Advanced Research + Invention Agency ARIA

Safeguarded AI:
TA1.2 Backend &
TA1.3 Human-Computer Interface
Call for proposals

Date: 11 February 2025

 ARIA Copyright © Advanced Research and Invention Agency 2025

Advanced Research + Invention Agency ARIA

Advanced Research + Invention Agency ARIA

SUMMARY

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 & technologically possible.

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

This Solicitation. This funding call is focused on Technical Areas (TAs) 1.2 and 1.3.

+ TA1.2 will develop the **'Backend'**: a professional-grade computational implementation of the mathematical modelling language developed in TA1.1.
+ TA1.3 will develop **'Human-Computer Interfaces'** across the programme, delivering a high-quality user experience to help diverse stakeholders interact with the systems being built in TA1.2 and TA2.

Logistical Summary. We expect to fund 3-6 teams across TA1.2 & TA1.3, for a total of £14.2m. The projects will last 27 months in the case of TA1.2, and 30 months for TA1.3, with a 'go/no-go' decision point after the first 12 months. Applicants can apply to either or both of these TAs at the same time.

Application deadline	April 09, 2025
Kickoff	Early June, 2025
Duration of TA1.2 & TA 1.3	27-30 months, with a 'go/no-go' decision point after 12 months
Total number of teams	3-6 teams across both TA1.2 & TA1.3
Total funding available	£14.2m

 ARIA Copyright © Advanced Research and Invention Agency 2025

SECTION 1: Programme thesis and overview

Today's AI is brilliant in many ways, but it is also unreliable. This unreliability imposes significant societal safety risks and limits our ability to govern these systems in robustly beneficial and legitimate ways. The Safeguarded AI programme is a £59m-backed R&D effort to develop 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. In doing so, we seek to demonstrate the viability of a new, alternative pathway for research and development toward safe and transformative AI.

Safeguarded AI envisions a R&D pathway for leveraging 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. A 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 cyber-physical application domains where reliability is key. It will also reduce the risks of frontier AI by providing high-assurance safety guarantees and building up large-scale civilisational resilience, thereby reducing humanity's vulnerability to potential future "rogue AIs" to an acceptable level within an acceptable time frame.

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 in other application areas and in other deployments around the world.

The Safeguarded AI programme is divided into three main Technical Areas (TAs).

+ **TA1 ('Scaffolding')** will build out the general-purpose scaffolding for the Safeguarded AI workflow. This includes developing a general-purpose modelling language ('syntax') (**TA1.1 'Theory'**), producing a computational implementation of that language (**TA1.2 'Backend'** – *this solicitation*), building the human-computer interfaces to help domain experts develop and refine formal world models and specifications about their domains of interest (**TA1.3 'Human-Computer Interfaces**

(HCI)' – *this solicitation*) and developing the socio-technical integrations to ensure that diverse groups of stakeholders can collectively deliberate about safety specifications and acceptable risk thresholds for AI (**TA1.4 'Socio-technical Integration'**).

+ **TA2 ('Machine Learning')** will develop the ML elements which harness frontier AI techniques into a general-purpose Safeguarded AI workflow.
+ **TA3 ('Applications')** will develop and prototype domain-specific applications of the Safeguarded AI workflow.

We fund several R&D teams—which, at ARIA, we call *Creators*—across all of these Technical Areas, and systematically facilitate interactions and collaborations between Creator teams and across the entire programme, along programme-specific interfaces. Please see Appendix A for a short summary of the programme, and visualisation for how the TAs fit together. Read the programme thesis [1] for a longer/technical explanation of the whole programme.

SECTION 2: Objectives of TA1.2 & TA1.3

TA1.2 – 'Backend'

The objective of TA1.2 is to develop the 'Backend', a professional-grade computational implementation of the mathematical language framework developed in TA1.1 (the 'Theory'). We do not anticipate any single applicant having the requisite variety of skillsets to design and implement the entire backend (though we are not excluding applications proposing to do so), but we hope to assemble the breadth & depth of the required expertise from a small number of applicants each contributing their own specialties.

The first task will be to elicit a detailed requirements document in collaboration with TA1.1 Creators. At a high level, the requirements will include:

+ a **distributed version control system** for structured artifacts (mathematical world models and safety specifications), implemented using principles from database theory
+ **pervasive security-by-design**, including the use of hardened cryptography (e.g. BoringCrypto) to enforce confidentiality and integrity of artifacts at rest and in transit, and a flexible capabilities-based permissions framework
+ a backend to mediate f**lexible interaction paradigms between humans and AI assistants** to collaborate on these artifacts, analogous to the backends for

contemporary environments for AI-assisted software engineering (such as Cursor or Windsurf)
- implementation of **type-checking** for mathematical world models and safety specifications
- implementation of **proof-checking** for various types of formal arguments of safety for hybrid cyber-physical system models, interfacing with multiple formats, such as
  - SMT-LIB/Alethe
  - eRHL
  - Neural barrier certificates
  - Certificates of positivity
  - Branch-and-bound certificates
  - a Lean 4 language server
  - Potential new proof systems developed in TA1.1
- **GPU-optimized implementation of category-theoretic operations** such as computing double colimits or double Grothendieck constructions

You can learn more about the ambitions of TA1.1 here & here, and learn about our TA1.1 Creators here.

In a later stage, in interaction with TA2, the Backend will also develop an interface that can be used by AI agents to verify probabilistic claims about specific checkpoints of domain-specific neural network controllers. The Backend should also be able to produce counterexamples or informative error messages for failed verifications. Depending on the relative capacity of TA2 and TA1.2 teams, the Backend could also be responsible for "compiling" neural networks in a black-box simplex architecture into a deployable executable package that has a high assurance of correctly implementing the exact mathematical function which was verified.

TA1.3 — 'Human-Computer Interfaces'

TA1.3 Creators will develop 'Human-Computer Interfaces' across the programme, delivering an exceptional user experience to help diverse humans interact with the systems being built in TA1.2 and TA2. This will require collaboration with Creators from other TAs (especially TA1.2, TA1.4, TA2 and TA3), to elicit and shape the requirements for these interfaces. We expect that, after an initial phase of orienting and building a toolkit (e.g. a component for direct manipulation of string diagrams), TA1.3 will involve a higher-bandwidth iteration cycle, with frequent prototypes and demos exploring novel paradigms.

Examples of HCI use-cases that will need design and implementation include:

+ Eliciting formal explainable safety specifications from diverse stakeholders;
+ Interactively collaborating with AI assistants in authoring world models and safety specifications;
+ Enhancing domain experts' ability to audit, review and edit scientific/mathematical models;
+ Updating model and safety specifications in light of counterexamples or shortfalls;
+ Reviewing proven guarantees and sample trajectories for spot/sense-checking or more comprehensive red-teaming;
+ Monitoring at run-time whether the incoming observational data is consistent with the mathematical model of the environment in order to spot potentially safety-relevant anomalies;
+ And more.

SECTION 3: Technical metrics

TA1.2 — 'Backend'

At a high level, over the full duration of this TA, success in TA1.2 will be measured according to the concrete requirements from TA1.1 Creators for the implementation of their theory (a mathematical language framework which includes a mathematical modelling language, a task-specification language, and a proof language), as well as the requirements from TA2 Creators concerning the interface with machine learning training loops.

After the first 12 months, there will be a 'go/no-go' decision point. At that time, we will assess the TA1.2 Creators based on their ability to comprehensively solicit and define the requirements for a computational implementation of the mathematical modeling language from TA1.1. If their work proves fruitful, they will then go on to develop some or all of the backend software for the programme for the remainder of the funding period.

TA1.3 — 'Human-Computer Interfaces'

In TA1.3, success will be determined based on user reviews from Creators across the entire programme, evaluating the relevance, usefulness and performance of the delivered HCI features. Successful TA1.3 Creators will (1) identify all high-impact human-computer interaction (HCI) use cases for the programme, (2) collect, test, and refine design specifications in collaboration with relevant Creators, and (3) deliver timely, high-quality solutions that meet these needs.

Once again, there will be a 'go/no-go' decision point at the end of the first 12 months. At this point, TA1.3 Creators will be evaluated based on their ability to identify, scope, design, and test a small set of HCIs for the programme. If successful, they will go on to develop the remaining HCI capabilities for the entire programme for the remaining duration of the funding period.

For both

We expect that the following qualities will be instrumental to / predictive for success in both TA1.2 and TA1.3:

+ A **thorough understanding of the requirements** through high-bandwidth interactions with users (other Creators on the programme), including in a context where those requirements will sometimes only be uncovered/determined in full as the project progresses;
+ **Fast iteration** loops between scoping, design, development, testing, and re-design;
+ **Product/output-orientation & striving for excellence**, including a willingness to go beyond the narrowly defined requisites in order to deliver excellent solutions which lay a solid basis for the Safeguarded AI R&D community, both during and beyond the programme;
+ **High standards of code quality and documentation**, including a focus on **reliability and security**

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

We are primarily looking for applications from software development organisations, especially teams with **strong mathematical backgrounds** (for work in TA1.2), and/or **strong design/HCI/UX capabilities** (for work in TA1.3). Ideal applicants would have a compelling track record for delivering innovative and bespoke software solutions, including working with multiple stakeholder groups to elicit requirements, test and iterate on solutions.

We do not *anticipate* any single applicant having the requisite variety of skillsets for the entirety of these TAs, but we hope to assemble the breadth & depth of the required expertise from a small number of applicants each contributing their own specialties. We also expect that, rather than from the get go having the full capacity required to deliver the proposed projects, some successful applicants will begin to recruit and hire additional staff during the early phase of the project.

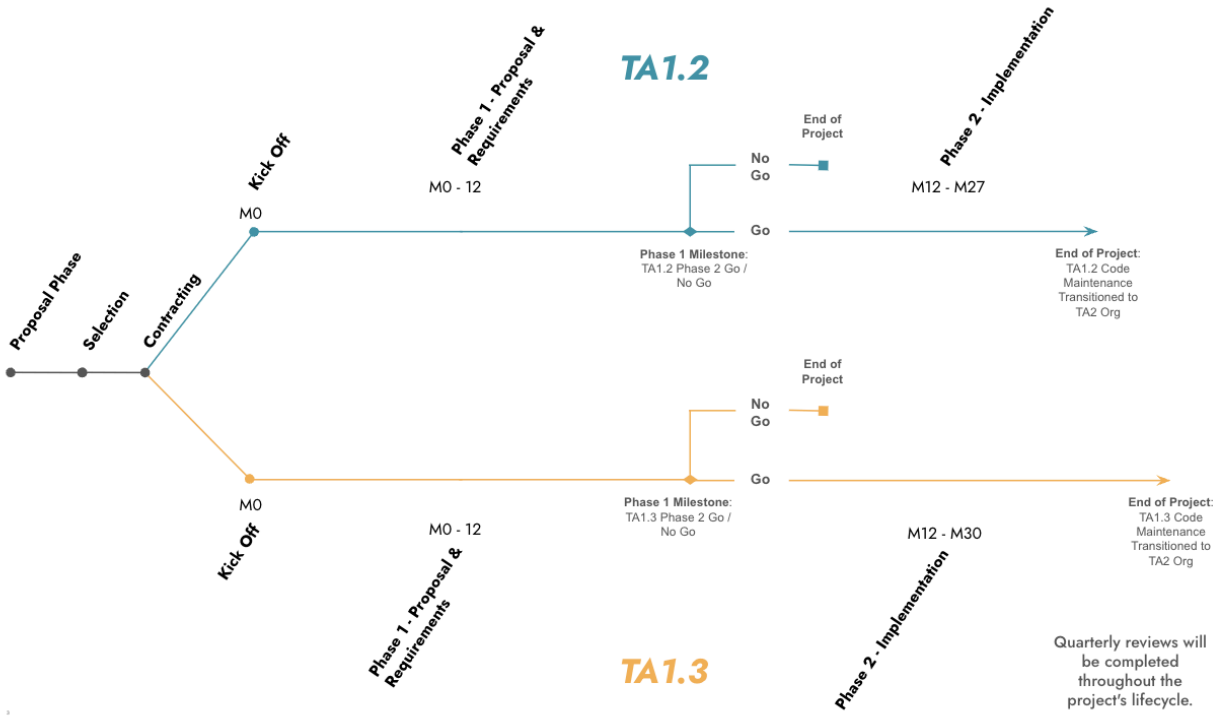
Applicants can apply for either or both of TA1.2 and TA1.3. Applicants applying to both TAs should submit a single application. For applications to a single TA, the page limit is 4; for applications to both TAs, the page limit is 7.

In the context of this funding call, we are looking to fund custom work to support the vision of a Safeguarded AI workflow that can be applied to *any* formalizable domain-specific task, as described in the programme thesis, the accompanying presentation, and the TA1.1 presentation. We are not looking to fund projects which are merely thematically or topically related, such as LLM agent scaffolds or LLM wrapper products for *individual* sectors or applications, for software development in general, nor for notions of AI safety that do not include formal quantitative guarantees.

SECTION 5: Programme duration and project management

Programme structure & duration

TA1.2 & TA1.3 will last 27 and 30 months respectively. We expect to fund 3-6 teams across both TAs, for a total funding of £14.2m. Applicants can apply to either or both of these TAs at the same time.



Creators will be selected, and funding will be awarded for the full duration of the TAs (27 and 30 months respectively). There will be a 'go/no-go' decision point after the first 12 months where each of the Creator teams will be evaluated based on specific success metrics, specified in Section 3. If successful, the teams will continue and potentially expand the scale of their work for the remaining funding period. The intention of this 12-month milestone is to review the progress of the project, ensure there is a viable way forward, pivoting where required and, where no viable way forward is present, close out the projects.

A critical part of the work in all TAs of this programme is collaborating with Creators from other TAs. The programme team will help facilitate these interactions. Among others, we will organise collaborative workshops ('Creator Events') approximately quarterly to bring together Creators from different parts of the programme.

Separately, ARIA will host regular Creator community events across all its 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.

Programme management and project milestones

A suitable project plan—including anticipated human resource leveling over the full 9 or 10 quarters, and an initial draft of quarterly project milestones and deliverables (based on the high-level information in section 2 of this solicitation and other programme materials)—will be proposed by Creators as part of their application, and revised in conversation with the programme team. Our standard project management requirements include light touch quarterly reporting on project progress and actual incurred costs, with more extensive reporting for a few key, agreed-upon milestones/deliverables. In addition to quarterly reports, the programme team will meet with Creators at relevant intervals to discuss their progress.

Each project's progress will be evaluated using clearly defined milestones. Decisions concerning the continuation, pivoting, or termination of a project will be determined by the Creator's ability to meet these (and potentially other, agreed-upon) milestones (where necessary to ensure alignment with broader goals and evolving priorities of the Programme).

Approach to intellectual property

In TA1 of this programme, we are pursuing a highly open approach. Intellectual property created by projects funded in TA1.2 & TA1.3 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.

These norms are chosen for all of TA1 for the purpose of facilitating flow of ideas but also because, in the ultimate vision, the TA1 scaffolding is the platform for a global assurance mechanism that enables multiple actors to verify statements about AI systems complying with internationally agreed norms. The open approach suggested here is critical for facilitating justified trust across the spectrum of stakeholders involved and affected.

SECTION 6: Application & Eligibility

Eligibility

We welcome applications from across the R&D ecosystem, including academia and non-profits.

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.

If your project is to primarily take place outside of the UK, we will ask you in your application to outline any proposed plans or commitments in the UK that will contribute to the programme within the project's duration. 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 1.2 & 1.3 consists of one stage which requires you to submit a detailed proposal 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.

The page limit for applications is 4 pages for applications to one of the TAs (i.e. either TA1.2 or TA1.3) and 7 pages for applications to both TAs at the same time.

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	**11 Feb 2025**
Full proposal submission deadline	**09 Apr 2025 (13:00 GMT)**
Full proposal review	**21 Apr 2025**

If you are shortlisted following full proposal review, you may be invited to meet with the Programme Director and/or Technical Specialist to discuss any critical questions/concerns prior to final selection — this discussion can happen virtually. This is likely to be the 24th and 25th April.

Successful/Unsuccessful applicants notified	07 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 ARIA's Programme Director (PD) and your lead researcher within 15 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.

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. **Please read the FAQs before submitting a question.**

Clarification questions should be submitted no later than 2nd April. 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

Short summary of full Safeguarded AI programme

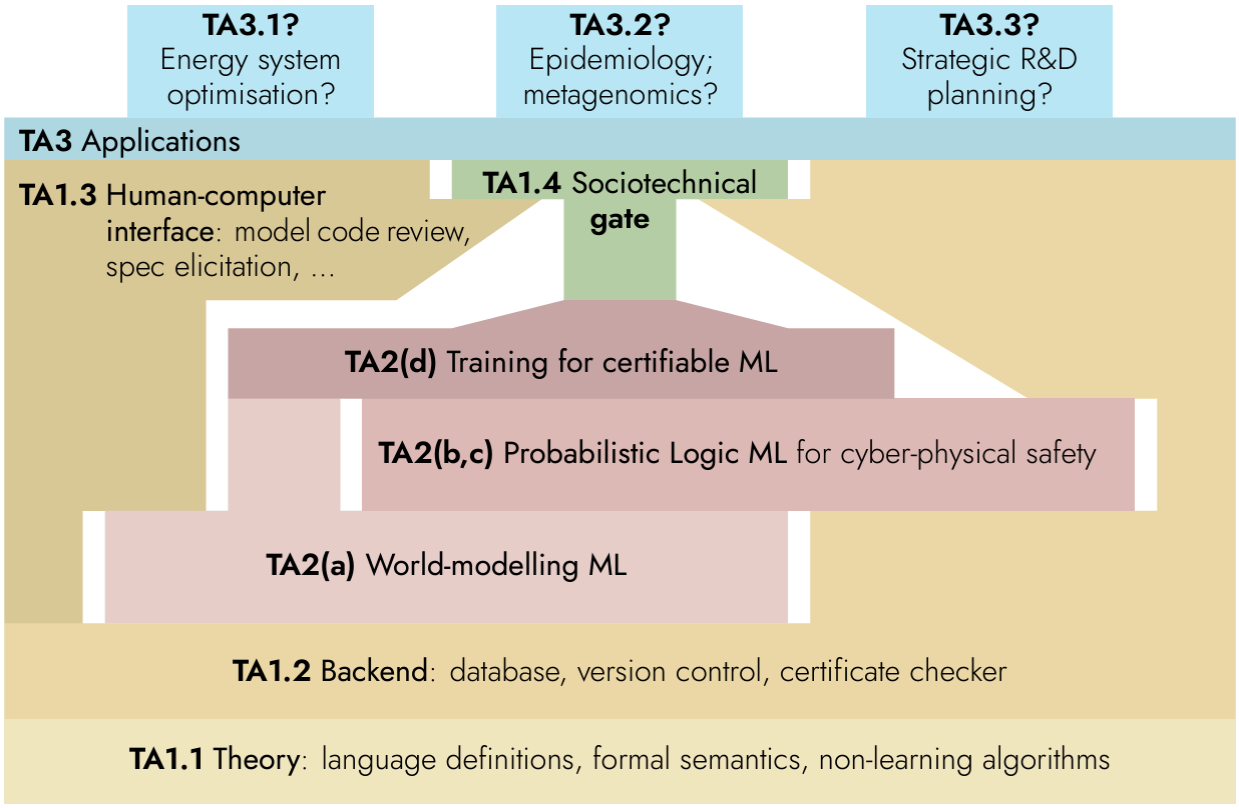
While this solicitation focuses on TA1.2 and TA1.3, the full programme can be found described in more detail in the Safeguarded AI programme thesis [1] (pages 7–13). Below, we provide a brief summary of each of the Technical Areas the programme is divided into.

+ **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 (this solicitation)**: 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 (this solicitation)**: 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 (Phase 1 call for proposals closed on 2.1.2025; you can find more information here)**: to leverage social choice and political theory to develop collective deliberation and decision-making processes about AI specifications and about AI deployment/release decisions, and later to evaluate Safeguarded AI's social impact.
+ **TA2 Machine Learning (Expressions of interest are open for individuals or organisations interested in getting involved in this effort.)**

- **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 (Phase 1 call for proposals closed on 2.10.2024; you can find more information here)**: 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

The following figure provides an overview of Technical Areas and their interfaces, shown visually as horizontal contacts.


TA3.1?
Energy system optimisation?
TA3.2?
Epidemiology; metagenomics?
TA3.3?
Strategic R&D planning?
TA3 Applications
TA1.3 Human-computer interface: model code review, spec elicitation, ...
TA1.4 Sociotechnical gate
TA2(d) Training for certifiable ML
TA2(b,c) Probabilistic Logic ML for cyber-physical safety
TA2(a) World-modelling ML
TA1.2 Backend: database, version control, certificate checker
TA1.1 Theory: language definitions, formal semantics, non-learning algorithms