Advanced Research + Invention Agency ARIA

Safeguarded AI: Cybersecurity

Call for Proposals

Date: 20 May 2026

V1.0

SECTION 1: Broader context and motivation

Artificial intelligence is advancing at extraordinary speed, enabling and accelerating both offensive and defensive applications. This is especially salient in cybersecurity, where AI is lowering the cost of offensive cyber operations and increasing the number and capability of threat actors.

Recent developments illustrate the scale of this shift. Anthropic's Claude Opus 4.6, for example, discovered 22 vulnerabilities in Firefox over two weeks, 14 of them rated high-severity by Mozilla;[1] Claude Mythos Preview found a 27-year-old vulnerability in OpenBSD (one of the most security-hardened operating systems in the world) that allowed an attacker to remotely crash any machine just by connecting to it, as well as several vulnerabilities in the Linux kernel that enabled attackers to escalate their privileges from ordinary user access to complete control of the machine.[2] These are not isolated findings.[3] With AI continuing to progress, the barriers to finding and exploiting software vulnerabilities are collapsing. The capabilities and population of threat actors will only grow as these tools become better and more widely available.

In the current paradigm, attackers hold a structural advantage: defenders must secure every entry point, while an attacker needs only one sufficient chain of exploits to cause large-scale damage. The result is a resource-intensive cat-and-mouse game in which offence is favoured even when capabilities are evenly matched.

We think this advantage is not fundamental, but an artefact of how we currently build software. The domain of pure cyber is fundamental defense-favoured, if we build the capabilities that unlocks this reality. To see why, it helps to scope the claim precisely. By pure cyber we mean operations conducted by sending bits through wires: the behaviour of code running on hardware, reachable over a network or a bus.[4] This means that the cyber attack surface is bounded and, in principle, fully characterisable. Cyberattacks exploit code that behaves in ways its designers did not intend. An attacker wins by finding an input or interaction that drives the system into an unintended state. If the space of possible behaviours can be specified and the implementation proven to stay within those bounds, there is nothing left for the attacker to find.

---

[1]Anthropic, "Claude and Firefox Security," red.anthropic.com/2026/firefox/ (March 2026)

[2]Anthropic, "Project Glasswing: Securing critical software for the AI era," anthropic.com/glasswing (April 2026)

[3]Other illustrative examples include: Nicholas Carlini, "Black-hat LLMs," [un]prompted (March 2026); AISLE, "AISLE Discovered 12 out of 12 OpenSSL Vulnerabilities" (January 2026); and Anthropic, "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign" (November 2025).

[4] Here, we are excluding attack vectors outside of the realm of pure cyber, such as social engineering, system misconfiguration, supply-chain compromise, and physical side channels. These attack surfaces that are real and consequential, and any serious security posture must address them through defence in depth. Nevertheless, if we are able to shrink the attack surface resulting from cyber alone, we allow defenders to focus their limited resources and raise the overall cost of landing a successful attack.

This is what formal methods deliver: mathematical proof that a system's behaviour is confined to its specification, eliminating entire classes of exploitable incorrectness rather than patching instances one at a time. We have a real-world demonstration that this is possible. Most famously, perhaps, DARPA's HACMS programme showed that formal methods hardened systems can withstand intense red-teaming conditions in real-world, in-use vehicle control systems, including a manned helicopter.[5]

Realising this vision is far from a solved problem, however. For one, formal methods have historically been too slow, expensive, and expertise-intensive to be broadly relied upon: verifying the seL4 microkernel took approximately 27 person-years for 9,300 lines of code, and CompCert required roughly 100,000 lines of Coq proofs and six person-years of effort.

AI is now changing this equation.[6] AI excels at precisely the tasks that have made formal verification prohibitively expensive: exploring vast search spaces for proof candidates. AI performance on relevant benchmarks has been improving. For example, a recent, large-scale 'vericoding' benchmark spanning 12,504 formal specifications across Dafny, Verus/Rust, and Lean found that, across off-the-shelf LLMs, they could successfully generate verified implementations from specs alone for 82% of Dafny tasks, 44% of Verus/Rust, and 27% of Lean.[7] This trend is further backed by growing interest and investment in relevant model capabilities, a growing number of datasets and benchmarks, and finally tooling to help elicit and fully leverage AI's growing capabilities.

Crucially, speed from AI does not require trust in AI. A compact proof-checking kernel can mechanically verify a proof regardless of how it was discovered. This asymmetry is the foundation of the entire field, and it means we can harness AI's speed for the search-intensive work while grounding trust in mathematics. The same frontier capabilities that are lowering the cost of attack can be turned, through this lever, toward producing formally verified defensive components at a pace that begins to match the threat.

Further leverage to defenders can be achieved by pushing the assurance boundary down the stack: from application source code through compiled binaries, to the instruction set architecture,

---

[5]Fisher et al., "The HACMS program: using formal methods to eliminate exploitable bugs," Phil. Trans. R. Soc. A, 2017. Also see: DARPA HACMS case study: darpa.mil/news/resources/case-studies/hacms.

[6]See, e.g., Yang, K. et al. "Position: Formal Mathematical Reasoning—A New Frontier in AI," Proceedings of the 42nd International Conference on Machine Learning, PMLR 267 (2025): 82384–82398; Kleppmann, M., 2025. "Prediction: AI will make formal verification go mainstream," martin.kleppmann.com (2025); Lin, S., Miyazono, E., Winham, D., "A Toolchain for AI-Assisted Code Specification, Synthesis and Verification", Atlas Computing (2025); Barrett, C. et al. "Certificates in AI: Learn but Verify," Communications of the ACM 69, no. 1 (2026): 66–75.; de Moura, L. "Proof Assistants in the Age of AI," leodemoura.github.io/blog/ (2026).

[7] Bursuc, Sergiu, et al. "A benchmark for vericoding: formally verified program synthesis." arXiv preprint arXiv:2509.22908, 2025.

and, ultimately, toward RTL and hardware semantics themselves. This matters because each layer below the verified one is a layer where an attacker can still operate. Even correct source code can compile to vulnerable binaries; compilers can introduce bugs absent from the original program; and at the hardware level, microarchitectural side channels and software/hardware boundary issues open further attack surface. As AI-driven offence matures, attackers will have an increasingly easy time to attack those lower layers too. Conversely, each layer the proof boundary descends shrinks the trust base and retires entire exploit classes at once, rather than patching their instances.

This is not something that's currently feasible out-of-the-box. In many cases we lack the formal semantics needed to state properties in the first place: ISA behaviour is typically specified only in prose documents hundreds of pages long, with ambiguities that surface when two implementations disagree; weak memory models are still an active research frontier; and proof effort has historically scaled poorly with system size. Closing these gaps requires investment on two fronts: expanding the class of systems we can formally model, and building tooling that lets proof effort scale with AI capabilities rather than human labour. If we succeed, the scope, scale and speed of formal verification expand beyond what is currently considered possible, and cybersecurity moves into a defence-favoured regime.

ARIA's Safeguarded AI programme seeks to accelerate this transition away from the perpetual, resource-hungry cat-and-mouse game between attackers and defenders, toward a regime where defenders hold the structural advantage, systematically retiring entire classes of exploits and building provably secure isolation boundaries to contain the blast radius of those that remain. The programme's Technical Area 1 invests in the underlying machinery: expanding the type of systems formal methods can address, and delivering a step change in the scale of systems we can verify. Technical Area 2, of which this funding call is part, pursues the complementary question: **given the machinery we have and are building, what are the most ambitious, security-critical systems we can verify now?**

The stakes are huge. The total cost of cyberattacks worldwide is estimated at multiple trillions of pounds annually.[8] Of this, the fraction that is pure cyber (i.e. not reliant on phishing, social engineering, or supply-chain compromise) is estimated at 30–40%.[9] On this basis, a robust solution to cyber defence would have a global public-good value at the order of at least a trillion pounds annually, even before accounting for the anticipated rise in threat actor capability.

[8]European Commission, "A cybersecure digital transformation," (2019). The Commission estimated the global annual cost of cybercrime at £5 trillion in 2020, representing a doubling compared to 2015. A naive projection of this trend would suggest a cost of £10 trillion by 2025.
[9]Google Cloud, M-Trends 2025 (2025); Verizon, 2025 Data Breach Investigations Report (2025).

Beyond economic value, hardened cyber infrastructure is load-bearing for protecting the health, privacy, security and prosperity of the public, and the trustworthiness of our institutions.

SECTION 2: About this funding call

Objectives

Backed by £20m, this funding call seeks to test and accelerate the hypothesis that AI-enabled formal methods can make high-assurance cyber defence practical at scale.

We will fund teams to build production-grade, security-critical software components whose key security properties are backed by machine-checked proofs, validated through coordinated red-team exercises. The programme targets high-leverage defensive primitives where proven correctness yields outsized resilience gains – from hardening critical infrastructure network perimeters, to securing the AI inference stack, to verifying the proof infrastructure itself. Targets are chosen in line with the high-level threat model outlined in section 1: a world where AI-driven increasing the offensive pressures on all sorts of critical software systems, with a growing population of threat actors with highly sophisticated capabilities.

Formal verification eliminates bugs relative to a specification, but the specification itself may be incomplete, the threat model may omit real-world attack paths, and the trusted computing base may contain assumptions that fail in deployment. This is why, alongside the Blue Teams, we are looking to work with a red-teaming partner to deliver rigorous adversarial evaluations of the resulting systems. The Red Team's job is to find exploits and document them for review by the Blue Teams' and the programme team.

Our ambition includes real-world adoption. We will evaluate teams for their credible interest and ability to carry developed capabilities towards large-scale deployment, whether themselves or through spinouts, partnerships, upstreaming, or other routes. We will actively support them in doing so through our networks, advice, and translation support and, where a team reaches a clear technical and deployment inflection point during the project period, may provide additional funding to move into an intensified delivery phase.

Two Tracks

The call is structured as two tracks: Track 1 (Blue Teams) for building and verifying security-critical components, and Track 2 (Red Team) for adversarial evaluation.

- **Track 1: Blue Teams.** We will fund 3–6 teams to build production-grade, security-critical software components whose key security properties are backed by machine-checked proofs, under clearly stated threat models and assumptions, using AI as

a central means of reaching levels of ambition for formal proof that would otherwise not be tractable.

- **Track 2: Red Team.** We are also looking for applications to fund one central Red Team. They will be responsible for executing, within each sprint cycle, a thorough pen-testing effort of every Blue Team's system, and document the results in a report. Just like for Blue Teams, the intelligent use of AI will play a central role in thorough Red Teaming.

Funding for the track 1 blue teams will be provided under our R&D funding terms (see here for more detail), while funding for the track 2 red team will be provided under our commercial service terms (see here for more detail).

Note that Red and Blue Teams need to be independent. While applicants are free to apply to both tracks, we will not appoint the same organisation or team to serve in both capacities.

Summary Table

Objective	Demonstrate that AI-enabled formal methods can make production-grade cyber defence practical by building and verifying high-leverage cybersecurity components
Total funding available	£20m
Number of teams	Track 1 (Blue Teams): 3–6 teams Track 2 (Red Teams): 1
Per project funding	Track 1: approx. £2.5–3.5m Track 2: approx. £2–3m
Application deadline	**1st July 2026 (14:00 BST)**; Track 1 Blue team - rolling thereafter until end of 2026
Project kickoff	1st September 2026; rolling thereafter
Project duration	10-15 months (from project start until end of November 2027)
Application link	Apply here
Application	Single stage; max. 5 page proposal track 1 (Blue Teams); max. 3 page proposals for track 2 (Red Teams)

Sprint Structure

As capabilities mature, the scale of systems within reach for formal methods will expand dramatically. To continuously calibrate our ambition to the evolving realities, projects will run in short technical sprints. Each 8-week cycle consists of:

- **Build and verify (6 weeks)**. Blue Teams build and formally verify a system or component against a declared specification and threat model.
- **Red-teaming (1 week).** The Red Team, which will have had full viewing access to the Blue Teams' development and verification process, attempts to break the system and delivers a detailed report covering attack attempts, outcomes, and reproducible artifacts for any successful breaks.
- **Review and retarget (1 week)**. Blue Teams review the red team's reports, evaluate how their processes can be improved, and define the target for the next sprint, in conversation with the ARIA programme team.

This rhythm allows us to iterate and learn quickly, remaining continuously calibrated to where AI-enabled formal methods capabilities have reached, and what the most security-critical targets are that we could tackle.

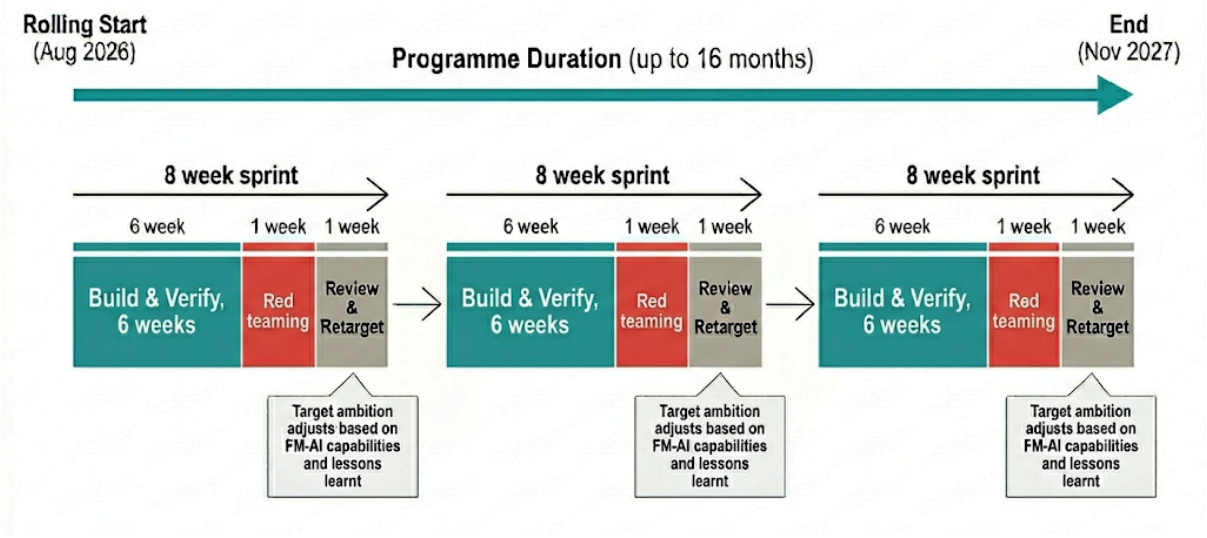

Fig. 1: Chart showing the overall programme duration timeline and the breakdown of individual 8-week sprint cycles.

Initial and follow-on targets. Track 1 applicants should propose a tractable but ambitious and valuable initial target, together with plausible candidates for more ambitious follow-on targets (possibly building on top of each other). Note that the final target selection will be subject to discussions with and approval from the programme team, both at the start of the project and for each sprint cycle.

Programme Supports

We will be working closely with teams on technical progress, strategic orientation and deployment pathways.

Collaborations & teaming. We understand that applicants may wish to find collaborators with complementary expertise across specific target domains, security, formal methods, and AI. To assist in this process, applicants are welcome to optionally submit their information (area of expertise, what you're looking for in collaborations, contact information) to our central teaming platform here, and identify others they may wish to reach out to.

Budgeting. We expect that the vast majority of costs of these projects will go towards funding a small, high-calibre technical team (typically 2-4 people) and compute costs. For compute cost, we estimate that teams will budget up to £200k per sprint *on average* (accounting for the fact that teams' ability to make effective use of computers will likely increase significantly over the course of the projects).

Learning & iteration. Sprint cycles are designed to enable fast learning and iteration. At the end of each cycle (week 8), Blue Teams assess the red-team report, draw out learnings, and select a new target. They then meet with the programme team to present their reflections and discuss the new target, before kicking off the next cycle. The process will be kept pragmatic and low-overhead, capturing the learning needed to retarget effectively while preserving the pace of the sprint cycle.

Stakeholder group. ARIA maintains a network of contacts across the cybersecurity community, critical national infrastructure operators, and other similar end users, whose operational knowledge is essential to getting specifications right and whose adoption of verified components is the programme's ultimate test of success. These stakeholders will help ground specifications and threat models in real operational conditions, and open pathways to adoption that would otherwise be hard to reach. We will work to make this network an active asset for Creator teams.

Creator events. We will organise two events throughout the project period where Creator teams will showcase their work to date, exchange methodological and strategic insights and connect to relevant stakeholders across cybersecurity, formal methods, AI, and real-world deployment environments. Separately, ARIA hosts regular Creator community events across all its programmes, for which attendance is encouraged but will not be mandatory.

Translation & adoption support. We will work with teams to identify and develop credible routes to real-world use, whether through direct deployment, upstreaming, partnerships, licensing, open-sourcing, or spinout formation. This may include support with end-user discovery, specification validation, stakeholder introductions and connections to relevant funders, investors, or adoption partners, drawing on ARIA's internal translation expertise, networks, and

Activation Partners. Where a team reaches a clear technical and deployment inflection point, ARIA may also consider deploying additional funding to support a shift into a more intensive delivery phase.

Timelines

The first batch of applications is open until July 1st. We expect to inform successful teams by July 31st.

Projects will kick off on September 1st, entering straight into the first 8-week sprint cycle. This sprint rhythm will continue through the programme, with short buffer periods around major programme moments: an interim milestone and workshop after the first three sprints, and a final showcase and reporting period towards the end of the programme. The programme will conclude at the end of November 2027.

We anticipate keeping the funding call open until early 2027, reviewing subsequent applications at an approximately monthly cadence. Any teams joining after the official start date of September 1st will slot into the sprint cycles at the earliest opportunity.

Approach to Intellectual Property

Our priorities with regard to the systems teams build and verify are **adoption** and **trust**. Our approach to IP is entirely downstream of those considerations.

- **Implementations.** Teams retain ownership of the systems they build and verify, and any tools they might build in the process of doing so. Where teams have not taken steps toward commercialisation or deployment of verified systems within 12 months of project completion, ARIA may choose to reassign the commercialisation rights, or require open-sourcing or permissively licensing to ensure the work reaches deployment.
- **Specifications and proof artefacts.** We require that all specifications and proof artefacts be published openly. Trust in formally verified systems depends on the ability of third parties to inspect the claims and their evidence.

SECTION 3: What we are looking for

Cybersecurity Targets

What sort of systems are we looking for Blue Teams to build and verify?

At a high level, we consider suitable cybersecurity targets to be software components where:

1. **correctness guarantees yield outsized impact** on systemic risk reduction. This includes (but is not limited to) components that:

    a. sit at a trust boundary or choke-point (gateways, verifiers, isolation boundaries, policy enforcement points, boot/attestation chains);
    b. protect many downstream systems (high reuse across sectors and environments, further down the compute stack, clear applicability to critical infrastructure);
    c. have a credible adoption pathway (drop-in retrofitability, upstreaming into widely used stacks, or a realistic deployment plan proportional to the security benefits).
2. we can state a **meaningful security specification** and produce machine-checkable evidence, with a clear account of the threat model, assumptions, and remaining trusted computing base;
3. the level of **ambition** of the targeted component grows commensurate with growing capabilities for AI-enabled formal methods;
    a. Ideal targets will often (but are not required to) admit a nested goal structure, or 'tractability ladder': early, pragmatic targets deliver standalone value while building the capabilities required for progressively more ambitious targets.

We are particularly interested in domains with outsized strategic importance, such as (but not exclusive to):

- **Securing critical infrastructure.** Critical infrastructure[10] is essential to both civilian and national security; its compromise could rapidly cause significant societal damage, including economic instability, disruption of social order, and loss of life. Given the difficulty of adopting in these systems (due to e.g. legacy code and uptime requirements), a particularly promising class of interventions are retrofittable components that harden key trust boundaries without requiring full system redesign, e.g. verified network-boundary and remote-access components, policy-enforcement and identity layers, and security-critical protocol implementations (e.g. TLS, mTLS, certificate validation, cryptographic libraries, code-signing and software update mechanisms) that protect high-value systems at especially sensitive points of failure.
- **Critical software infrastructure:** foundational software components whose failure or compromise would propagate across large parts of the digital ecosystem, including operating systems, browsers, hypervisors, package managers, compilers, cryptographic libraries, container runtimes, and software update mechanisms. We are especially interested in components that sit on security-critical trust boundaries — such as sandboxing, memory isolation, authentication, parsing, dependency resolution, and secure distribution — where AI-enabled formal methods could produce stronger guarantees than today's testing- and review-heavy approaches.

[10] See: https://www.npsa.gov.uk/about-npsa/critical-national-infrastructure

- **Securing the AI inference stack.** The AI inference stack is increasingly becoming critical infrastructure in its own right. Compromise could expose confidential data, enable theft or tampering of weights, cause cross-tenant leakage, manipulate outputs, or disrupt critical services. We are therefore interested in systems that can provide strong assurance for properties such as confidentiality and integrity of weights, prompts, and outputs; tenant isolation; and safe, correct mediation of shared compute.
- **Verifying the proof stack itself.** In a world where we increasingly rely on AI-generated guarantees for safety-critical systems, the soundness of the proof kernel is a foundational dependency. Furthermore, proof checkers will face greater adversarial pressure, as AI-driven proof search will relentlessly find any corner cases, implementation flaws, and soundness breaks in the proof stack itself. Verifying the soundness of this stack (i.e. ensuring that only valid proofs are accepted as such) is therefore increasingly load-bearing for the credibility of downstream assurance claims.

To further illustrate the sort of targets we consider interesting, as well as the level of specificity we expect in proposals, we include an example list in the appendix. We encourage applicants to read these. The list is merely *indicative* — applicants need not confine themselves to these targets.

What are we looking for - Blue Teams (Track 1)

We are looking for teams that are strongly motivated and exceptionally capable of pushing the frontier of AI-enabled formal methods for cybersecurity. Teams should combine deep expertise in security and systems engineering (including the technical judgement and discipline to choose the right threat model, proof boundary, and make precise assurance claims), as well as formal methods and AI-assisted engineering (e.g. developing scalable yet trustworthy workflows for AI-enabled spec validation, code and proof generation). They should also be able to iterate quickly and flexibly, working closely with ARIA as targets, methods, and opportunities evolve over the life of the programme. We welcome proposals from single organisations as well as teams spanning multiple institutions, including across academia, industry, non-profits.

Strong teams will usually show some combination of:

- prior delivery of substantial systems or infrastructure software where correctness, performance, or reliability mattered;
- strong security judgement about threat models, attack surfaces, and what claims would actually matter in deployment;
- genuine depth in formal methods, proof engineering, semantics, verification, compilers, runtimes, or closely adjacent areas;

- concrete evidence that AI is a load-bearing part of the workflow for specification, implementation, proof, testing, or red teaming;
- the ability to work in short, evidence-led cycles, revise targets, and collaborate closely with ARIA and the central red-teaming partner.

We are also interested in teams that are motivated and well positioned to get successful outputs into real-world use — whether through direct deployment, upstreaming into existing stacks, partnership with likely operators, licensing, or, where appropriate, the creation of a spinout, etc.

What we are *not* looking for - Blue Teams (Track 1)

We are not interested in applications for:

- projects without meaningful *formal* assurances;
- projects where the assurance claim is vague, inflated, or poorly scoped;
- projects whose defensive value is marginal as opposed to transformative;
- projects with no plausible deployment path, or which lack the ambition to move beyond research prototypes to real-world deployments;
- pure tooling projects without an end-to-end security component as the main output.

This is specifically a call to test and accelerate the potential of *AI-enabled formal methods*. Proposals that do not make AI central to how they will achieve scale, speed, or assurance are unlikely to be competitive.

How we judge success - Blue Teams (Track 1)

For individual projects, we will judge success by the overall strength of the assurance case and the cyberresilience value of the resulting components. In particular, we will look at:

- whether the target secures a high-leverage system, trust boundary, choke-point, or defensive primitive;
- the precision and completeness of the specification;
- the realism of the threat model;
- the strength, scope, and machine-checkability of the formal claim;
- the clarity of the proof boundary, remaining trusted computing base, and residual assumptions;
- how well the system performs under independent red teaming and security review;
- whether AI materially expands what the team can specify, build, prove, or attack within the programme's time and budget;
- the practicality, performance, maintainability, and deployability of the resulting system; and

- the credibility of the path to adoption.

Technical success alone is not enough to improve cyber resilience. The components built in this programme need to be adopted in settings where they materially reduce risk. We will therefore assess not only whether teams can build and verify strong systems, but also whether those systems can credibly be integrated, upstreamed, or deployed in practice, and whether the team has a realistic plan for making that happen.

What are we looking for - Red Teaming Partner (Track 2)

We are looking for one partner to serve as the programme's central adversarial evaluation capability. The red-teaming partner's core responsibility is to stress-test the full assurance case around each blue-team system on every sprint cycle, and report what they find. They will have view access to each blue team's work throughout the build phase, using it to understand the system and prepare attacks, before executing focused exercises in the evaluation phase. Their reports should be rigorous enough that the blue team can act on them, and clear enough to serve as a credible external record of how each system held up under attack. The red-teaming partner will also advise the ARIA programme team on cross-team patterns, evaluation design, and end-of-programme assessments.

- **Sprint cycle.** Blue teams operate on 8-week sprint cycles: 6 weeks build-and-verify, 1 week red-teaming, 1 week review-and-retarget. The Red Team partner embeds into this cadence — observing during the build phase, executing focused exercises during the red-teaming week, and delivering reports that feed into the review phase.
- **Concurrent engagements.** The partner will be working with 3–5 blue teams simultaneously, each on different components, proof stacks, and threat models. The role demands evaluative independence and adversarial sharpness across all concurrent engagements, plus the operational discipline to manage information flow cleanly between them.
- **Duration.** Engagement runs from project start (expected August 31 2026) through the end of November 2027. The partner will join the cadence as blue teams come online.
- **Indicative budget:** £2–3m total, covering the full engagement period.
- **Contract type:** Service contract with ARIA, with deliverables, acceptance criteria, and payment milestones tied to sprint cycles. This differs from the R&D grant agreements used for blue teams.

What We Are *not* Looking For - Red Teaming Partner (Track 2)

The ideal partner combines deep technical attack capability with strong judgement about assurance claims, and intellectual honesty. They should be comfortable operating across multiple proof stacks and system types, using AI effectively in their own workflow, and maintaining evaluative independence and adversarial sharpness across concurrent engagements with multiple blue teams.

We are *not* looking for:

- penetration-testing firms without genuine engagement with formal methods;
- teams who are not using AI to augment their workflows and adversarial testing;
- proposals that frame the role as primarily an embedded advisory function rather than rigorous adversarial evaluation;
- teams without credible plans for managing concurrent engagements while maintaining independence.

SECTION 4: Application and eligibility

Application process

The guidance on what to include in your proposal, and the criteria against which your proposal will be assessed, differs depending on the track you apply for:

Track 1, see here for guidance on what to include in your proposal and how proposals will be assessed under the Track 1 selection criteria. Successful Track 1 applicants will receive R&D funding agreements.

Track 2, see here for guidance on what to include in your proposal and how proposals will be assessed under the Track 1 selection criteria. Successful Track 2 applicants will be contracted on commercial services terms.

Eligibility, timelines and how to apply guidance below applies to both track 1 and 2.

Eligibility

We welcome applications from across the R&D ecosystem, including startups, industry and academia.

We typically require the majority of the project work to be conducted in the UK (i.e. >50% of project costs and personnel time). Exceptions are possible where a strong case can be made for

the benefit to the UK, and where the team is uniquely qualified to deliver the intended project. This case should be included as part of the funding application.

Red and Blue Teams must be independent. Applicants are free to apply to both tracks, but the same organisation or team will not be appointed to serve in both capacities.

SECTION 5: Timelines

We intend to select the teams as soon as possible within the initial timeline set out below. However, if we do not identify suitable teams, we will continue to accept and assess proposals on a rolling basis using the application process detailed above until we find the right teams.

Applications open	**20 May 2026**
Full proposal submission deadline	**1 July 2026 (14:00 BST)**
Full proposal review	**3 July 2026 - 22 July 2026**
If you are shortlisted following full proposal review, you may be invited to meet with the Programme Directors to discuss any critical questions/concerns prior to final selection — this discussion can happen virtually. This is likely to be the 21st and 22nd of July.	
Successful/Unsuccessful applicants notified	**31st July 2026**

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 on the 3rd and 4th of August.

We expect contract/grant signature to be no later than 4 weeks from successful/ unsuccessful notifications. During this period the following activity will take place:

- Due diligence will be carried out
- The PD and the applicant will discuss, negotiate and agree the project activities, milestones and budget details
- Agreement to the set Terms and Conditions of the Grant/Contract. ARIA will not negotiate the terms of the funding.

SECTION 6: 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.

For the specific selection criteria, please see here for Track 1 and here for Track 2.

SECTION 7: 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, please use the chat function on the funding call page for the quickest response. It can guide you to the right information or connect you with the ARIA team if needed. Clarification that cannot be answered by the AI chat function should be submitted no later than 26th June 2026. 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 and create your account before the application deadline.

If you are disabled or have a long-term health condition, we can offer support to help you engage with ARIA, navigate our funding application process, or carry out your project, you can find more information here.

APPLY HERE

APPENDIX A: Example Targets

Example 1: Verified WireGuard Control Plane

- **What would be built.** A verified control plane for a WireGuard-based overlay network. Given device identities, access policy, and current network state, it computes and distributes exactly the peer, key, and routing configuration needed to realise the authorised connectivity graph. The control plane handles the full device lifecycle: enrolment, approval, key rotation, policy change, and revocation.
- **Threat model.** The underlying WireGuard data plane provides encrypted, authenticated transport. The attacker targets the layer above: observing or tampering with control-plane traffic, attempting to enrol an unauthorised device, or compromising one legitimate node and trying lateral movement or stale access after revocation, or compromising one legitimate node and using stale configuration state to pivot laterally.
- **Why it matters.** WireGuard-based overlays are increasingly used as the effective access perimeter for cloud, developer, and critical infrastructure environments. A verified control plane would make a precise zero-trust claim: only authorised devices learn about, and can be configured to reach, the peers and assets that policy permits. That is a concrete zero-trust gain: shifting trust away from network location and toward explicit identity- and policy-based authorisation.
- **Specification sketch.** For any enrolled devices, identities, and access policy, the control plane computes per-device peer visibility and configuration state such that only authenticated and authorised devices can learn about and reach the relevant peers and resources. Registration, approval, key rotation, and revocation preserve that invariant over time.
- **Verification target.** The core proof target is the control-plane state machine: device registration, identity binding, peer authorisation and visibility, policy compilation, key distribution, rotation, and revocation; with both the static authorisation invariant and its preservation under all lifecycle transitions. The WireGuard tunnel protocol itself is assumed, ideally reusing existing verified WireGuard implementations rather than reproving the data plane. Identity issuance is treated as a trusted input.
- **Prior art.** Owl has been used to formally verify core security properties of the WireGuard handshake, and Tamarin has been used to verify the protocol level. None of this work addresses the control plane or end-to-end overlay system, which is where most operational risk in WireGuard-based zero-trust deployments actually lives. Tailscale and Cloudflare's WARP are widely deployed control-plane designs but are not formally verified.

Example 2: Verified SSH Boundary

- **What would be built.** At minimum, a verified SSH boundary: the packet framing and parsing, the RFC 4253 transport layer, and the RFC 4252 user-authentication state machine, including the concurrency and signal-handling discipline of the daemon. A more ambitious version would include the channel-opening and multiplexing machinery (RFC 4254). The stretch goal is a full protocol-compatible, drop-in SSH server.
- **Threat model.** A remote, unauthenticated attacker sends arbitrary bytes to the SSH port, attempting to trigger undefined behaviour, violate the protocol state machine, or gain unauthorised system access.
- **Why it matters.** SSH is the skeleton key to virtually all server infrastructure. A single bug in an SSH daemon can give root access to millions of machines (see, for example, CVE-2024-6387 ("regreSSHion")). A verified SSH boundary would eliminate the dominant classes of remote pre-auth code execution and pre-auth state confusion for a protocol that is used essentially everywhere.
- **Specification sketch.** For any network input and any interleaving of concurrent events including signal delivery, the implementation either parses and processes input according to the SSH specification or rejects it; malformed inputs cannot trigger memory unsafety; the transport and authentication state machines transition only along paths permitted by RFC 4253 and RFC 4252; and an authenticated session is exposed to the host only when a valid authentication transcript has been completed for an authorised principal.
- **Verification target.** The network-facing SSHv2 boundary: packet framing/parsing, the RFC 4253 transport layer, the RFC 4252 user-authentication state machine, and the concurrency/signal-handling discipline that surrounds them. The proof goal is memory safety, state-machine correctness under adversarial input and the absence of unintended transitions through concurrent or asynchronous control flow. Correctness of underlying cryptographic primitives (e.g. HACL*, EverCrypt) and the OS kernel are assumed.
- **Prior art.** Vest provides verified high-performance binary parsers; HACL* and EverCrypt supply verified cryptographic primitives; Owl demonstrates that verified secure-channel protocol designs can be compiled into interoperable implementations. Earlier work by Poll and Schubert formally specified SSH state machines and found that even mature implementations could violate parts of the standard. No existing system covers the full SSH boundary (parser, both state machines, and concurrency discipline) under a single machine-checked proof, which is the open frontier.

Example 3: Verified TLS 1.3 Server

- **What would be built.** A protocol-compatible, drop-in TLS 1.3 server with a machine-checked proof of wire-format parsing, the handshake and record-layer state machines, downgrade resistance, and faithful exposure of authentication outcomes to the calling application. This includes both server and client roles (the latter required for mTLS), a curated cipher-suite set, session resumption, 0-RTT, and post-handshake authentication.
- **Threat model.** A network attacker with full control of the wire attempting to violate session confidentiality or integrity, impersonate a server, force weakened parameters via downgrade or cross-protocol attacks, trigger memory unsafety through malformed messages, or cause the server to expose an authenticated session to the application when authentication has not in fact succeeded.
- **Why it matters.** TLS is the universal trust boundary of the internet: every HTTPS connection, mTLS deployment, service-mesh identity layer, and code-signing verification path depends on it. A single bug in a TLS stack can compromise traffic across millions of services (see, for example, Heartbleed; FREAK; the Triple Handshake attack). A verified TLS 1.3 server would eliminate the dominant classes of state-machine confusion, downgrade, and pre-auth memory bugs, for the protocol that sits at the outer envelope of essentially all secure communication.
- **Specification sketch.** For any sequence of network inputs, the implementation either processes input according to RFC 8446 or rejects it; malformed inputs cannot trigger memory unsafety; the handshake and record-layer state machines transition only along paths permitted by RFC 8446; the negotiated parameters are the strongest mutually supported by the configured policy; and an authenticated session is exposed to the application only when a valid handshake transcript has been completed with an authenticated peer under those parameters.
- **Verification target.** Wire-format parsers, handshake state machine, record layer, downgrade-resistance logic, and the application-facing API contract — covering both cryptographic soundness of the protocol (under standard assumptions on the primitives) and faithfulness of the implementation to it. Cryptographic primitives are assumed correct, reusing HACL*/EverCrypt. Certificate-path validation is an assumed oracle, or a separately verified component. OS, hardware, side channels, and TCP stack are out of scope.
- **Prior art.** Project Everest and miTLS demonstrated end-to-end verification of a TLS 1.3 reference implementation in F*, covering the record layer and a substantial fragment of the handshake. HACL* and EverCrypt provide verified primitives; EverParse provides verified binary parsers; Rustls is a memory-safe but not formally verified production stack. The open frontier is bringing miTLS-style assurance to production-grade performance and

feature completeness, and shipping it as a drop-in replacement at meaningful trust boundaries.

Example 4: Verified Output Mediation for AI Inference

- **What would be built.** A verified output path in which every token stream, tool call, or actuation request must pass through a specified chain of checks before it can leave the inference system.
- **Threat model.** The attacker (which may be the AI system itself, via prompt injection or adversarial input) tries to smuggle unsafe outputs past filters, bypass monitoring, or invoke tools and actions without approval. The threat is not that the model "thinks wrong," but that there is no hard barrier between model output and real-world effect.
- **Why it matters.** Many real AI failures happen not at inference time but at the output boundary, when unfiltered model outputs reach tools, APIs, or users without mediation. Verified output mediation chain would guarantee that every externally visible output must traverse the approved check sequence, and no path exists to bypass it. This check is structural, and not (on its own) semantic.
- **Specification sketch.** Complete mediation and non-bypass. Every externally visible output — token streams, tool invocations, API calls, actuation requests — must traverse the declared monitor chain in order. Only outputs that pass all checks are released. No output path exists that circumvents the chain.
- **Verification target.** The proof target is the output-mediation architecture: the routing of all output paths through the monitor chain, the non-existence of bypass paths, and the correct sequencing of checks. The correctness of individual filter logic (e.g. content classifiers) is outside the proof boundary — the claim is structural (every output is checked) rather than semantic (every check is correct).
- **Prior art.** Reference monitor concepts from operating systems security (complete mediation, tamperproofness, verifiability) provide the theoretical foundation. Existing AI guardrail systems (e.g. NeMo Guardrails, Guardrails AI) implement monitor chains but without formal verification of the non-bypass property.

Example 5: Verified Proof-Checking Kernel

- **What would be built.** A formally verified proof-checking kernel with a small trusted base, suitable for checking proofs produced by tactics, automation, and AI systems.
- **Threat model.** An adversary — including an AI proof-generation system — submitting crafted "proofs" that the kernel erroneously accepts. The kernel must be robust to adversarial input. This is the adversarial robustness to AI proof generation requirement:

as AI's heuristic-based proof search (and in particular RL training) create strong adversarial pressure to find loopholes in the checker.

- **Why it matters.** If we are building a world where AI agents produce formally verified artefacts at scale, the proof checker is the single point of trust. A bug in the proof checker undermines everything built on top of it.
- **Specification sketch.** The proof kernel is sound: if it accepts a proof, the corresponding theorem genuinely follows from the axioms. More precisely, the kernel implements a well-defined type theory, and the implementation is proved faithful to that theory.
- **Verification target.** The kernel alone. The theorem to prove is soundness of the implementation with respect to a precisely defined logic. The verification target is the proof-checking kernel only — not the tactic framework, automation, or AI systems that produce proofs for it to check.
- **Prior art.** Existing proof assistants (Lean, Coq, Isabelle) have small kernels by design. Candle (Abrahamsson, Kumar, Myreen, Owens) provides a verified HOL Light kernel compiled to machine code via CakeML. MetaCoq has formalised substantial parts of Coq's metatheory and a verified type-checker for a large fragment. The open frontier is extending this assurance to richer dependently-typed logics (Lean 4, Coq) at production performance, and to kernels designed under adversarial pressure from AI proof search.

Example 6: Verified Capability-Mediated Runtime for AI Agents

- **What would be built.** A verified object-capability runtime for AI agents, in which each agent runs in its own compartment and holds only explicit, unforgeable capability tokens for the tools, files, network endpoints, and UI surfaces it may use. A stronger extension would add an authority-safe programming layer for agent-generated code.
- **Threat model.** The agent itself is assumed compromisable — through prompt injection, jailbreak, or adversarial input — and may try to exfiltrate data, escalate privileges, pivot to other systems, or misuse tools. The goal is not to make the agent benign, but to ensure that a compromised agent can act only with the authority it already holds.
- **Why it matters.** As AI agents proliferate, the blast radius of a compromised agent becomes a systemic risk. Verified containment turns this from a trust problem into a confinement problem: the question is not whether the agent behaves well, but whether its authority is strictly bounded. Object-capability security is a principled foundation for this because least privilege, delegation, and attenuation are built into the model itself.
- **Specification sketch.** The system satisfies the object-capability discipline: no ambient authority; capabilities are unforgeable, delegable, and attenuable; the capability graph is

the complete authority graph. An agent can invoke only what it holds capabilities for, and any derived code or sub-agent exercises only explicitly delegated authority.

- **Verification target.** The OCap reference monitor and capability runtime: no capability implies no access; attenuation and delegation preserve least privilege; compartment boundaries enforce confinement.
- **Prior art.** seL4 is a capability-based kernel with strong formal assurance: capDL provides a declarative way to specify capability distributions and reason about future access possibilities: Microkit provides a practical framework for building statically structured seL4 systems. The main open gap is the agent-specific layer above that substrate: a small verified broker architecture for tools, files, network, and UI surfaces, composing kernel-level and runtime-level OCap guarantees end-to-end.

Annex A - Safeguarded AI: Cybersecurity — Track 1 Blue Team Applicants

Funding terms

Successful **Track 1 Blue team proposals** will be funded under **ARIA's R&D grants/contracts** (more information on the terms can be found here).

Application Process and Criteria

The application process for Track 1 Blue Teams applicants consists of a single stage which requires you to submit a detailed proposal of up to 5 pages.

All proposals should include:

- **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 work, their expertise, and why you/the team are motivated to solve the problem.
- **Answers to administrative questions** to help ensure we are responsibly funding R&D. Questions relate to budgets, IP, potential COIs, etc.
    - *This information is provided directly via the application portal and does not count towards the page limit.*

Application details — Track 1

Project & Technical information

Each proposal should make clear:

1. What component you propose to target *initially*, including:
    - what the component is and why securing it matters;
    - the threat model being assumed;
    - the exact security properties to be specified and proved;
    - the execution semantics or proof level against which the claim is stated (e.g. source, IR, bytecode, Wasm, assembly), and what this implies for the meaning and limits of the claim;
    - the trusted base, key dependencies outside the proof boundary, and the residual assumptions that would remain;
    - the intended deployment context of the component, and any important integration, performance, operational, or maintenance constraints that shape the target; and

    - any relevant prior art, and how your proposed approach differs from or improves on it.
2. What more ambitious follow-on targets you would be interested in pursuing if the initial work succeeds, and why those would be valuable next steps. *Subsequent targets will be selected sprint-by-sprint with the programme team, in light of progress and the evolving capability frontier. Applicants should show a credible direction of travel, not a fixed multi-cycle plan. These don't need to be spelled out at the same level of detail as the initial target — we're looking for a sense of the kinds of targets you would find valuable, and why.*
3. How you plan to execute and scale AI-enabled formal methods across the full workflow, including:
    - how you would use AI to accelerate specification, implementation, proof, testing, or related parts of the workflow;
    - how you would develop and validate the specification, especially where the hard part is ensuring that the claim being proved is the claim that matters in practice;
    - how you would preserve assurance quality as those workflows scale; and
    - the main technical bottlenecks or tractability risks, and why you believe the proposed work is feasible within the programme.

Information about the team

Use this section to help us judge whether your team can execute the proposed work. Please tell us:

- who the core contributors are and their expected availability;
- how responsibilities are divided across the team (e.g. system design, spec definition and validation, AI workflow design, proof and validation, deployment/adoption);
- the most relevant things you have previously built, verified, or deployed;
- how you approach scaling AI-enable formal methods for real-world cybersecurity, and evidence of the team's ability to use AI productively in the proposed technical workflow;
- why this particular team is well matched to this particular target; and
- how the team is positioned to drive real-world adoption, including any relevant prior experience, relationships, and plans for commercialisation or open-sourcing.

We are primarily looking for evidence of execution, technical depth, security mindset, and ambition. We care more about demonstrated capability, ownership, and learning speed than about conventional signals prestige.

Administrative questions:

This section includes information about the budget, intellectual property that you intend to rely

on, any perceived conflicts of interest and for non-UK applicants how the proposed project may benefit the UK.

In completing your application you must also provide answers to the following questions. Answers to these questions are not included in the page cap. You should complete these questions in the application portal so there is no need to format these specifically.

Application	Guidance
How much funding do you need?	*Please provide a cost breakdown by completing the spreadsheet here. In your proposal you may submit your budget using yearly, quarterly, or monthly phasing. Prior to completing this template you should review ARIA's Eligible cost guidance here. If your proposal is successful, prior to contract signature when the scope of work has been agreed, you will be required to provide a monthly cost breakdown.*
Are you proposing to contribute funding?	*If you or your organisation are proposing to contribute funding to the project please let us know how much funding you plan to contribute, who is contributing the funding, is the funding already secured and any other relevant details.* *ARIA will fund 100% of project costs and contribution of funding is not essential however, we welcome proposals that contribute funding in cases when such funding will strengthen the potential success. In these cases, this funding contribution will be considered as part of the overall strength of the project proposal.*
Does your proposal depend on background IP (pre existing)?	*If Yes, give us an Indication of: What background IP is required, Whether you currently have rights to that IP.*
Have you already secured funding for a similar project or are you currently in the process of seeking support from other funding sources for the same project?	*If yes, tell us more about the funding you already have or are applying for.*

Any other factors or restrictions that might impact your freedom to operate and deliver the project?	*Please provide a detailed description of any perceived conflicts of interest with the programme director, import/export or security restrictions that you are aware of*
Are you proposing to perform the majority of the proposed project outside of the UK?	*Our primary focus is on funding those who are based in the UK. For the vast majority of applicants, we therefore require the majority of the project work to be conducted in the UK (i.e. >50% of project costs and personnel time).* *However, we can award funding to applicants whose projects will primarily take place outside of the UK, if we believe it can boost the net impact of a programme.* *In these instances, you must outline any proposed plans or commitments in the UK that will contribute to the programme within the project's duration (note the maximum project duration is 3 years).* *Please provide a detailed description of any proposed plans (including a timeline) or commitments).*
Has a suitably authorised member of your Organisation approved the submission of this proposal?	*In the application portal, please select the option that best describes your situation and provide details where required.*
Have you read and understood our funding terms?	*Our goal is to ensure your research can get going quickly, so we want to ensure a fast negotiation and award process. We aim to have agreements signed within 6 weeks, which we recognise can be much faster than standard at some organisations. Before proceeding, please confirm that you have read and understand our funding terms. If you are unsure which terms apply to you, you can find more guidance here.*
Additional questions about you/your organisation that can be found in the application portal.	

Evaluation Criteria

In conducting a full review of the Track 1 Blue Team Applicants proposal we'll consider the following criteria:

1) **Worth Shooting For** – The proposal demonstrates strong judgement about what counts as a high-leverage cybersecurity target: components where machine-checked correctness yields outsized resilience gains, sitting at trust boundaries, choke-points, or defensive primitives whose verification durably retires classes of exploit. We are not looking for teams to define a detailed multi-cycle plan, but we are looking for credible signs of ambition, taste, and technical judgement to keep selecting such targets sprint-by-sprint as the programme's capabilities and the AI-enabled formal methods frontier evolve.

2) **Well defined** – The proposed initial target is precisely scoped: a clear security specification, realistic threat model, well-defined proof boundary, and clean account of the trusted computing base and residual assumptions. The team's broader direction of travel is credible, and the proposed composition, technical workflow, costs, and timelines are realistic for the work described.

3) **Differentiated** – The proposal makes AI a load-bearing part of the workflow in ways that would not be tractable through human effort alone, pushing what AI-enabled formal methods can deliver rather than applying them in conventional ways.

4) **Responsible** – The proposal identifies major ethical, legal, or regulatory risks and presents clearly defined and feasible mitigation plans, as applicable. The team demonstrates genuine commitment to intellectual honesty in scoping its assurance claims and to presenting specs, proofs, and TCB assumptions in ways that external stakeholders can understand and validate.

5) **Intrinsic motivation and team strength** – Given that targets will evolve sprint-by-sprint, team quality and ambition are the central assets we are investing in. We are looking for teams with deep expertise across security, systems engineering, formal methods, and AI-assisted workflows, the technical taste to identify and pursue ambitious targets, and the ability to operate in short, evidence-led cycles. Strong proposals will show genuine motivation to work on the problem and a credible orientation toward real-world adoption, be that through direct deployment, upstreaming, partnership, or spin-out.

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.

Annex B - Safeguarded AI: Cybersecurity — Track 2 Red Team Applicants

Background

As discussed, ARIA's Safeguarded AI programme is funding a portfolio of 'blue teams' to build and formally security-critical software components whose key security properties are backed by machine-checked proofs. As part of this effort, we are looking for a **red teaming partner** who will conduct regular adversarial evaluations across all funded blue-team projects. This document discusses the scope of work and application process and requirements for this Red Team partner. We strongly recommend potential applicants to read the main solicitation document alongside this document.

Successful **Track 2 red team proposals** will be funded under **ARIA's commercial services terms** (The full terms and conditions, which may be refined during negotiations, can be viewed here). This reflects the nature of the work being carried out under Track 2, which is expected to be delivered on a more service-based basis, rather than as research and development.

Application Process and Criteria

The application process for Track 2 Red Team Applicants consists of a single stage which requires you to submit a detailed proposal of up to 3 pages.

All proposals should include:

- **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 work, their expertise, and why you/the team are motivated to solve the problem.
- **Answers to administrative questions** to help ensure we are responsibly funding R&D. Questions relate to budgets, IP, potential COIs, etc.
    - *This information is provided directly via the application portal and does not count towards the page limit.*

Project & Technical information

If you are applying to serve as the programme's red-team / adversarial evaluation partner, your proposal should address:

- your adversarial-evaluation methodology, including how you would assess formally verified systems where conventional implementation bugs may be less central but specification gaps, trusted-computing-base assumptions, and integration and deployment failures remain critical;

- how you envision operating within the sprint cycle, including how you would use the build phase to understand each system, prepare and execute a thorough red teaming exercise, and communicate your findings in a way that is clear, rigorous and useful;
- what will allow you to operate effectively across 3-5 concurrent blue teams working on different system components and using different proof stacks;
- how you use AI to scale vulnerability discovery, exploit development, analysis, and evaluation design while preserving rigour and judgement; and
- how you would advise the ARIA programme team on cross-team patterns, evaluation design, and end-of-programme assessments, distinct from your per-team evaluation work.

Information about the team

Use this section to help us judge whether your team can execute the proposed work. Please tell us:

Who the core contributors are, their expected availability, and the roles they would play;

- how responsibilities are divided across the team;
- the most relevant things you have previously red-teamed, evaluated, attacked, or otherwise stress-tested, and any other evidence that you can deliver the kind of adversarial evaluation proposed here;
- what evidence you have that the team can use AI productively in the proposed technical workflow; and
- any external partners, domain access, technical infrastructure, or other assets that would materially strengthen the realism, quality, or credibility of your evaluation work.

We are primarily looking for evidence of execution, technical depth, security mindset, and the ability to operate effectively in a demanding, high-ambiguity programme environment. We care more about demonstrated capability, ownership, and learning speed than about conventional signals prestige.

Administrative questions:

This section includes information about the budget, intellectual property that you intend to rely on, any perceived conflicts of interest and for non-UK applicants how the proposed project may benefit the UK.

In completing your application you must also provide answers to the following questions. Answers to these questions are not included in the page cap. You should complete these questions in the application portal so there is no need to format these specifically.

Application	Guidance
How much funding do you need?	*Please provide a cost breakdown by completing the spreadsheet here. In your proposal you may submit your budget using yearly, quarterly, or monthly phasing.* *Prior to completing this template you should review ARIA's Eligible cost guidance here.* *If your proposal is successful, prior to contract signature when the scope of work has been agreed, you will be required to provide a monthly cost breakdown.*
Are you proposing to contribute funding?	*If you or your organisation are proposing to contribute funding to the project please let us know how much funding you plan to contribute, who is contributing the funding, is the funding already secured and any other relevant details.* *ARIA will fund 100% of project costs and contribution of funding is not essential however, we welcome proposals that contribute funding in cases when such funding will strengthen the potential success. In these cases, this funding contribution will be considered as part of the overall strength of the project proposal.*
Does your proposal depend on background IP (pre existing)?	*If Yes, give us an Indication of: What background IP is required, Whether you currently have rights to that IP.*
Have you already secured funding for a similar project or are you currently in the process of seeking support from other funding sources for the same project?	*If yes, tell us more about the funding you already have or are applying for.*
Any other factors or restrictions that might impact your freedom to operate and deliver the project?	*Please provide a detailed description of any perceived conflicts of interest with the programme director, import/export or security restrictions that you are aware of*
Are you proposing to perform the majority of the proposed project outside of the UK?	*Our primary focus is on funding those who are based in the UK. For the vast majority of applicants, we therefore require the majority of the project work to be conducted in the UK (i.e. >50% of project costs and personnel time).*

	However, we can award funding to applicants whose projects will primarily take place outside of the UK, if we believe it can boost the net impact of a programme. *In these instances, you must outline any proposed plans or commitments in the UK that will contribute to the programme within the project's duration (note the maximum project duration is 3 years).* *Please provide a detailed description of any proposed plans (including a timeline) or commitments).*
Has a suitably authorised member of your Organisation approved the submission of this proposal?	*In the application portal, please select the option that best describes your situation and provide details where required.*
Have you read and understood our funding terms?	*Our goal is to ensure your research can get going quickly, so we want to ensure a fast negotiation and award process. We aim to have agreements signed within 6 weeks, which we recognise can be much faster than standard at some organisations. Before proceeding, please confirm that you have read and understand our funding terms. If you are unsure which terms apply to you, you can find more guidance here.*
Additional questions about you/your organisation that can be found in the application portal.	

Evaluation Criteria

In conducting a full review of the Track 2 Red Team proposal we'll consider the following criteria:

1. **Approach and operational execution** – The proposal sets out a credible methodology for adversarial evaluation of the systems in question, making AI a central part of the workflow to ensure realistic adversarial capabilities and scalability. The proposal also demonstrates a credible plan for executing intense, prep-heavy red-teaming weeks across 3–5 concurrent blue teams every 8 weeks, sustaining quality across the full sprint cadence, and producing reports that are crisp, on-point, and serve both the blue teams and programme teams.
2. **Demonstrated ability to do the work** – The team brings deep technical attack capability and adequate breadth across the range of system types blue teams are likely to

be working on (e.g. network stacks, kernels, compilers, AI infrastructure, cryptographic libraries). We are looking for evidence of past delivery comparable in technical depth and operational intensity to what this engagement requires. We care more about evidence of capability than conventional signals of prestige.

3. **Responsible** – The proposal demonstrates thoughtfulness in identifying risks specific to such adversarial evaluation work and proposes appropriate mitigations where necessary. The team demonstrates intellectual honesty about what evaluation can and cannot establish.
4. **Intrinsic motivation** – The team is genuinely motivated by the goals of this call: making AI-enabled formal methods work for cybersecurity at the level of ambition the programme requires, and bring the ambition and judgement that consistently raises the quality of what gets done.
5. **Commercial terms** – Alignment with ARIA's culture and values – Genuine alignment with the programme's mission, and the orientation to engage as a mission-aligned partner rather than a transactional vendor.