Advanced Research + Invention Agency ARIA

Safeguarded AI: TA2 Phase 1 – Machine Learning Organisation

Call for proposals

Date: 02 April 2025

Advanced Research + Invention Agency ARIA

Advanced Research + Invention Agency ARIA

SUMMARY OF CALL FOR PROPOSALS

What Is ARIA? ARIA is an R&D funding agency created to unlock technological breakthroughs that benefit everyone. Created by an Act of Parliament, and sponsored by the Department for Science, Innovation, and Technology, we fund teams of scientists and engineers to pursue research at the edge of what is scientifically and technologically possible.

The Safeguarded AI Programme. Backed by £59 million, the Safeguarded AI programme aims to combine scientific world-models and mathematical proofs to develop quantitative safety guarantees for AI. We want to show how we might construct "gatekeeper" AI capabilities specifically designed to identify and mitigate the safety risks of other AI agents. By demonstrating 'proof of concept' we intend to establish the viability of a new, alternative pathway for research and development toward safe and transformative AI.

About This Solicitation. 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. Because of the global significance of developing these capabilities (if successful), and potential externalities in case of insufficient risk management and/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. As such, TA2 will be funded in two Phases. In Phase 1 (this solicitation), we will award grants to up to 5 teams at roughly 1-2 FTE to spend 3.5 months developing a full proposal for Phase 2, including a technical roadmap & an organisational plan for the hosting entity. In Phase 2 – applications for which will open in June 2025 – we plan to award a single £18m grant to one entity to host the entire TA2 research agenda.

Ideal Applicants. We welcome applications from new founding teams or existing entities looking to create a new affiliated non-profit. Applications will be evaluated based on their credible ability to advance the TA2 R&D agenda, and on rigorous plans for implementing top-tier governance and security practices to ensure that Safeguarded AI capabilities will be developed and used for the benefit of humanity at large.

Application for Phase 1 closes	30 April 2025

Application Phase 2 opens	25 June 2025
Application for Phase 2 closes	01 October 2025
Total funding Phase 1 (total)	£1m (to up to 5 team)
Total funding Phase 2 (total)	£18m (single award)

SECTION 1: Programme Thesis and Overview

Today's AI is brilliant in many ways, but it is also unreliable. This lack of reliability sharply limits AI's usefulness, especially in safety-critical domains. Our Safeguarded AI programme [1] is a £59m-backed R&D effort to construct a general-purpose AI workflow for producing domain-specific AI agents or decision-support tools for managing cyber-physical systems with quantitative guarantees, improving upon both performance and robustness compared to existing operations.

To do this, we will employ a system that includes both state of the art, "frontier AI", as well as human expertise to construct a "gatekeeper": AI capabilities specifically designed to identify and mitigate the safety risks of other AI agents. This gatekeeper consists of a formal world-model and safety specifications about the application domain, and several ML components responsible for proposing effective task policies and generating verifiable safety guarantees, among others. The resulting Safeguarded AI system will unlock the raw potential of state of the art machine learning models in a wide array of business-critical or safety-critical application domains, while also reducing the risks of frontier AI by providing high-assurance safety guarantees.

The programme will develop the toolkit for building such Safeguarded AI workflows, and demonstrate it in a range of applications in a critical cyber-physical operating context – such as energy, transport, telecommunication, healthcare, supply chains, R&D planning and more – where the value attainable by full deployment is estimated to be billions of pounds. This would, first, act as a proof of concept, demonstrating that it's possible to realise the benefits of AI in safety critical applications through quantitative safety guarantees; and second, catalyse further R&D to replicate and scale the results in diverse application areas across the world.

The Safeguarded AI programme is divided into three main Technical Areas (TAs). TA1 ("Scaffolding") builds the open-source tooling that domain experts can use to author and refine formal world-models and safety specifications about their application domain; TA2 ("Machine Learning") develops the ML elements which harness frontier AI techniques into a general-purpose Safeguarded AI workflow; and TA3 ("Applications") prototypes domain-specific applications of the Safeguarded AI workflow.

For more details, see our programme thesis Safeguarded AI: constructing guaranteed safety [1] as well as a summary of the programme in Appendix A.

SECTION 2: What And How Are We Funding TA2?

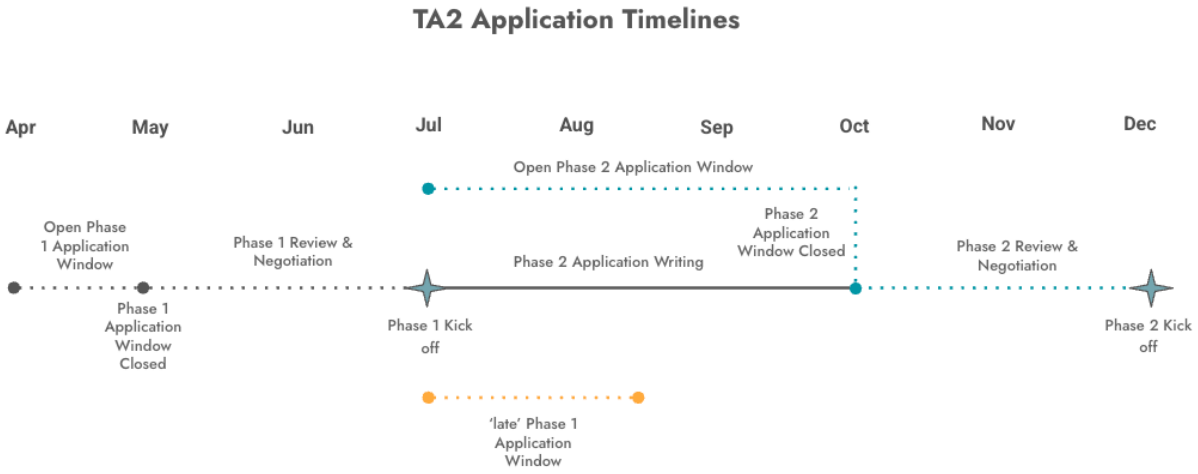
Technical Area 2 (TA2) will develop the machine learning (ML) elements which harness frontier AI techniques into a general-purpose Safeguarded AI workflow. We discuss the overarching objectives of TA2 in Section 3.

We plan to fund TA2 through a £18m award to a single, new or existing, non-profit entity to deliver the TA2 research agenda. (There is no requirement that the entity will work on the TA2 agenda at the exclusion of other activities.) If successful, the TA2 research agenda will develop ML capabilities of a safety-crital and potential dual-use nature. Therefore, the organisation delivering the TA2 research agenda will need to push the frontier on organisational governance and security standards in order to ensure, across time, the safe development and deployment of these technologies. This will constitute, next to the applicant's credible ability to successfully deliver on the technical objectives, one of the central selection criteria on the basis of which we will award the TA2 funding. We discuss this in more detail in section 3B.

Ahead of the £18m award (here 'Phase 2'), we will run a Phase 1 (this solicitation) where we award grants to up to 5 teams at roughly 1-2 FTE. These teams will spend 3.5 months developing a full proposal for Phase 2, including a technical roadmap & organisational plan. We chose this two-phased structure to enable more teams to develop a thoughtful, in-depth Phase 2 proposal and to help derisk the large Phase 2 award. That said, **applications to Phase 2 will remain open to everyone**, including teams who have not received a Phase 1 funding.

The full Phase 2 proposal will require detailed discussion of the technical R&D plan, as well as of a set of key questions about organisational design and governance. The funding call

for Phase 2 will be launched later, in June 2025. (In Appendix B, we discuss the prospective Phase 2 structure, application process and evaluation criteria). Phase 1 applicants should familiarise themselves with the structure, goals and application process of Phase 2, since this will inform their plans in Phase 1. Given the magnitude of the full TA2 award, we will have a high bar for awarding it, and, indeed, we may choose to fund zero teams, if we are not able to reach sufficient levels of confidence in any of them.



For those applicants that do not meet the Phase 1 application deadline, to make TA2 funding accessible to as many strong applicant teams as possible, we will accept (shortened) Phase 1 proposals throughout the first part of Phase 1 (until 17th Aug 2025). 'Late' Phase 1 applications will **not** be eligible for Phase 1 funding, if successful, they will be able to meet with the Safeguarded AI Programme team, including the Scientific Director to discuss their thinking. You can find more information on the process for 'late' phase 1 proposals on page 18.

SECTION 3: TA2 Objectives

A. Technical Research Agenda

Technical Area 2 seeks to challenge the claim that "even if it were possible to specify real-world safety, it wouldn't be economically competitive to train an AI system that produces solutions which *provably* satisfy such a spec". To do so, TA2 aims to demonstrate that we can (1) use human-level frontier AI to act as assistants in helping domain experts build best-in-class mathematical models of real-world complex dynamics in relevant applications

domains (Figure 1, yellow triangles), and (2) leverage securely boxed AI (Figure 1, red triangle) to train autonomous control systems that can be verified with reference to those models, resulting in improvements to both performance and robustness compared to non-certified baselines (e.g. LLM agent loops).

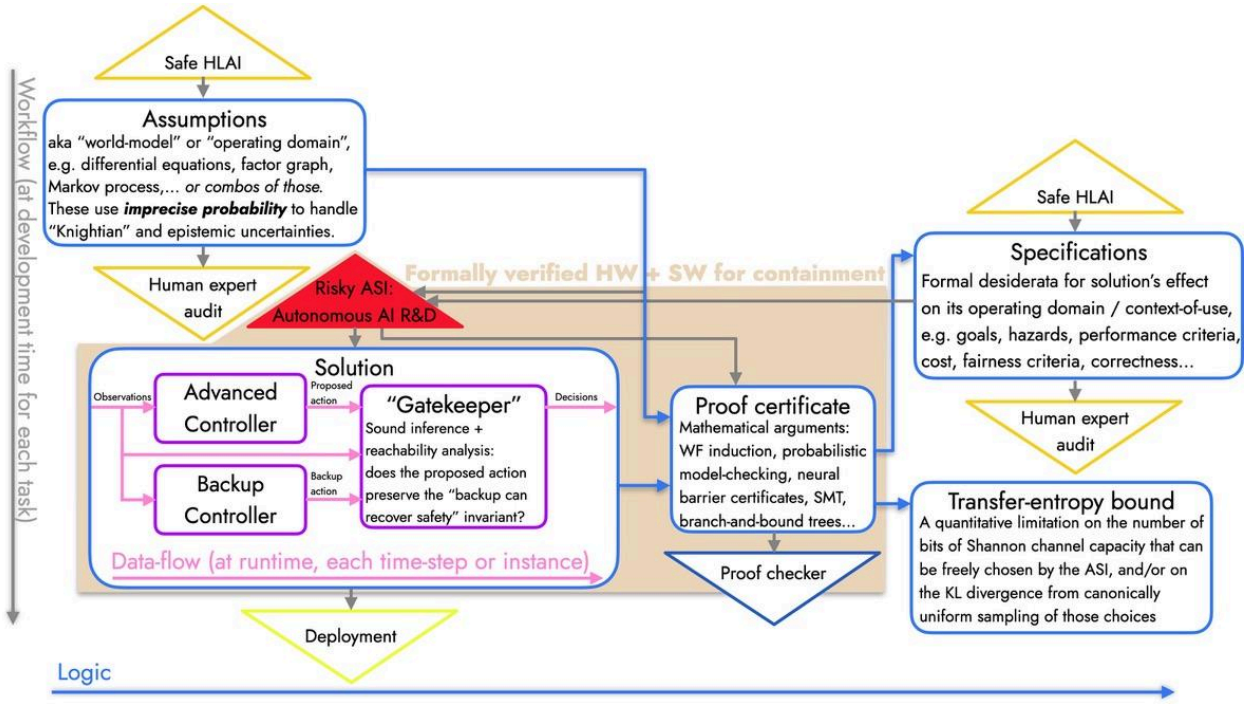


***Figure 1**: Conceptual illustration of a Safeguarded AI workflow. The yellow triangles depict (sub-)human-level AI systems acting as assistants to human experts in authoring and refining the world-models and safety specifications. The red triangle represents a highly-advanced AI system which is unsafe to deploy directly but can, if securely boxed, be leveraged to synthesise task-specific solutions and proof certificates attesting the correctness and uniqueness of the solution relative to the world-model and safety specifications.*

While the "gatekeeper" workflow will primarily build on mainstream pre-trained frontier AI, it involves forking a frontier model and fine-tuning or "post-training" it in a few different ways in parallel, to assemble a workflow which can produce verifiably safeguarded AI solutions for specific tasks. The corresponding four core technical objectives of this R&D effort can be summarised as follows. (Note that the following R&D objectives are interdependent, and should not be tackled in isolation nor in sequence.)

+ **Objective TA2(a): "World-modelling ML."** The objective is to represent human knowledge in a formalised way that admits explicit reasoning, including accounting

for various forms of uncertainty. In this way, we want to iteratively construct a mathematical model of the task-relevant aspects of the real world, and build on this to define quantitative specifications of safety criteria (as well as of task success). Said world-model should be human auditable, cohere with our current leading scientific understanding, and be able to make predictions about the future effects of any given, relevant intervention. Depending on the chosen approach, objective TA2(b) may be needed to make sure that the world-model is self-consistent.

+ **Objective TA2(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 about and derive correct conclusions from the world-model. Exact reasoning may be intractable but can be approximated efficiently with methods such as amortised inference with neural networks, or by using neural networks as powerful heuristics to guide NP-complete algorithms in order to make them practical. Either of these methods provide ways to leverage Transformers and similar architectures to automate reasoning with reliable coherence, unlike purely end-to-end approaches which do not include explicit coherence constraints.

+ **Objective TA2(c): "Safety verification ML."** A central use case of the world-model of TA2(a,b) is to verify that a given action or plan is safe according to the given safety specification. Possible avenues include (but are not limited to) proof certificates which can be obtained by adapting in-context learning to drive a proof search and which are later checked by a proof-checker (that is itself formally verified); or the use of amortised inference whose guarantees of correctly estimating desired conditional probabilities are asymptotic in computational resources. Because of uncertainty in the world-model, a bound on the probability of violating that specification may be estimated, with the goal of rejecting any action or plan for which that probability is above a given threshold.

+ **Objective TA2(d): "Policy training."** An agent policy should be trained to achieve task performance together with finite-horizon safety guarantees, taking advantage of the capabilities developed in TA2(a,b,c). In particular, there should be a "backup" policy to switch to when safety verification fails, with stronger guarantees that the backup policy will satisfy the safety specification in a wider variety of situations (trading off task performance). Here we are primarily interested in the case where the harm that might occur would be caused by the AI system itself.

As part of this R&D roadmap, collaborations with Creators from other TAs will be essential. Some notable programmatic intersections include:

+ **TA1.2 ("Backend")** will build a computational implementation of the mathematical modelling language developed in TA1.1, as well as – in collaboration with and based on the requirement specs provided by TA2 – an interface for AI-driven machine learning training loops for verifying probabilistic claims about policies with respect to these world-models, and producing counterexamples or informative error messages for failed verifications.
+ **TA1.3 ("Human-Computer Interfaces")** will develop the interfaces that enable human domain experts to interact with the AI-powered "assistants" (developed in TA2) in writing world-models and safety specifications, as well as interfaces for run-time monitoring.
+ **TA1.4 ("Sociotechnical Interfaces")** will develop processes for diverse groups of stakeholders to make collective deliberations about acceptable risks and safety specifications, suggest quantitative bargaining solutions to facilitate multi-objective certifiable ML; and provide red-teaming of TA2's organisational processes for making go/no-go decisions about any new deployment, release, or publication.
+ **TA3 ("Applications")** builds domain-specific models and specifications by using and testing tooling developed by TA1 & TA2, and develops benchmark metrics for performance in those specific application areas, against which TA2 technical success can be evaluated.

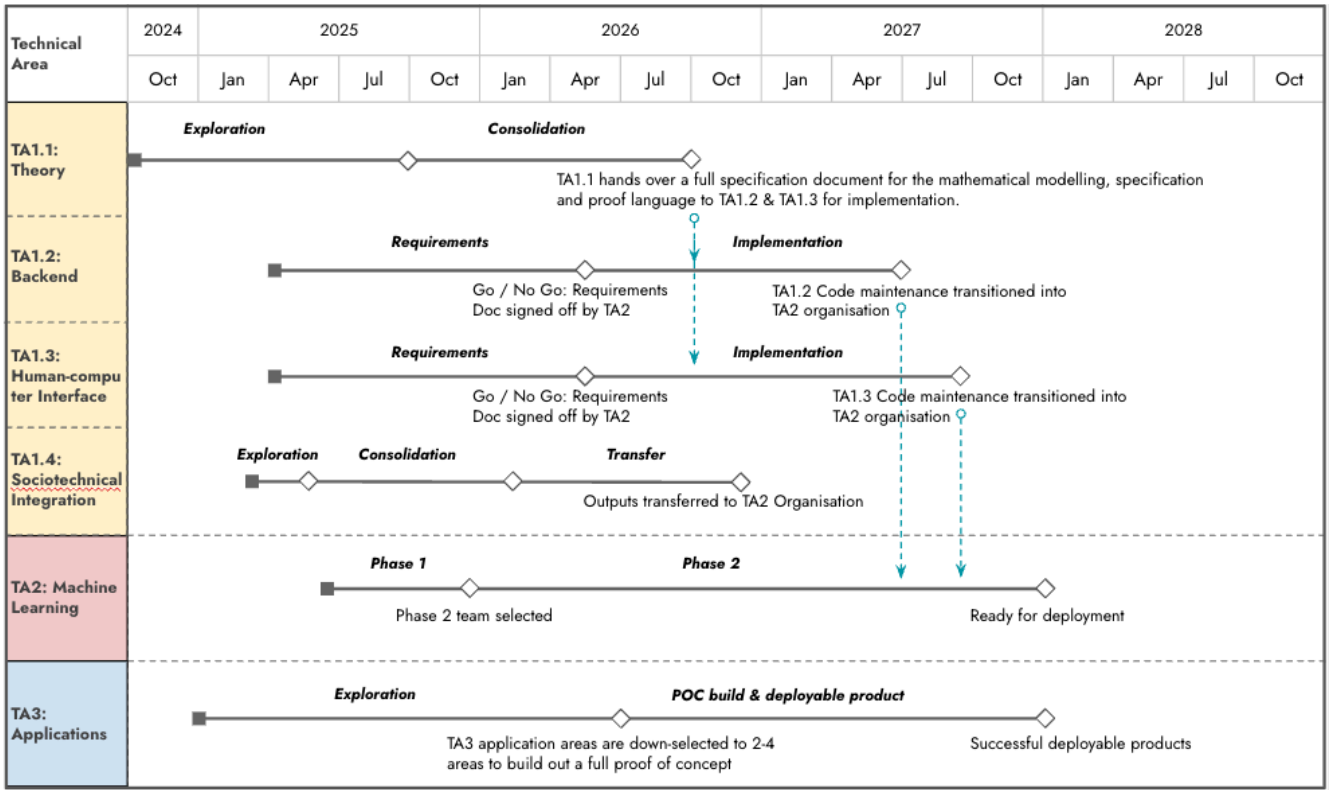
More detail about the technical scope of TA2 can be found in the Programme Thesis, pages 9-11. The Creators for TA1.2, 1.3 and 1.4 are still in the process of selection/negotiation. You can find out more about the call for proposals related to TA1.2 and 1.3 here. You can read more about the TA3 Creators here.

While the technical objectives of the TA2 effort are set out above, we don't yet know what approaches will prove most effective in achieving them. Figuring this out is the point of the TA2 research agenda. Thus, as part of the Phase 2 application, we will ask applicants to address the following questions as part of their technical proposal:

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 Structure, Governance and Security

As mentioned, the organisation delivering this work will have to push the frontier of robust governance and security mechanisms in order to ensure, across time, the safe development and deployment of these technologies. This is because, if successful, the ML capabilities developed as part of TA2 would come with substantial misuse risks if applied to harmful or nefarious use cases. Thus, the outputs must be carefully governed to ensure legitimate and net-positive impact. Among other things, the TA2 entity needs a robust process for reviewing and evaluating decisions to release e.g. publications, API access, commercial licensing, weights, etc., to TA3 entities and others. Additionally, it must adopt first-of-class (cyber) security, which can credibly prevent TA2 capabilities from proliferating unintentionally and irreversibly.

In line with the mission to ensure that AI systems are developed and deployed in service of humanity at large, we also envision that the entity delivering the TA2 research agenda will engage in multilateral technical cooperation, both nationally and internationally. The goal is to ensure interoperability and the sharing of safety-critical information to enable global deployment of Safeguarded AI workflows and to avoid undue incentives to race ahead at

the expense of safety. We hope, through this programme, to nucleate opportunities to develop such an international collaboration (see also page 5 of the Programme Thesis).

We don't have the answers to what the most robust governance structure looks like. But we have a set of questions which we believe any serious proposal will have thoughtful answers to. It will be up to TA2 applicants to deliberate them and propose their own best answers to, as part of their Phase 2 application.

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 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?

***

You can learn more about the vision for the TA2 by listening to this conversation between Programme Director David 'davidad' Dalrymple, Scientific Director Yoshua Bengion, Technical Specialist Nora Ammann and external moderator Adam Marblestone.

SECTION 4: About TA2 Phase 1

Project Deliverables and Management

Phase 1 will last for 3.5 months. As discussed earlier, the purpose of Phase 1 is to support teams in developing a full proposal for Phase 2. In other worlds, the deliverables of Phase 1 are functionally equivalent to the application requirements for the Phase 2 application. The Phase 2 application asks for, among others, comprehensive answers to the technical and governance questions in Section 3A and Section 3B, as well as evidence of 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). You can learn more about Phase 2 application requirements in Appendix B.

Throughout the Phase 1 project period, teams will have the option to meet at least twice (with each meeting approx. 90 minutes) with David 'davidad' Dalrymple (Programme Director) and Yoshua Bengio (Scientific Director) to discuss their thinking. Beyond that, there will be light touch reporting at the end of Phase 1 on progress and cost information.

Approach to Intellectual Property

TA2 Phase 1 will not produce outputs which would require notable intellectual property measures. Our default IP policies apply.

Note that, in service of identifying the best governance structure for the entity delivering the TA2 research agenda, we will retain the ability to share the concept-level elements of answers to the questions about governance structures with other applicants, except where this information pertains to the comparative advantage of a specific group, such as considerations with respect to recruitment/team composition. If you have concerns with this, please explain those concerns in your response to the administrative questions. We will then decide, on a case by case basis, whether deviating from our default policy is warranted.

Applicants are welcome to publish the answers to the questions outlined in Section 3B, if they choose to do so.

SECTION 5: Applying to TA2 Phase 1: eligibility & application process

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.

A non-exhaustive list **types of applicants** 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[1] 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[1] non-profit entity to become a pioneering supplier of guaranteed-safe AI capabilities; or

[1] 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[1] non-profit entity[2], where TA2 R&D can be pursued under conditions of first-of-class information- and cyber-security

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 ARIAs standard eligibility criteria here, the following types of entities are **not** eligible for funding to deliver TA2[3]:

+ For-profit companies
+ Universities directly hosting TA2

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.

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

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

Phase 1 Application process

Proposals for TA2 Phase 1 should not exceed 4 pages and address:

- **Project information**, including:
  - Your articulation of the TA2 mission and technical objective
  - Information about the technical team, and/or about how you plan to attract key technical talent, and any other evidence pertaining to your ability to successfully pursue the TA2 research agenda in Phase 2
  - Broad-stroke answers to how you are currently thinking about the governance questions listed in section 3B, and information about how you concretely plan to go about developing full answers to these questions over the course of Phase 1
  - A discussion of what the major obstacles are you foresee, that could prevent successful execution of the Phase 1 deliverables
- **Information about the team**, including
  - An overview of all team members involved in Phase 1, their respective areas expertise; evidence of your ability to lead a successful organisation, including evidence of your ability to attract additional funds; evidence of the team's commitment to the public interest, and discussion of why you are motivated to solve this problem
- **Administrative questions** to help ensure we are responsibly funding R&D. Find an overview of all questions here . These will be answered via the application portal and do not count towards the 4 page limit. When completing the cost information, please refer to ARIA's standard eligible cost guidance. However, note that certain costs typically classified as ineligible under the standard guidelines may be eligible for this Phase 1. These include costs associated with legal, accountancy, regulatory, and professional advice related to incorporation.

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

- Page count: 4 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 Section 7 for guidance on how to submit your application.

'Late' Phase 1 Proposals

For those applicants that do not meet the Phase 1 application deadline, to make TA2 funding as accessible as possible to as many strong applicant teams, we will accept (shortened) Phase 1 proposals until 17th Aug 2025. These proposals will **not** be eligible for Phase 1 funding and will be reviewed against the same Phase 1 evaluation criteria (see below). If successful, these teams will be invited to meet with the Safeguarded AI Programme team, including the Scientific Director to discuss their thinking.

To apply, follow the same instructions as for Phase 1 (see above), but limit the submission to 3 pages instead of 4 pages. Email clarifications@aria.org.uk, and we'll provide you with a link via which you will be able to submit your application. If you are not selected for meeting with the Programme Director as part of this process, you will still be eligible to submit a proposal in response to the Phase 2 call document.

Phase 1 Evaluation Process & Criteria

To build a programme at ARIA, each Programme Director directs the review, selection, and funding of a portfolio of projects, whose collective aim is to unlock breakthroughs that impact society. As such, we empower Programme Directors to make robust selection decisions in service of their programme's objectives ensuring they justify their selection recommendations internally for consistency of process and fairness prior to final selection.

We take a criteria-led approach to evaluation, as such all proposals are evaluated against the criteria outlined below. We expect proposals to spike against our criteria and have different strengths and weaknesses. Expert technical reviewers (both internal and external to ARIA) evaluate proposals to provide independent views, stimulate discussion and inform decision-making. Final selection will be based on an assessment of the programme portfolio as a whole, its alignment with the overall programme goals and objectives and the diversity of applicants across the programme.

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

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 1 proposals we'll consider the following criteria:

1) **Credible ability to deliver the TA2 technical R&D agenda**, as evidenced by the applicant's understanding of the TA2 mission and technical objectives, a credible technical team for delivering the TA2 research agenda, and/or a compelling plan to hire such a team.

2) **Credible ability to embed the technical work in a compelling organisational structure, including the legal & economic model, governance and security,** as evidenced by the applicant's plan for Phase 1, their experience leading successful organisations, demonstrated commitment to the public interest by the team, and evidence of their ability to attract additional funds.

3) **Fit and intrinsic motivation in alignment with the TA2 objectives**, as evidenced by 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. Strong cases for benefit to the UK include proposals that:
    1. are led by an applicant within the UK who will perform the majority (>50% of project costs spent in the UK) of the project within the UK
    2. are led by an applicant outside the UK who seeks to establish operations inside the UK, perform a majority (>50% of project costs spent in the UK) of the project inside the UK and present a credible plan for achieving this within the programme duration.

SECTION 6: Timelines

This call for project funding will be open for applications as follows. Note that we may update timelines based on the volume of responses we receive.

Advanced Research + Invention Agency ARIA

Phase 1 Applications

Applications for Phase 1 open	02 April 2025
Phase 1 submission deadline	30 April 2025 (13:00 BST)
Review of Phase 1 proposals	19 May 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 1 applicants notified	30 May 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 ARIAs 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.

SECTION 7: 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

APPENDIX A – Short summary of full Safeguarded AI programme

While this solicitation focuses on TA2, the full programme can be found described in more detail in the Safeguarded AI programme thesis [1] (pages 7–13). Below, we provide a brief summary of each of the Technical Areas the programme is divided into.

+ **TA1 Scaffolding**
    - **TA1.1 Theory:** to research and construct computationally practical mathematical representations and formal semantics for world-models, specifications, proofs, neural systems, and "version control" (incremental updates or patches) thereof.
    - **TA1.2 Backend**: to develop a professional-grade computational implementation of the Theory, yielding a distributed version control system for all the above, as well as computationally efficient (possibly GPU-based) type-checking and proof-checking APIs.
    - **TA1.3 Human-computer interface**: to create a very efficient user experience for eliciting and composing components of world-models, goals, constraints, interactively collaborating with AI-powered "assistants" (from TA2), and run-time monitoring and interventions.
    - **TA1.4 Sociotechnical integration**: to leverage social choice and political theory to develop collective deliberation and decision-making processes about AI specifications and about AI deployment/release decisions.
+ **TA2 Machine Learning (this solicitation)**
    - **TA2(a) World-modelling ML**: to develop fine-tuned AI systems to represent human knowledge in a formalised way that admits explicit reasoning, including accounting for various forms of uncertainty.

- **TA2(b) Coherent-reasoning ML**: to develop efficient ways to reason about the world-model thereby allowing us to practically leverage the world-model to guarantee safety in a complex environment.
- **TA2(c): Safety-verification ML**: to develop fine-tuned AI systems to verify that a given action or plan is safe according to the given safety specification.
- **TA2(d): Policy training:** to fine-tune AI systems to learn an agent policy that achieves finite-horizon safety guarantees, taking advantage of the capabilities developed in objectives TA2(a,b,c).

+ **TA3 Applications**: to elicit functional and nonfunctional requirements, test problems and evaluation suits in a particular application domains, and to ultimately demonstrate deployable solutions, leveraging TA1 and TA2 tools, to solve specific, economically valuable challenges in cyber-physical systems

APPENDIX B – About TA2 Phase 2

Phase 2 Structure & Management

Applications for TA2 Phase 2 will be launched in June 2025, and remain open until the end of September 2025, when Phase 1 project phase ends. TA2 Phase 2 will kick off around EOY 2025 (calendar year) and last until the end of the programme, EOY 2027.

As part of the application, we will agree on technical milestones against which the successful TA2 will report throughout the duration of the programme. Throughout this period, we will also facilitate interactions with Creators across the rest of the programme, among others by means of quarterly Creator events. Finally, there will be ARIA's standard project management requirements, including light touch quarterly reporting on progress and cost information.

Phase 2 Application & Evaluation Criteria

More information about how to apply to Phase 2 will be released in June 2025. While we retain the option to make updates to the final application criteria and guidance for TA2 Phase 2, the present discussion should be considered highly indicative of what we will be looking for in applications to TA2 Phase 2.

The Phase 2 application is expected to be 30-50 pages long, including a technical roadmap & organisational plan. We plan to select the Phase 2 team based on 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. Strong cases for benefit to the UK include proposals that:
    a. are led by an applicant within the UK who will perform the majority (>50% of project costs spent in the UK) of the project within the UK
    b. are led by an applicant outside the UK who seeks to establish operations inside the UK, perform a majority (>50% of project costs spent in the UK) of the project inside the UK and present a credible plan for achieving this within the programme duration.

The selection process for Phase 2 applicants will differ from Phase 1. More information will be provided in the Phase 2 funding call documentation.

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.[4] 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 (see section 3B).

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

