

Safeguarded AI: TA3 Applications Call for proposals

Date: 7 August 2024

V1.0

| | |
|--|-----------|
| SUMMARY OF CALL FOR PROPOSALS..... | 3 |
| SECTION 1: Programme thesis and overview..... | 4 |
| SECTION 2: TA3 objectives..... | 5 |
| SECTION 3: Technical metrics..... | 6 |
| SECTION 4: What are we looking for/what are we not looking for..... | 7 |
| SECTION 5: Programme duration and project management..... | 11 |
| Programme structure..... | 11 |
| Programme duration..... | 11 |
| Programme management and project milestones..... | 12 |
| Approach to intellectual property..... | 12 |
| Community events..... | 12 |
| SECTION 6: Application & Eligibility..... | 13 |
| Eligibility..... | 13 |
| Application process..... | 13 |
| SECTION 7: Timelines..... | 14 |
| SECTION 8: Evaluation criteria..... | 14 |
| Proposal evaluation principles..... | 14 |
| Proposal evaluation process and criteria..... | 15 |
| SECTION 9: How to apply..... | 17 |
| SECTION 10: References..... | 18 |
| APPENDIX..... | 18 |
| Short summary of full Safeguarded AI programme..... | 18 |

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 understand how we might construct a “gatekeeper”: a targeted AI whose job it is to understand and reduce 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.

This Solicitation. Safeguarded AI’s success will depend on showing that the gatekeeper actually works in a safety-critical domain – the focus of Technical Area 3 (TA3). The ~10 teams selected through this call for proposals will work with other programme teams, global AI experts, academics, and entrepreneurs, in setting the groundwork for being the first to deploy Safeguarded AI in one or more areas.

Ideal Applicants. We’re looking for both product developers (Track 1) and end-users (Track 2) in safety-critical domains where AI is currently too unreliable to be deployed at scale. We’re open to all forms of creators including individuals, startups, SMEs, large companies, and nonprofit R&D organisations.

Logistics Summary. TA3 Applications will include 2 phases with this solicitation focusing on the first phase (£5.4m). A second phase (£8.4m) will launch in approximately 1 year:

| | |
|--------------------------------|------------------------|
| Application deadline | 02 October 2024 |
| Kickoff | November/December 2024 |
| Total funding available | £5.4m |
| Total number of teams | 10 teams |

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. The [Safeguarded AI programme \[1\]](#) 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.

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 system which monitors and ensures safe behaviour of other AI agents. This gatekeeper will consist 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 applications domains where reliability is key.

The programme will develop the toolkit for building such a Safeguarded AI workflow, and demonstrate it in a range of applications domains such as energy, transport, telecommunication, healthcare, and more. This would, first, act as a proof of concept, proving 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 across application areas in the world.

The Safeguarded AI programme is divided into three main Technical Areas (TAs). This call for proposals is in Technical Area 3 (TA3), where we will fund teams to develop and prototype domain-specific applications of the Safeguarded AI workflow. Successful proposals will get the opportunity to access, inform and prototype the use of novel, "edge-of-the-possible" AI solutions in their specific domain of application. Meanwhile, TA1 is focusing on building the tools that domain experts can use to develop and refine formal world models and specifications about their domains of interest; and TA2 will develop the ML elements which harness frontier AI techniques into a general-purpose 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: TA3 objectives

The programme's Technical Area 3 (TA3) aims to challenge the claim that “guaranteed safe AI will provide no additional economic value beyond mainstream AI.”

We do this by developing compelling domain-specific AI decision support tools or safeguarded autonomous AI control systems. We seek to demonstrate that a “gatekeeper” workflow can be used to solve economically valuable challenges, by using AI to create and maintain tools and systems that can improve both performance and robustness, compared to existing operations. Examples of relevant application domains include optimisation processes in transport, communication, energy, supply chains, R&D, clinical trials, and others (see section 4 for a more detailed list). The stretch goal within the programme (i.e. by 2028) is for an operator in a relevant application domain to be using a Safeguarded AI solution in actual practical operational use. We want to do this in a critical cyber-physical operating context where the value attainable by full deployment is estimated to be billions of pounds. We give examples of potential application domains in section 4.

TA3 is seeking two types of Creators (i.e. individuals and teams that ARIA will fund and support). Throughout the rest of this document, we will refer to them as Track 1 and Track 2.

- + **Track 1 Creators** are product developers, i.e. individuals, entrepreneurs and existing organisations (including startups, SMEs, large companies, and nonprofit R&D organisations) interested in leveraging near-future Safeguarded AI technology to develop products tailored to the needs of customers/end users in a variety of specific application domains (see section 4 for more details). As part of track 1, we are also open to funding operators who are looking to build a small team to do this R&D work in-house.
- + **Track 2 Creators** are customers/end users, i.e. organisations who already want to participate in the co-creation of solutions—not by building an in-house solution R&D team, but by actively participating in the customer discovery processes run by Track 1 Creators whom we may fund in the same sector/domain.

TA3 will include two phases. In Phase 1 of TA3 (this solicitation), we will cast a wide net, funding a number of “pilot” efforts to deeply understand customer needs, elicit requirements, begin to source datasets and/or simulators, design evaluation suites to validate the performance of predictive models and autonomous and semi-autonomous controllers, etc. in the chosen application areas.

This phase will involve drafting models and specifications manually, or identifying existing ones (as opposed to using AI tooling, as will be done in Phase 2). These models should be developed as a 'curriculum', progressing from simpler to more complex "example problem" (within the same application domain). Work in this phase will include interacting with Creators from TA1 and TA2 to communicate aspects of the application domain that might importantly inform their work. These specifications will serve as benchmark evaluations for the entire research programme and ARIA will make sure that the AI systems developed in the context of the programme will be specifically tailored to solve these particular problems. Later, in Phase 2, TA3 Creators will be able to use the tooling prototypes made available by Creators in TA1 and TA2, in order to accelerate and scale up the creation of models and specifications.

Creators will have the opportunity to access, inform and prototype the use of the cutting-edge Safeguarded AI solutions emerging from this programme. At the same time, they will be able to deepen their customer understanding, build out internal AI know-how, and connect to ARIA's wider community of creators and innovators.

By the end of Phase 1, we are looking for teams to have established:

- + A deep understanding of the domain requirements and a credible deployment path (see section 3 for more detail), and
- + The suitability of their application domain & customer problems as a demonstration of the capabilities developed in TA1 and TA2, and for the programme goals overall

The reason for launching TA3 early into the overall life-cycle of the programme is that we want to prioritise understanding the needs of application areas sooner rather than later, thereby helping to de-risk the practical applicability of the tools developed in other TAs. This will allow TA3 Creators to elicit user-inspired assessments and requirements, and share them with Creators in other TAs to inform their work.

SECTION 3: Technical metrics

As discussed in the previous section, the objective at the end of TA3 (Phase 2) is to use work in other Technical Areas of the programme to develop compelling domain-specific AI decision support tools or autonomous control systems.

In Phase 1 (pilot projects), we will assess the following indicators of progress:

For Track 1 Creators:

- + Ability to produce (a curriculum of) “simplified test problems” (from simple to increasingly more complex) to support the work in TA1.1
- + Deep understanding of customer needs and other safety specifications
- + Access to data needed to build models and specifications
- + Development of suitable evaluation suites to validate & improve the performance of predictive models and autonomous and semi-autonomous controllers
- + Success in establishing a credible deployment path
- + Fit with emerging capabilities in TA2 and TA1

For Track 2 Creators:

- + Ability to work with Track 1 Creators to provide information, data and feedback to help build the curriculum of test problems, predictive models and specifications
- + Ability to provide information needed to establish the baseline of current operations, and inform the design of evaluations suites to validate & improve predictive models and autonomous and semi-autonomous controllers
- + Ability to help establish credible deployment paths

Depending on the proposals submitted, we expect some Track 1 and Track 2 Creators will collaborate and share some level of information with each other. During contract negotiations, the Programme Director will agree with each Creator what information they can comfortably share with other Creators. This information will be captured in the scope of work as “Specification information,” with a requirement that information marked as such will not be shared beyond the programme participants, without the express permission of the source.

SECTION 4: What are we looking for/what are we not looking for

TA3 solicits potential *product developers* (which could be individuals, entrepreneurs, existing entities, small, medium or large enterprises, nonprofit research institutions, etc.) as well as *customer* organisations (that is, operators in a relevant application domain) interested in using our gatekeeper AI workflow to build and apply safeguarded products for specific tasks in a specific sector. Phase 1 will provide the funding to run initial “pilots” for 12-15

months, from which we will choose a smaller number of projects to fund with a larger contribution in Phase 2.

Application areas suitable for an early demonstration (i.e. within our programme duration of 4 years) are likely fit all the following criteria:

- (a) scalability**— an ideal application area affords to specify a ‘curricula’ of problems ranging from more simple to more complex “instances” of the same problem type (e.g. by increasing the size or number of entities being modelled, etc.)
- (b) known in principle**— the primary difficulties involved in this problem area should **not** include yet unsolved scientific problems, indicated by e.g.
 - + a lack of a solid informal scientific consensus understanding of substantial aspects of the predictive model
 - + difficulty of making sufficiently detailed measurements or observations of the phenomenon.

Instead, the difficulties should be more of the type “there’s just a lot of moving parts” or (less preferably) “it’s just really inefficient to compute.”

- (c) predictable in principle**— not swamped by sensitivity to initial conditions
- (d) need for high trust**— because of their safety-critical or mission-critical nature, automation and AI solutions for this application are currently facing serious barriers to adoption due to lack of reliability, which our methods could directly address
- (e) absence of bias, or bias mitigation strategy**— either we have data which is unaffected by systemic bias, or we have no reason to expect existing large language models distilled from internet text to bring in systemic bias, or we have a plan in place to avoid the default outcome in which these biases become encoded and amplified by the model
- (f) large-scale relevance**— the total value attainable by successful deployment of Safeguarded AI in this domain is estimated to be billions of pounds
- (g) existing baseline predictor or controller(s)**— there’s some approximate and/or costly ways that large instances are dealt with in practice today, to which new predictive mathematical models and new decision-making systems could be quantitatively compared

The full set of specific application domains is to be determined in the review of responses to the TA3 solicitation, but some areas we're currently exploring include:

- + energy system optimisation, e.g.
 - real-time power dispatch to balance supply and demand and respond to perturbations (this is especially complex with energy storage and more renewable supply)
 - monitoring and stabilising grid phase, beyond flow matching
 - probabilistic demand forecasting, on various time scales
 - keeping network models up-to-date (e.g. with new rooftop solar installations)
 - long-term planning of transmission network capacity improvements
- + transport network
 - adaptive routing/dispatch and control of mass rapid transit
 - dynamic control of railway signals
 - adaptive routing/dispatch of buses
 - long-term planning of transport network improvements
- + telecommunications network optimisation
 - real-time allocation of beamforming subchannels to optimise transmitter energy consumption
 - management of upstream/backhaul capacity
 - long-term network expansion planning
- + supply chain and inventory management
 - probabilistic demand forecasting
 - distribution requirements planning
 - last-mile delivery routing
- + supply chain management for vaccines & drugs (with specifications informed by public-health outcomes rather than by profits)
- + operational monitoring and tuning of bioreactors that produce vaccines & drugs
 - e.g. monoclonal antibodies
- + medical device control systems
- + dynamic optimisation of clinical trials
- + epidemiology
 - especially under various intervention scenarios, for decision support
 - incorporating diverse data sources, including metagenomics
 - including potential models of non-infectious disease states in population
- + complex business process management, dispatch and triage
 - e.g. for healthcare

- + data integration, in contexts where it is usually done by hand to avoid mistakes
 - e.g. for healthcare
- + construction project management and scheduling
- + climate and weather prediction
 - especially under various intervention scenarios
- + aircraft and spacecraft flight dynamics
 - fully autonomous aircraft
 - improving autopilot performance
 - improved efficiency of airspace/traffic control
- + R&D planning
 - roadmapping
 - short-term project management
 - medium-term forecasting
 - long-term R&D portfolio planning
- + control systems for robots in human environments
- + personalised educational curriculum design based on Bayesian assessments
- + industrial (e.g. chemical) process operations
- + water and wastewater system operations
- + risk assessment & risk management

Note that **verified software systems** is an area which is highly suitable for a simplified gatekeeper workflow, in which the world-model is implicit in the specification logic. However, in the context of ARIA's mission to "change the perception of what's possible or valuable," we consider that this application pathway is already perceived to be possible and valuable by the AI community. As such, this programme focuses on building capabilities to construct guaranteed-safe AI systems in *cyber-physical* domains. That being said, if you are an organisation which specialises in verified software, we would love to hear from you outside of this solicitation about the cyber-physical challenges that are just at the edge of the possible for your current techniques. You can contact us by submitting your input [here](#). We plan to open a separate funding call in the near future which will be open to proposals in this domain.

The selection of Creators for the remainder of the technical areas (TA1.2 - 1.4, and TA2) will be subject to separate competitive solicitations due to be released in the coming months. Applications for TA1.1, 1.2, 1.3, 1.4 and 2 should not be submitted in response to this call; instead parties interested in participating in these elements beyond TA3 are encouraged to

register their interest by sending an email to clarifications@aria.org.uk; we'll notify you when the other TA solicitations go live.

SECTION 5: Programme duration and project management

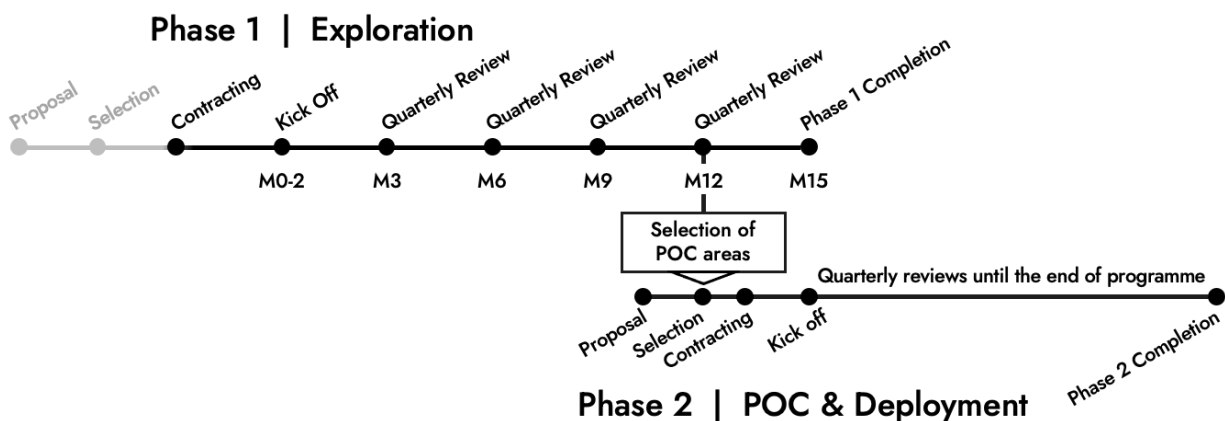
Programme structure

TA3 Phase 1 will be funded with £5.4m (inclusive of VAT where applicable) distributed across approximately 10 teams. These could be part-time efforts depending on the details of the proposal.

Following on from Phase 1, Phase 2 will likely consist of further £8.2 million of funding distributed between 2-4 teams full-time, each pursuing a different application area using the domain-general tools developed in TAs 1 and 2.

Programme duration

We expect to fund TA3 Phase 1 projects through the end of March 2026. Application proposals should not exceed this timeframe. We intend to make decisions about Phase 2 in early 2026, based on the outputs from the initial phase. The second phase will focus on building a proof of concept followed by a deployable product. Phase 1 Creators will be asked to submit proposals for Phase 2 funding, and the solicitation will also be open to new applicants to submit their proposals.



Programme management and project milestones

We are seeking to fund you and your team's research time/efforts towards the creation of a curriculum of test problems, functional and nonfunctional requirements, relevant datasets or simulators, evaluation suites, a stakeholder map & proposal of a plausible path to deployment etc. These outputs will inform the design of tools in other TAs that will eventually be deployed in the domains which we select for Phase 2. Track 2 Creators in particular will instead provide input and feedback on the respective deliverables of relevant Track 1 Creators.

Alongside our standard project management requirements (i.e. light touch quarterly reporting on progress and cost information), we will meet with all Creators on a quarterly basis to discuss the progress, and facilitate interactions with Creators from other TAs as required.

Approach to intellectual property

Work created in TA3 (such as domain-specific models, libraries, techniques, and control systems using TA1 and TA2 software tools) will be treated according to ARIA's usual terms, as the proprietary IP of its Creators. However, we do request a non-exclusive, non-commercial license for the data provided by you, solely for the purpose of conducting pre-competitive research within the context of TA1 and TA2.

Community events

In an effort to foster a collaborative research environment, ARIA will host regular Creator community events across programmes to allow participants to exchange updates, ideas, and feedback on best paths forward. Attendance at these events is encouraged but will not be mandatory.

SECTION 6: Application & Eligibility

Eligibility

We welcome applications from across the R&D ecosystem, including individual entrepreneurs, startups, and established companies.

Our primary focus is on funding those who are based in the UK. For the vast majority of applicants, we therefore require the majority of the project work to be conducted in the UK (i.e. >50% of project costs and personnel time).

However, we can award funding to applicants whose projects will primarily take place outside of the UK, if we believe it can boost the net impact of a programme.

In these instances, you must outline any proposed plans or commitments in the UK that will contribute to the programme within the project's duration (note the maximum project duration is 15 months). If you are selected for an award subject to negotiation, these plans will form part of those negotiations and any resultant contract/grant.

More information on the evaluation criteria we will use to assess benefit to the UK can be found later in the document [here](#).

Application process

The application process for Technical Areas 3 consists of one stage which requires you to submit a detailed proposal (max. 4 pages) including:

- **Project & Technical information** to help us gain a detailed understanding of your proposal.
- **Information about the team** to help us learn more about who will be doing the research, their expertise, and why you/the team are motivated to solve the problem.
- **Administrative questions** to help ensure we are responsibly funding R&D.
Questions relate to budgets, IP, potential COIs etc

You can find more detailed guidance on what to include in a full proposal [here](#). We strongly recommend you read this document as it contains information critical to proposal submission.

For more details on the evaluation criteria we'll use, [click here](#).

SECTION 7: Timelines

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

| | |
|---|------------------------------------|
| Applications open | 07 August 2024 |
| Full proposal submission deadline | 02 October 2024 (12:00 BST) |
| Full proposal review | 6 November 2024 |
| <p>If you are shortlisted following full proposal review, you may be invited to meet with the Programme Directors to discuss any critical questions/concerns prior to final selection — this discussion can happen virtually. This is likely to be the 8th and 11th November.</p> | |
| Successful/Unsuccessful applicants notified | 18 November 2024 |
| <p>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 Programme Director (PD) and your lead researcher within 15 working days of being notified.</p> <p>We expect contract/grant signature to be no later than 8 weeks from successful/unsuccessful notifications. During this period the following activity will take place:</p> <ul style="list-style-type: none"> • 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 8: Evaluation criteria

Proposal evaluation principles

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 ARIAs proposal review process can be found [here](#).

Proposal evaluation process and 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.

Proposals that pass through the initial screening and compliance review will then proceed to full review by the Programme Director and expert technical reviewers.

In conducting a full review of the proposal we'll consider the following criteria:

- 1) **Worth Shooting For** – The proposed project uniquely contributes to the overall portfolio of approaches needed to advance the programme goals and objectives. It has the potential to be transformative and/or address critical challenges within and/or meaningfully contribute to the programme thesis, metrics or measures.
- 2) **Differentiated** – The proposed approach is innovative and differentiated from commercial or emerging technologies being funded or developed elsewhere.

- 3) **Well defined** – The proposed project clearly identifies what R&D will be done to advance the programme thesis, metrics or measures, is feasible and supported by data and/or strong scientific rationale. The composition and planned coordination and management of the team is clearly defined and reasonable. Task descriptions and associated technical elements provided are complete and in a logical sequence with all proposed stage-gates and deliverables clearly defined. The costs and timelines proposed are reasonable/realistic.
- 4) **Responsible** – The proposal identifies major ethical, legal or regulatory risks and that planned mitigation efforts are clearly defined and feasible.
- 5) **Intrinsic motivation** – The individual or team proposed demonstrates deep problem knowledge, have advanced skills in the proposed area and shows intrinsic motivation to work on the project. The proposal brings together disciplines from diverse backgrounds.
- 6) **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.

For all other applicants we will evaluate the proposal based on its potential to boost the net impact of the programme in the UK. This could include:

3. A commitment to providing a direct benefit to the UK economy, scientific innovation, invention, or quality of life, commensurate with the value of the award;
4. The project's inclusion in the programme significantly boosts the probability of success and/or increases the net benefit of specific UK-based programme elements, for example, the project represents a small but essential component of the programme for which there is no reasonable, comparably capable UK alternative.

When considering the benefit to the UK, the proposal will be considered on a portfolio basis and with regard to the next best alternative proposal from a UK organisation/individual.

SECTION 9: How to apply

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 26th September. 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](#)

SECTION 10: References

[1] Dalrymple, D. (2024). *Safeguarded AI: constructing guaranteed safety*. aria.org.uk. Available at:

<https://www.aria.org.uk/wp-content/uploads/2024/01/ARIA-Safeguarded-AI-Programme-Thesis-V1.pdf>.

[2] Dalrymple, D. (2023). *Mathematics and modelling are the keys we need to safely unlock transformative AI*. aria.org.uk. Available at:

<https://www.aria.org.uk/wp-content/uploads/2024/04/ARIA-Mathematics-and-modelling-are-the-keys-we-need-to-safely-unlock-transformative-AI-v01.pdf>.

APPENDIX

Short summary of full Safeguarded AI programme

While this solicitation focuses on TA3, 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 (Phase 1 call for proposals closed 28.05.2024; you can find more information [here](#)):** 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 (Phase 1 call for proposals are to open later this year)**
 - **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 (This Solicitation):** 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