Advanced Research + Invention Agency

ARIA

Safeguarded AI: TA1.1 Theory

Call for proposals

Date: 11 April 2024

v1.0

Table of Contents

Section 1: Programme thesis and overview **3**

Section 2: Programme objectives ***5***

Section 3: TA1.1 Technical metrics **7**

Section 4: What are we looking for/what are we not looking for **11**

Section 5: Project duration and management **12**

Project milestones 12

Approach to intellectual property 13

Programme and project management 14

Community events 14

Section 6: Eligibility and application process **14**

Eligibility 14

Application process 14

Non-UK applicants only 15

Section 7: Timelines **16**

Section 8: Evaluation criteria **17**

Proposal evaluation principles 17

Proposal evaluation process and criteria 17

Section 9: How to apply **19**

Section 10: References **19**

Section 1: Programme thesis and overview

This solicitation is derived from the programme thesis Safeguarded AI: constructing guaranteed safety [1], in the opportunity space Mathematics and modelling are the keys we need to safely unlock transformative AI [2]. This section summarises some of the essential context from the programme thesis.

Most of the potential value of artificial intelligence is the ability to solve specific problems with levels of quality, speed, and cost that are not jointly feasible with teams of humans. On the other hand, the most serious dangers of artificial intelligence come from the potential that AI systems may very effectively or rapidly solve "wrong" problems, i.e. problems which deviate from what was intended by the operator, and/or from what is acceptable by society (these are sometimes known as AI accident risks and AI misuse risks respectively).

The Safeguarded AI programme thesis [1] lays out an approach to ensuring that AI systems solve the "right" problems: problems which were intended — not just by a single human operator, but by a multi-stakeholder collective deliberation process. Ideally, this should be incrementally updated, to ensure that stakeholders can continually adjust the specification, adapting in a sociotechnical way to shifting needs and unforeseen complexities.

The first step in the envisioned workflow would be for groups of humans and human-level AIs to collectively construct formal probabilistic descriptions corresponding to the aspects of reality that are relevant to the task. That includes actions and observations available to a prospective autonomous system, and the ways in which actions and observations interact with the dynamics of the world. Only with a rich vocabulary for properties of uncertain trajectories of world-states can the "right" problems be specified in a fully grounded way. Even probabilistic uncertainty alone lacks sufficient epistemological humility; as suggested by Jeannette Wing [3], we aim to construct system descriptions which define classes of probabilistic models. Correctness means that for all models in the class, the probability of harmful outcomes caused by deployment is bounded by a societal risk criterion.

It is important that the modelling framework used to specify cyber-physical systems be expressive enough to succinctly describe systems from across diverse areas of science and engineering, such as electrical transmission lines, radio beamforming channels, movement of physical goods, epidemic transmission, and even the delay characteristics of interrelated business processes (with semantics related to stochastic timed Petri nets). For the "cyber" aspect, the same modelling framework must be capable of reasoning about computational processes (including neural networks, described declaratively and succinctly, like the Open

Neural Network Exchange format). We are more concerned with description length than computational complexity.

Ultimately, the most ambitious goal for this modelling framework is to house a coherent composite of models for all the processes that are critical for the safety of humanity from catastrophic risks. This composite model would be used to define safety for a global network of "safeguarding" AI systems, which would each ensure that all other AI systems (including each other) are not taking actions that would lead to catastrophic outcomes exceeding the societal risk criteria — as well as actively mitigating other catastrophic risks such as climate change and pandemics.

While ambitious, in prospective futures where some AI systems have such advanced capabilities as to pose a risk that humanity will lose control of the future, it is plausible that they may also have sufficient capabilities to ensure that we do not—as long as humans and AIs can communicate in a fully grounded way about what exactly that problem specification is.

For more context, please read the programme thesis, Safeguarded AI: constructing guaranteed safety [1].

Section 2: Programme objectives

This solicitation focuses on TA1.1 of the Safeguarding AI Programme. TA1.1 seeks R&D Creators, which are individuals and teams that ARIA will fund and support, to research and construct computationally practicable mathematical representations and formal semantics to support world-models, specifications about state-trajectories, neural systems, proofs that neural outputs validate specifications, and "version control" (incremental updates or "patches") thereof.

The aspirational aim of TA1.1 as a whole is to define "syntax" (algebraic construction operators, and version-controllable serialisable data structures), and formal semantics, for language(s) that can be used by teams of humans (and, later, AI systems) to define "world models", probabilistic specifications, neural network controllers, and proof certificates (which present efficiently checkable arguments verifying that a controller composed with a world model satisfies its specification). Primary responsibility for an industrial-strength implementation will rest with an organisation (or organisations) in TA1.2, which will be selected in a solicitation to follow later in the year (More information on the wider Safeguarded AI programme and TAs can be found below).

We anticipate that from the collection of various test problems and solution approaches, a cohesive solution will emerge. Therefore within TA1.1, we hope to produce a single artefact that is a dissertation-length definition of these languages, to be used as a reference in other areas of the programme. Development and delivery of the artefact will be led by a single Creator (acting as the lead author) who will be selected as part of the Phase 1 review process and delivered in Phase 2 of the project (more information on the approach to phased funding can be found below).

To maximise utility from the outputs, we plan to require all work in TA1 (including TA1.1, TA1.2, and TA1.3) to be open-sourced, permissively licensed and free from patent encumbrances. We envision a community of practice forming around the modelling tools, and we believe that a next-generation open-source framework for "probabilistic models in the large" can have value not just in enabling the rest of this programme but beyond: e.g. in climate modelling, computer-aided engineering, risk assessment for insurance and asset management, and macroeconomic modelling. In the best case, the TA1 approach to uncertain knowledge representation could have an impact analogous to Codd's 1970 relational model [4] and the resulting Structured Query Language, i.e. a new foundation for "data base management" suited to the age of AI, as Codd's model was perfectly suited to the age of networked client-server applications with deterministic data.

Whilst this solicitation focuses on TA1.1, as laid out in the programme thesis [1] (pages 5-12), the wider Safeguarded AI programme is divided into several technical areas (TAs), as follows:

TA1 Scaffolding

- **TA1.1 Theory: this solicitation**
- **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.

TA2 Machine learning

- **TA2.1 World-modelling ML**: to develop fine-tuned AI systems that are fluent in the TA1.1 language of world-models, and interact with users as assistants.
- **TA2.2 Proof-search ML**: to develop fine-tuned AI systems as search heuristics for automated proving techniques that interact in the TA1.1 language of proofs with the TA1.2 proof-checker.
- **TA2.3 Training for certifiable ML**: to develop an automated "training loop" for autonomous systems that can be certified as meeting their specifications. The most promising approach is training "backup controllers" that can take over and certifiably ensure safety anywhere in a local neighbourhood of state space containing the reachable set over a short time horizon under an advanced AI system's policy, see also the Black-Box Simplex Architecture [5].
- **TA2.4 Sociotechnical**: to leverage social-choice theory to develop collective deliberation and decision-making processes about AI specifications and about AI deployment/release decisions.

TA3 Applications: to elicit functional and nonfunctional requirements from customers in a particular sector, design simplified test problems on a spectrum of complexity, and ultimately to demonstrate deployable solutions leveraging TA1 and TA2 tools.

We are beginning with TA1.1 because the bulk of our programme thesis depends on whether attempts toward the TA1.1 objectives gain traction. The selection of Creators for TA1.2-1.3, 2.1 - 2.3, 2.4, and 3 will be subject to separate competitive solicitations due to be released in the coming months. The TA3 solicitation will be the next of these, due to be released in May. Applications for TA1.2, 1.3, 2 and 3 should not be submitted in response to this call; instead applicants interested in participating in these elements beyond TA1.1 should register their interest by sending an email to clarifications@aria.org.uk and we'll notify you when the other TA solicitations goes live.

Section 3: TA1.1 Technical metrics

Each Creator in TA1.1 is intended to work on a problem which is ***plausibly critical*** for achieving the overall vision of TA1.1. What we mean by this is that for each Creator, it should appear reasonably likely that the optimal plan for achieving TA1.1 (whatever that may be) might well require solving their problem. This does not require that a complete plan to achieve TA1.1 exists, let alone is detailed in the Creator's proposal.

Creators in TA1.1 should, in their proposal:

- Define their problem with multiple formal criteria (of the kind that could *in principle* be encoded in a proof assistant such as Lean 4), likely in addition to informal criteria
- Mention related past work and identify on which criteria they fall short
- Differentiate the problem from the normal course of their research going forward were it not for ARIA, and explain why
- Identify and contrast two or three potential approaches to the problem

It is important to note that a given Creator is not expected to directly address, or even to be familiar with, all the technical concepts invoked in the below desiderata. The fragmentation of the mathematical modelling landscape into clusters which are not familiar with each other's abstractions is part of the problem to be solved, and part of the solution is to nucleate a community of Creators, who will address the desiderata collectively. We plan to transition from an exploratory mode in Phase 1 of TA1.1 to a mode that is more coherently focused on a single cohesive solution in Phase 2 of TA1.1.

Desiderata for the eventual language(s) include:

- Many, if not all, of the following kinds of system could be representable as "world models":
    - Petri nets (PNs)
    - Differential equations
        - ordinary (ODEs)
        - partial (PDEs)
        - stochastic (SDEs, SPDEs)
        - random (RODEs, RPDEs)
        - jump-diffusion
    - Markov processes
        - discrete-time Markov chains (DTMCs)
        - continuous-time Markov chains (CTMCs)
        - Markov decision processes (MDPs)
        - Markov automata (MA)
        - open games
        - n-player stochastic games
        - decentralised partially observable Markov decision processes (Dec-POMDPs)
        - mean-field games
    - Hybrid systems
        - Generalised stochastic hybrid systems (GSHS)
    - Probabilistic models
        - probabilistic graphical models (PGMs)
            - Bayesian networks (BNs)
            - structural causal models (SCMs)
            - Markov random fields (MRFs)
            - factor graphs
        - corecursive programs in a purely functional probabilistic programming language (PPL)
            - including, notably, autoregressive large language models (LLMs), cf. (Dohan et al., 2020) [6]
        - probabilistic logic programs
            - including probabilistic answer set programs with interval-valued annotated disjunctions
        - score-based generative models (SBGMs)

- World models should be able to combine the following kinds of uncertainty:
    - stochasticity (probability)
    - nondeterminism
    - partiality
    - Note that credal sets (convex sets of probability measures) are an elegant baseline combination of probabilistic and nondeterministic uncertainty.
- World models (critically, potentially of heterogeneous kinds) should be composable in the following ways:
    - alternative "modes" or "phases" or "regimes", with specified (stochastic and/or nondeterministic) potential transitions between them
        - one potential direction for this is lax colimits
    - parallel or concurrently existing "systems", with specified "interfaces" or "boundaries" on which they must be "coherent" or "compatible" or "agree"
        - one potential direction for this is lax limits
    - input-output composition: the behaviour of one system determining the parameters of another
- The language could support stating and verifying the following kinds of relationships between models:
    - exact simulation (in the sense of, and including, bisimulation)
    - coarse-graining, sound abstraction (behavioural containment)
    - approximation (bounded behavioural distance, perhaps Hausdorff-Kantorovich distance)
- The language could support stating and verifying proof certificates of universally quantified disjunctions of conjunctions of linear inequalities between probabilities of predicates at different points in time
- This could involve "ambiguity tubes" in the sense of (Wu et al., 2022) [7]
- The language *may* support a richer probabilistic temporal logic
- The language could support checking at least two different kinds of proof certificates, and be extensible to others, such as:
    - Alethe and/or LFSC certificates from SMT solvers
    - Certificates based on certified abstract interpretation or bound propagation
    - Positivstellensatz certificates for containment of polyhedra
    - Branch-and-bound trees
    - Neural certificate techniques, such as:
        - Neural control barrier functions
        - Reach-avoid supermartingales
    - Proofs by well-founded induction

- The language could manipulate a large (10^8-parameter) neural network with an efficient on-disk and in-memory data structure, but without abstracting it as a black box (i.e. still able to reason about the piecewise-affine structure and algebraic properties of the architecture).
- The language could represent autonomous controllers which involve conditional composition of neural networks (as in simplex architectures), and constructions such as backtracking tree search, not merely monolithic end-to-end neural controllers.
- The language should support a natural concept of version control, in which incremental changes to components can be easily and efficiently propagated to incremental changes of entire systems, and require only incremental changes to neural systems proof certificates.

We welcome solutions in the following themes:

- string diagrams
- Functorial semantics
- generalisations of probabilistic programming, e.g. probabilistic logic programming, probabilistic answer set programming
- generalisations of Petri nets, e.g. stochastic dynamic coloured Petri nets
- multiple-categorical structures
- fibrations
- categorical systems theory
- probabilistic temporal logics such as probabilistic μ-calculus
- imprecise probability via convex sets of subprobability measures and/or lower probabilities
- composition via lax colimits and/or related constructions
- hierarchical hypernets

However, as stated above, we are open to solutions that take other approaches to addressing some or many of the desiderata; this is meant to encourage people working in the above areas, rather than to discourage researchers who are not familiar with any of them.

We are looking to fund this Technical Area 1.1 with up to £3.5M in total for the first year. We expect to make 10 to 16 awards in this Technical Area.

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

What we do expect to fund (but not strict conditions):

- **In addition to researchers, we are interested in proposals from creators of pedagogical materials (e.g. maths education videos, interactive tutorials, high-production-value expository blog posts, etc.)** who are willing to offer a substantial amount of their time to collaborate with other researchers in the programme to produce explanations of some of the more arcane mathematical concepts that may be in use, to upskill a broader population of engineers and scientists to be able to derive value from using the programme's modelling framework.
- We are looking for *forward-looking* proposals, i.e. proposals to do work that has mostly not already been done (by you or others). While preliminary results are a good signal, it should be clear how you hope to extend the results and/or to remove necessary preconditions.
- We are looking for proposals to solve theoretical problems that seem tractable (albeit perhaps speculative or with ambitious stretch goals), where the solutions will be relevant and important for ultimately addressing many of the objectives below. Tractability means that you are confident, and can give us confidence that there are means to spend your time in a way that likely translates to progress on the problem.
- We will seriously consider proposals that differ substantially from the specific solution themes outlined in the programme thesis, but make a compelling case for their differing approach to solve a similar scope of problems.
- We do encourage proposals that, for as many members of a group as are interested and available, include a majority of their time (we do not encourage joint proposals across more than one recipient institution, and/or across groups that are not likely to actually work together on a regular (~weekly) basis).
- We are keen to explore structures not typical in academic research such as supporting early career researchers as project leads or funding >80% of senior academics' time so that they can focus fully on their ARIA project.

Proposals we do not expect to fund:

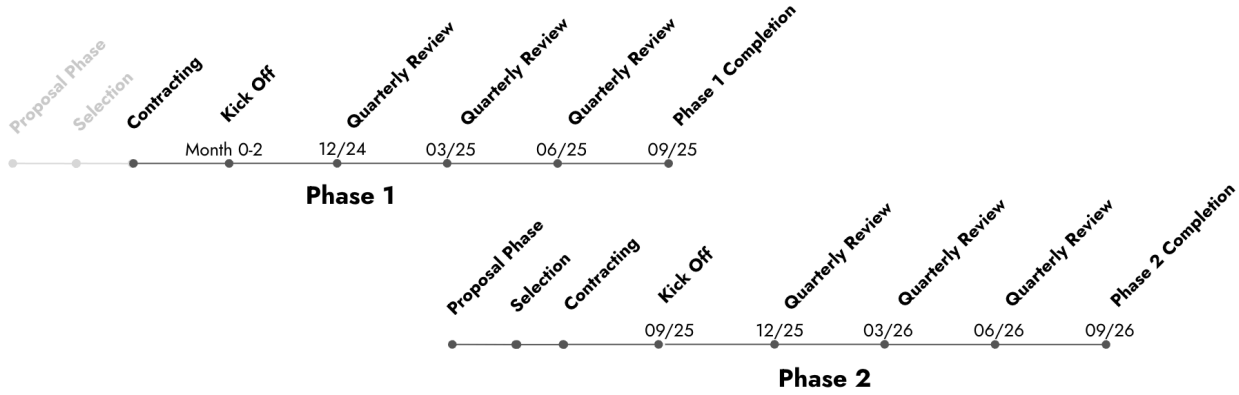
- We are not looking for *complete* proposals that purport to address all the objectives above, but rather identifying key subproblems that are plausibly critical in order to do so.

- In this solicitation we are not yet looking for proposals to build software, except as part of "experimental mathematics", development of theory using a proof assistant, or as early proof-of-concept research code to demonstrate a new idea.
- We are unlikely to be interested in proposals for approaches that have substantial conceptual friction with a probabilistic approach or with a causal approach to modelling.

Section 5: Project duration and management

We expect to commit funding for TA1.1 projects until September 2025, as such applicant proposals should not exceed this timeframe. Based upon the outputs of the initial funding we intend to make decisions in March 2025 about a second phase of funding, to be delivered from September 2025 until September 2026 (Phase 2). Where we decide to release funding for Phase 2, existing Creators will be required to submit proposals for Phase 2 funding. At the same time we will release a solicitation for TA1.1 Phase 2 inviting new applicants to submit new proposals.

The estimated budget for Phase 2 is £1.9m.



Project milestones

We're looking to fund the research effort/time of you and your team toward solving problems that are *plausibly critical* toward achieving desiderata along the lines of the themes above.

TA1.1 is not primarily driven by particular figures of merit in technical performance. Rather, applicants should propose the concrete problem(s) which their initial research direction aims to solve, including specific success criteria, which will likely be theoretical (e.g. if we

can prove these 3 properties of a construction, that is a success), and define a time-frame on which they expect to be able to resolve that question if the answer is positive. Teams that are successfully selected for award will enter into a contracting phase with ARIA where the specific scope of work will be finalised.

Throughout the course of theoretical projects, it may become clear that the project's initial problem statement isn't quite "right" (or isn't tractable), and it's better to change course and try to solve a different problem. In these circumstances suggested pivots must be discussed with the Programme Director, and likely with other Creators as well.

In addition to the the success criteria above, we'll review collaboration amongst TA1.1 Creators and consider:

- Are your outputs being leveraged by other TA1.1 Creators?
- Are you leveraging the outputs from other Creators?
- How is your collaborativeness rated by peers?
- Is there excitement about your outputs from outside the programme?

We will measure these factors by conducting a biannual confidential peer-review survey of all programme Creators.

What these metrics are collectively attempting to approximate is how much Shapley Value your contributions have regarding the expected value of TA1.1 as a whole achieving its technical objectives and its objective of nucleating a research community beyond the programme.

At such time as TA3 Creators are identified and have begun to produce "simplified test problems", we may be able to define more quantitative metrics about the number of problems to which your methods are applicable, or the description length of problems in a particular language fragment, but such metrics are speculative at this stage.

Approach to intellectual property

In TA1.1 of this programme we are pursuing a highly open approach. Intellectual property created by projects funded in TA1.1 shall be:

- Published under a Creative Commons Attribution (CC-BY) licence, if not software
- Dual-licenced under an MIT licence and an Apache 2 licence, if software
- Subject to a patent non-aggression pledge (example), if patented

The intent of the dual-licence requirement above is to provide users with a concise selection of licensing options in order to maximise the openness of the source code. This approach

offers users the flexibility to select either the MIT or Apache 2 licence downstream of the initial development. By doing so, a broader spectrum of users can benefit from the material because it expands the compatibility with various other licences, while also affording users the freedom to choose based on their preferences and needs.

Programme and project management

Alongside our standard project management requirements (light touch quarterly reporting on progress and cost information), we expect our portfolio of TA1.1 Creators to meet each other at quarterly meetings and spend substantial time discussing potential interfaces or synergies between their work and ways that they might be able to adjust their abstractions to be more compatible, either by feeding into each other, or as alternative approaches for solving the same crystallised problem statement.

In due time, we also expect TA1.1 Creators to meet and work with creators from other TAs: to work toward using their research results to formalise test problems from Creators in TA3 problem domains; to assist TA1.2 creators in eliciting and formalising requirements for computational implementations; to educate TA2 creators about their frameworks and abstractions; and potentially other collaborations.

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: Eligibility and application process

Eligibility

We welcome applications from across the R&D ecosystem, including individuals, universities, research institutions, small, medium and large companies, charities and public sector research organisations.

Application process

The application process for technical areas 1.1 consists of one stage:

Stage 1 - Full proposals

This step requires you to submit a detailed proposal including:

- **Project and 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.

For more details on the evaluation criteria we'll use, click here.

Non-UK applicants only

Our primary focus is on funding those who are based in the UK. However, funding will be awarded to organisations outside the UK if we believe it can boost the net impact of a programme in the UK. If you are a non-UK applicant, you must therefore outline any proposed plans or commitments that will contribute to the programme in the UK within the project's duration.

If you are successfully selected for an award subject to negotiations this proposal will form part of those negotiations and any resultant contract/grant.

More information on the evaluation criteria we will use to assess your answers can be found later in the document here.

If you are a non-UK applicant we have provided some additional guidance in our FAQs including available visa options.

Section 7: Timelines

This call for project funding will be open for applications as follows (we may update timelines based on the volume of responses we receive):

Applications open	**11.04.24**
Full proposal submission deadline	**28.05.24 (12:00 BST)**
Full proposal review	**24.06.24**

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

Successful/Unsuccessful applicants notified	**10.07.24**

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 8 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 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 (based on the objectives of TA1.1. We expect resultant awards to use the Basic Grant Agreement template).

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. The costs and timelines proposed are reasonable/realistic.

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.

4) **Responsible** – The proposal identifies any 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 – Applicable to non-UK applicants only –** There is a clear case for how the research will benefit the UK. Proposals originating from applicants outside the UK who seek to establish operations inside the UK, perform a majority of the research inside the UK and present a credible plan for achieving this within the programme duration will be deemed 'UK Applicants' (note this will be reflected in your contract terms).

   For all other non-UK applicants we will evaluate the proposal based on its potential to boost the net impact of the programme in the UK. 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 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

Section 10: References

[1] Dalrymple, D. (2024). *Safeguarded AI: constructing guaranteed safety*. [online] 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.

[3] Wing, J. (2021). *Trustworthy AI*. Communications of the ACM. Available at: https://cacm.acm.org/research/trustworthy-ai/.

[4] Codd, E.F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6), p.387. Available at: https://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf.

[5] Sanaz Sheikhi, Mehmood, U., Bak, S., Smolka, S.A. and Stoller, S.D. (2024). The black-box simplex architecture for runtime assurance of multi-agent CPS. *Innovations in systems and software engineering*. doi:https://doi.org/10.1007/s11334-024-00553-6.
[6] Dohan, D., Xu, W., Lewkowycz, A., Austin, J., Bieber, D., Lopes, R.G., Wu, Y., Michalewski, H., Saurous, R.A., Sohl-dickstein, J., Murphy, K. and Sutton, C. (2022). *Language Model Cascades*. arXiv.org. doi:https://doi.org/10.48550/arXiv.2207.10342.
[7] Wu, F., Villanueva, M.E. and Houska, B. (2022). Ambiguity Tube MPC. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2206.09085.