Safeguarded AI:
TA1.4 Sociotechnical Integration
Call for proposals

Date: 15 October 2024

V1.0

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, mathematical proofs and frontier AI to develop quantitative safety guarantees for AI. We seek to build "gatekeepers": targeted AIs whose job is to understand and reduce the safety risks of other AI systems. 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. It is crucial that the development and deployment of Safeguarded AI is informed and governed in societally beneficial and legitimate ways, through public, representative input. In Technical Area 1.4 Phase 1, we are looking to support teams from the **economic, social, legal & political sciences** to work on projects to ensure the sound **socio-technical integration** of Safeguarded AI systems. TA1.4 Phase 2 – which will be subject to a separate, future funding call – will support projects focused on evaluating the societal impacts of Safeguarded AI systems.

Logistics Summary. TA1.4 Phase 1 is supported by a total of £3.4m, to be distributed across 2-6 teams over a duration of up to 18 months. Phase 2 will be supported by an additional £1m.

Application deadline	2 January 2024
Kickoff	March 2024
TA1.4 Duration	18 months
Total funding available	£3.4m
Total number of teams	2-6 teams

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** will build out the general-purpose scaffolding for the Safeguarded AI workflow. This includes building the tools to help domain experts develop and refine formal world models and specifications about their domains of interest, as well as — the focus on this solicitation — developing the socio-technical interfaces through which diverse groups of stakeholders can collectively deliberate about safety specifications and acceptable risk thresholds for AI.

- **TA2** will develop the ML elements which harness frontier AI techniques into a general-purpose Safeguarded AI workflow.
- **TA3** will develop and prototype domain-specific applications of the Safeguarded AI workflow.

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: TA1.4 objectives

The goal of TA1.4, broadly speaking, is to ensure that Safeguarded AI systems will be developed and deployed in service of humanity at large. We are looking to support teams with relevant expertise in the economic, social, law and political sciences to develop mechanisms, processes, and tools which can be integrated into the Safeguarded AI scaffolding to enable the governance of Safeguarded AI, in both development and deployment.

TA1.4 is divided into two phases. Phase 1 (this solicitation) is focused on research which has the potential to critically inform the design of Safeguarded AI systems in their early stages of development. This phase lasts for up to 18 months, and is supported by £3.4m. Phase 2 will focus on the evaluation of the societal impacts of Safeguarded AI systems. This phase is subject to a separate funding call (expected to go live around late 2026), and will be supported by an additional £1m. Due to the nature of the work we are looking to fund in Phase 2, these projects will only start during the later phase of the Safeguarded AI programme, when the Safeguarded AI technologies are starting to take more concrete shape. This present solicitation is targeted at TA1.4 Phase 1 only.

Examples of open problems we are particularly interested for Creators in TA1.4 Phase 1 to work on include:

- **Qualitative deliberation facilitation**: What tools or processes best enable representative input, collective deliberation and decision-making about safety specifications, acceptable risk thresholds or success conditions for a given application domain, to be integrated into the Safeguarded AI scaffolding (e.g. Lee et al., 2019 [2]; Martin et al., 2020 [3]; Small et al., 2021 [4]; CIP, 2023 [5]; Keswani et al., 2024 [6]; Oldenburg & Xuan, 2024 [7])? How can limitations of existing approaches be avoided or overcome (e.g. Feffer et al., 2023 [8]; Lambert et al., 2023 [9]; Boerstler et al., 2024 [10]; Zhi-Xuan et al. 2024 [11])?

- **Quantitative bargaining solutions**: What social choice mechanisms or quantitative bargaining solutions could best navigate irreconcilable differences in stakeholders' goals, risk tolerances, and preferences, in order for Safeguarded AI systems to serve a multi-stakeholder notion of public good (Cornitzer et al., 2024 [12])? This could include, for example, aggregation methods for reach-avoid specifications that results in an overall trajectory-scoring function (e.g. Wolpert & Bono, 2010 [13]; Leahy et al., 2023 [14]; Watanabe, 2024 [15]), or frameworks for exploring the Pareto frontier using offline multi-objective RL and finding a policy that implements the bargaining solution in a reasonable number of iterations (e.g. Chen et al., 2019 [16]; Roijers et al., 2021 [17]; Lu et al., 2024 [18]; Dima et al., 2024 [19]; Yuan, et al., 2024 [20]).
- **Governability tools for society**: How can we ensure that Safeguarded AI systems are governed in societally beneficial and legitimate ways (Grossi et al., 2024 [21]; Lazar, 2024 [22])? For example, what processes and tools can best elicit societal risk curves[1], which can be used to guide safety evaluations, fairness criteria, and deployment decisions of domain-specific applications of Safeguarded AI? How can we foster other sources of societal legitimacy such representation, accountability (e.g. Raji et al., 2020 [23]) or explanation (e.g. Lazar, 2024 [24]; Munch & Bjerring, 2024 [25])?
- **Governability tools for R&D organisations**: Organisations developing Safeguarded AI capabilities have the potential to create significant externalities – both risks and benefits – to society and humanity. What set of decision-making and governance mechanisms (broadly construed) are best to ensure that entities developing or deploying Safeguarded AI capabilities have and maintain these externalities as appropriately major factors in their decision-making, especially for decisions about deployments, releases, publications, or experiments which could pose a risk of leaking powerful malware (e.g. Cihon, 2021 [26]; Schuett, 2023 [27]; Schuett et al., 2024 [28]; Hendrycks, 2024 [29])?
- **Stewardship towards safe & beneficial socio-economic futures:** Safeguarded AI, if successful, will unlock the automation of economically valuable areas of work. How can we ensure that the resulting benefits are justly redistributed across society, and

[1] Societal risk curves (also called "Farmer's diagrams") are typically represented as F-N curves, mapping the cumulative frequency (F) of hazard events and the gravity of a given hazard measured in the number of fatalities (N) caused. They are commonly used in domains like nuclear energy or aviation to assess and manage acceptable risk levels, and we are interested in extending and adapting their use for high-stakes AI.

that societal downsides, e.g. unemployment risks, are effectively mitigated or balanced out?[2]

We are also open to applications proposing other lines of work which illuminate critical socio-technical dimensions of Safeguarded AI systems, and propose solutions to ensure they will reliably be developed and deployed in service of humanity at large. This might include questions about how Safeguarded AI should be embedded in, and interface with, existing democratic, socio-economic, legal or geopolitical structures. A broader discussion of relevant topics can be found in Critch & Krueger, 2022 [30]).

Depending on the nature of the project, TA1.4 Phase 1 Creators may work in collaboration with Creators from other Technical Areas of the programme. For example, TA1.4 Creators might collaborate with Creators for TA1.3 who develop the computational implementation for the world-models and safety specifications, as well suitable human-computer interfaces for their (iterative) development and version controlling. Here, TA1.4 Creators might provide input and specifications regarding the socio-technical affordances that such tooling should provide. In other cases, TA1.4 Creators could offer to analyse, simulate, “stress-test,” “game out,” or “red-team” the governance structures for the TA2 R&D organisation, as proposed by TA2 Creators.

SECTION 3: Technical metrics

Creators in TA1.4 will work on problems that are plausibly critical to ensuring that the technologies developed as part of the programme will be used in the best interest of humanity at large, and that they are designed in a way that enables their governability through representative processes of collective deliberation and decision-making. For TA1.4 Phase 1 projects, success is measured in terms of their ability to shape the design and deployment of Safeguarded AI systems both during and (importantly) beyond the duration of the programme. This shaping would likely take place through one or more of four types of output:

- *Mathematical theories* which e.g. can be used as problem specifications against which safety-verification ML will be trained as part of TA2(c)

[2] For example, one potential approach might be sector- or occupation-specific insurance products, proposals for which would ideally be developed and operationalised to an extent that relevant stakeholders (e.g. insurance companies, pension schemes, etc.) could take them up for implementation.

- *Computational/algorithmic solutions* which e.g. can be invoked in the computational implementation of world-models and safety specifications as part of TA1.2/TA1.3
- *Legally implementable decision-making processes* which e.g. can be adopted by the TA2 organisation (and similar R&D organisations) via their bylaws
- *Crystallised philosophical or strategic insights* which e.g. substantially inform new crucial considerations for decision-making about Safeguarded AI systems

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

We are open to applications from across the R&D spectrum, including academia, nonprofits and for-profits. We are looking for applications from individuals or groups with strong expertise in the economic, social, legal and political sciences, as relevant to the objectives outlined above. Creator teams might also include software development capacity, if their projects include tooling as part of their deliverables.

We're focused on work that is plausibly critical in ensuring that Safeguarded AI technologies will be developed and used responsibly and to the benefit of humanity at large. Notably, while TA1.4 focuses on open socio-technical problems about Safeguarded AI systems, the solutions to these problems do not have to involve AI (although we are also interested in solutions that do directly leverage AI).

We welcome proposals for research projects which span the full 18 months of the TA1.4 period, as well as projects that will conclude sooner.

Work to evaluate the societal impacts of Safeguarded AI systems is out of scope for this solicitation, and will instead be the focus of a future funding call on TA1.4 Phase 2.

SECTION 5: Programme duration and project management

Programme structure & duration

TA1.4 Phase 1 will be supported by a total of £3.4M across 2-6 teams and over a period of up to 18 months.

Each project's progress will be evaluated using clearly defined success criteria. Across all projects, success will be measured in terms of the project's ability to shape the design and deployment of Safeguarded AI systems both during and (importantly) beyond the duration of the programme. Further project-specific success criteria will be defined by the applicant prior to the start of a project and agreed upon by ARIA. Success, pivoting, or termination decisions for each project will be determined by the applicant's ability to meet these agreed-upon criteria.

Programme management and project milestones

Our standard project management requirements include light touch quarterly reporting on progress and cost information. Furthermore, we will meet with all Creators on a quarterly basis to discuss the progress, and facilitate interactions with Creators from other TAs as required. Depending on the nature of work, TA1.4 Creators will interact more or less closely with Creators from other Technical Areas.

Suitable project milestones and deliverables will be proposed by Creators, and decided upon in conversation with the programme team.

Approach to intellectual property

The output of the work carried out as part of TA1.4 will be open-access. Where software or tooling are produced, these will be made open-source, including code and documentation.

These norms are chosen 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.

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 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 (note the maximum project duration is 18 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 1.4 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	**15 October 2024**
Full proposal submission deadline	**02 January 2025 (12:00 GMT)**
Full proposal review	**23 January 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 27th and 28th January.

Successful/Unsuccessful applicants notified	**11 February 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.

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] Lee, M.K., et al. (2019). *WeBuildAI: Participatory framework for algorithmic governance*. Proceedings of the ACM on human-computer interaction 3.CSCW: 1-35 doi:https://dl.acm.org/doi/10.1145/3359283

[3] Martin Jr, D, et al. (2020). *Participatory problem formulation for fairer machine learning through community based system dynamics*. arXiv. doi:https://arxiv.org/pdf/2005.07572

[4] Small, C, et al. (2021). *Polis: Scaling deliberation by mapping high dimensional opinion spaces*. Recerca: revista de pensament i anàlisi 26.2. Available at: https://www.demdis.sk/content/files/2022/11/Polis-manusript.pdf

[5] Collective Intelligence Project. (2023). *Alignment Assemblies*. Available at: https://www.cip.org/alignmentassemblies

[6] Keswani, V, et al. (2024). *On the Pros and Cons of Active Learning for Moral Preference Elicitation*. arXiv. doi:https://arxiv.org/abs/2407.18889

[7] Oldenburg, N, and Zhi-Xuan, T. (2024). *Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games*. arXiv. doi:https://arxiv.org/pdf/2402.13399

[8] Feffer, M, et al. (2023). *From preference elicitation to participatory ML: A critical survey & guidelines for future research*. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. doi:https://dl.acm.org/doi/10.1145/3600211.3604661

[9] Lambert, N, Gilbert, T.K., and Zick, T. (2023). *Entangled preferences: The history and risks of reinforcement learning and human feedback*. arXiv. doi:https://arxiv.org/pdf/2310.13595

[10] Boerstler, K, et al. (2024). *On the stability of moral preferences: A problem with computational elicitation methods*. arXiv. doi:https://arxiv.org/abs/2408.02862.

[11] Zhi-Xuan, T, et al. (2024). *Beyond Preferences in AI Alignment*. arXiv. doi:https://arxiv.org/pdf/2408.16984

[12] Conitzer, V, et al. (2024). *Social choice for AI alignment: Dealing with diverse human feedback*. arXiv. doi:https://arxiv.org/abs/2404.10271

[13] Wolpert, D.H., and J.W. Bono. (2010). *A theory of unstructured bargaining using distribution-valued solution concepts*. No. 2010-14. Available at:

https://aura.american.edu/articles/report/2010-13_Distribution-valued_solution_concepts/23892576?file=41890680

[14] Leahy, K, Mann, M, and Serlin, Z. (2023). *Safety-Aware Task Composition for Discrete and Continuous Reinforcement Learning*. arxiv. doi:https://arxiv.org/abs/2306.17033

[15] Watanabe, K. (2024). *Pareto Fronts for Compositionally Solving String Diagrams of Parity Games*. arXiv. doi:https://arxiv.org/pdf/2406.17240

[16] Chen, X, et al. (2019). *Meta-learning for multi-objective reinforcement learning. International Conference on Intelligent Robots and Systems* (IROS). Available at: https://ieeexplore.ieee.org/abstract/document/8968092

[17] Roijers, D.M., et al. (2021). *On following pareto-optimal policies in multi-objective planning and reinforcement learning. Proceedings of the Multi-Objective Decision Making (MODeM) Workshop*. Available at: https://modem2021.cs.universityofgalway.ie/papers/MODeM_2021_paper_3.pdf

[18] Lu, H, Herman, D, and Yu, Y. (2023). *Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. The Eleventh International Conference on Learning Representations*. Available at: https://openreview.net/forum?id=TjEzIsyEsQ6

[19] Dima, S, et al. (2024). *Non-maximizing policies that fulfill multi-criterion aspirations in expectation*. arXiv. doi:https://arxiv.org/pdf/2408.04385

[20] Yuan, Y, et al. (2024). MODULI: Unlocking Preference *Generalization via Diffusion Models for Offline Multi-Objective Reinforcement Learning*. arXiv. doi:https://arxiv.org/pdf/2408.15501

[21] Grossi, D, et al. (2024). *Enabling the Digital Democratic Revival: A Research Program for Digital Democracy*. arxiv. doi:https://arxiv.org/abs/2401.16863

[22] Lazar, S. (2024). *Automatic Authorities: Power and AI*. arXiv. doi:https://arxiv.org/abs/2404.05990

[23] Raji, I.D., et al. (2020.) *Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 conference on*

fairness, accountability, and transparency. doi:https://dl.acm.org/doi/pdf/10.1145/3351095.3372873

[24] Lazar, S. (2024). *Legitimacy, Authority, and Democratic Duties of Explanation*. In: *Oxford Studies in Political Philosophy Volume 10*. Edited by: David Sobel and Steven Wall, Oxford University Press. Available at: https://academic.oup.com/book/56337/chapter-abstract/445461225?redirectedFrom=fulltext

[25] Munch, L, and J.C., Bjerring. (2024). *Can large language models help solve the cost problem for the right to explanation?. Journal of Medical Ethics.* Available at: https://jme.bmj.com/content/early/2024/09/12/jme-2023-109737.abstract

[26] Cihon, P, Schuett, J and S.D., Baum. (2021). *Corporate governance of artificial intelligence in the public interest. Information* 12.7: 275. Available at: https://www.mdpi.com/2078-2489/12/7/275

[27] Schuett, J. (2023). *AGI labs need an internal audit function*. arxiv. doi:https://arxiv.org/abs/2305.17038

[28] Schuett, J, Reuel, A, and Carlier, A. (2024). *How to design an AI ethics board. AI and Ethics*: 1-19. Available at: https://link.springer.com/article/10.1007/s43681-023-00409-y

[29] Hendrycks, D. (2024.) *Section 8.4 Corporate Governance, in Introduction to AI Safety, Ethics and Society. Taylor & Francis, (forthcoming)*. ISBN: 9781032798028. Available at: https://www.aisafetybook.com/textbook/governance

[30] Critch, A, and Krueger, A. (2020). *AI research considerations for human existential safety (ARCHES)* arxiv. doi:https://arxiv.org/pdf/2006.04948

APPENDIX

Short summary of full Safeguarded AI programme

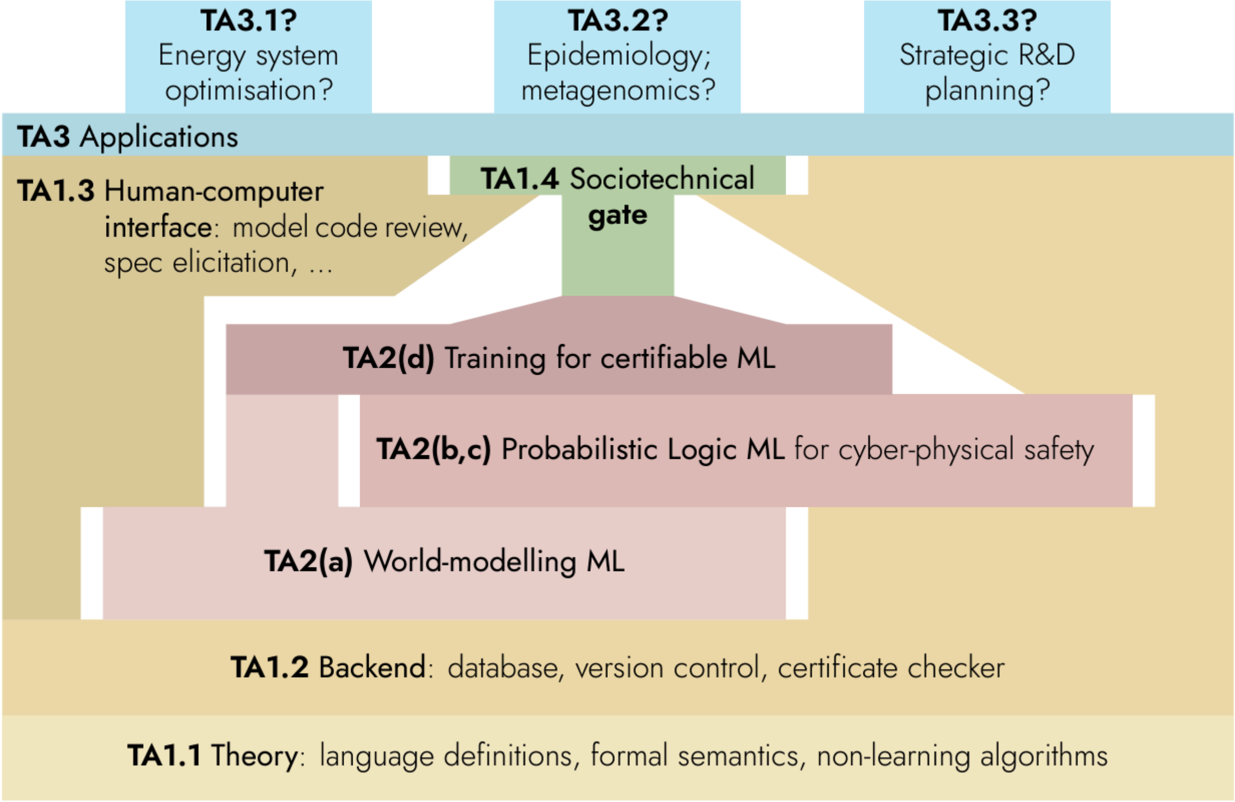
While this solicitation focuses on TA1.4, the full programme can be found described in more detail in the Safeguarded AI programme thesis [1] (pages 7—13). Below, we provide a brief summary of each of the Technical Areas the programme is divided into.

- **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 (including this solicitation)**: 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 deployale 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