you think should be especially protected?
5. How are you going to set the **incentive structures** of the organisation and its various stakeholders such that they robustly align with the mission of the organisation, and the net benefit of humanity at large? Among others:
    a. How do you envision setting those incentives for your team/staff?
    b. How do you envision accessing sufficient compute and financial means for the successful pursuit of the TA2 R&D effort, without compromising the integrity of your governance structures?

6. What is your plan for **recruiting** the required top-tier scientific and engineering talent? If the applicant team does not already have them, what is your plan for recruiting a full-time CEO?
7. How do you envision the **management & operational model** of the organisation, including internal structures, roles and responsibilities, project management, culture and office location?
8. How will you ensure that the organisation will (continue to) be **an asset to the UK**? Among others:
    a. How do you envision relationships with other relevant organisations in the UK, for example the UK AI Security Institute (AISI) or the AI Research Resource (AIRR)?
    b. How will you ensure that the majority of TA2 R&D work & its cost pursued by the organisation will (continue to) be physically located in the UK?
9. How do you envision productive avenues for **international cooperation** 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? Among others:
    a. How do you envision relationships with other relevant non-UK organisations, including international AI Safety/Security Institutes, international counterpart organisations with compatible mission statements, and leading for-profit AI labs?
    b. How do you envision the potential for collaborations with interested ATAS-exempt countries, e.g. in the form of joint workshops, extended visits, joint working groups or information-sharing arrangements?

Evaluation Criteria

Proposals will pass through an initial screening and compliance review to ensure proposals conform to the format guidance and they are within the scope of the solicitation. At this stage we will also carry out some checks to verify your identity, review any national security risks and check for any conflicts of interest. Prior to review of applications, Programme Directors and all other reviewers are required to recuse themselves from decision making related to any party that represents a real or perceived conflict. Where it is clear that a proposal is not compliant and/or outside the scope, these proposals will be rejected prior to a full review on the basis they are not compliant or non-eligible. In conducting a review of the Phase 2 proposals we'll consider the following criteria:

1. **Credible ability to deliver the TA2 technical R&D agenda**, based on:
    a. the applicant's answers to the technical questions in section 3A
    b. the technical team's fit and relevant track record, and credible ability to hire further top technical talent
2. **Credible ability to embed the technical work in a compelling organisational structure, including the legal & economic model, governance and security**, based on:
    a. the applicant's answers to the the governance questions in section 3B
    b. the applicant's ability to secure additional (conditional) funding commitments of a minimum of 20% of ARIA's Phase 2 funding (i.e. at least £3.6m external funding committed conditional on ARIA's Phase 2 funding being awarded)
3. **Intrinsic motivation and team fit**, based on:
    a. the applicant's ability to demonstrate deep problem knowledge, advanced skills in the proposed area and intrinsic motivation to pursue the project and broader mission of TA2
4. **Benefit to the UK** – There is a clear case for how the project will benefit the UK through social and economic impact. Strong cases for benefit to the UK include proposals that:
    a. are led by an applicant within the UK who will perform the project within the UK
    b. are led by an applicant outside the UK who seeks to establish operations inside the UK, perform the project inside the UK and present a credible plan for achieving this within the programme duration.

Further information on ARIA's proposal review process can be found here.

Application Process

The selection process will be supported by a set of internal and external evaluators. It will consist of (a) the review of the proposal materials, and (b) a technical interview with a shortlisted set of candidates, which will take place in November. We expect to inform candidates of our final selection in December 2025.

Applications for Phase 2 open	25 June 2025
Phase 2 submission deadline	01 October 2025 (13:00 BST)

Review of Phase 2 proposals **27 October 2025**

If you are shortlisted following full proposal review, you will be invited to meet with the Programme Directors to discuss any critical questions/concerns prior to final selection — this discussion can happen virtually.

Successful/Unsuccessful Phase 2 applicants notified **17 November 2025**

At this stage you will be notified if you have or have not been selected for an award subject to due diligence and negotiation. If you have been selected for an award (subject to negotiations) we expect a 1 hour initial call to take place between ARIA's PD and your lead researcher within 10 working days of being notified.

We expect contract/grant signature to be no later than 4 weeks from successful/ unsuccessful notifications. During this period the following activity will take place:

- Due diligence will be carried out
- The PD and the applicant will discuss, negotiate and agree the project activities, milestones and budget details
- Agreement to the set Terms and Conditions of the Grant/Contract. You can find a copy of our funding agreements here.

How To Submit Your Application

Before submitting an application we strongly encourage you to read this call in full, as well as the general ARIA funding FAQs.

If you have any questions relating to the call, please submit your question to clarifications@aria.org.uk.

Clarification questions should be submitted no later than 4 days prior to the relevant deadline date. Clarification questions received after this date will not be reviewed. Any questions or responses containing information relevant to all applicants will be provided to everyone that has started a submission within the application portal. We'll also periodically publish questions and answers on our website, to keep up to date click here.

Please read the portal instructions below and create your account before the application deadline. In case of any technical issues with the portal please contact clarifications@aria.org.uk.

Application Portal instructions

APPLY HERE