Advanced Research + Invention Agency ARIA

Programme Thesis

Scaling Trust

v1.0

CONTEXT

This document presents the core thesis underpinning a programme that is currently in development at ARIA. We share an early formulation and invite you to provide feedback to help us refine our thinking.

This is not a funding opportunity, but in most cases will lead to one — sign up **here** to learn about any funding opportunities derived or adapted from this programme formulation.

An ARIA programme seeks to unlock a scientific or technical capability that

+ changes the perception of what's possible or valuable
+ has the potential to catalyse massive social and economic returns
+ is unlikely to be achieved without ARIA's intervention.

PROGRAMME THESIS, SIMPLY STATED

Every day we spend time, effort, and resources coordinating with each other – finding the right people or services to work with, negotiating the terms of the interaction, and enforcing the agreements made. These are fundamental costs of modern life we pay, what economists call 'Coasean transaction costs', and AI is collapsing them. Their reduction signals a new era for our society and our economy.

We envision this era as one where AI agents are our faithful representatives, continuously aligning to our preferences, able to go out into both digital and physical worlds and cheaply mobilise, negotiate, and verify on our behalf. As opposed to an all-powerful, all-knowing AI, this is a world of many-agents, each holding our private information, operating with their own objectives, and constantly interacting with institutions, humans and other agents.

This is an exciting vision, one of human flourishing and augmentation via technology, that preserves our plurality and uniqueness.

We believe new security primitives like programmable cryptography and secure hardware represent a unique opportunity to usher in this new world by:

1. Creating a scalable trust infrastructure for agents across digital and physical worlds, thereby increasing the complexity of secure interactions agents can engage in, and the pool of potential parties they can engage in them with;
2. Enabling new forms of secure interactions previously impossible for humans or traditional software, unlocking new valuable markets and societal value.

To this end, we plan on launching three core initiatives. First, we will fund open-source applied tools needed to build this capability robustly, safely and for all of humanity. Second, we will fund fundamental research to build a stronger theoretical foundation for this field, and to uncover new security primitives agents can harness for secure coordination. Third, we will launch a series of challenges open to all to benchmark the tools built and research developed, with multi-million-pound prize funds for the best teams.

It will take many disciplines and stakeholders to make this vision a reality. The road ahead is riddled with risks, but getting it right is transformative. We're calling on all of you – hackers, roboticists, game theorists, cryptographers, AI security engineers, and everything in between – challenge our plans, join the community, and help us Scale Trust.

Figure 1. Comic: Imagining the silent trust infrastructure of tomorrow

PROGRAMME THESIS, EXPLAINED

Why now?

Three trends are converging:

+ AI agents are growing more capable
+ AI is increasingly directly interacting with the physical world
+ Engaging in advanced security protocols with programmable cryptography and trusted hardware is becoming practical

We believe an opportunity lies at their intersection.

AI agents are growing more capable

As more resources (compute, data, engineering hours) get poured into the development of AI, models have kept on improving and are predicted to keep improving. While some wonder if we will plateau – either from running out of data, or finding the limits of the current architectures, we are today in the infancy of really understanding how to use these models. Whether models get more powerful or we learn more about how to scaffold requests and how to get the best from them, we believe progress is going to continue at a pace for some time. This is a trend we need to internalise deeply in our thinking as it is easy to anchor on our current capabilities.

AI is increasingly directly interacting with the physical world (Embodied AI)

Embodied AI refers to the specific intelligence that powers a physical system, serving as the "mind" or "brain" that enables it to operate within a complex environment. These systems—which include general-purpose robots, autonomous vehicles (AVs), and smart warehouse facilities—act as the "body." This intelligence uses sensors (like cameras or LiDAR) to perceive the world and actuators (like motors or steering) to act within it. This creates a controllable feedback loop, allowing the AI to reason, make decisions, and see its actions directly influence its future sensory input.

With verticals such as AI for science, self-driving cars, intelligent factory robots, and modern warfare drones, the field of Embodied AI has exploded over the last 10 years. As more and more people turn their attention to AI, many see Embodied AI as the next and final frontier. As the hundreds of millions of venture funding poured into this field gets turned into work, autonomous digital systems will have more touchpoints into the physical world.

Engaging in advanced security protocols with programmable cryptography and trusted hardware is becoming practical

Programmable cryptography and secure hardware are two families of techniques to achieve new kinds of secure interactions between parties – computations on encrypted data, proofs of computations, cryptographic commitments to specific actions – that can open up new markets.

Programmable cryptography refers to advanced cryptographic primitives that allow for general-purpose computation on encrypted or private data. Instead of traditional cryptography, which secures data *at rest* (like encryption), programmable cryptography secures data *in use*. While not practical yet for many applications, programmable cryptography is quickly improving and being accelerated at both software and hardware levels. It has seen tremendous progress and investment over the last ten years, fuelled by use-cases in the cryptocurrency industry and adoption by large tech players (eg. Tiktok, Google).

Similarly, computer hardware can be designed to safeguard computations, preventing adversaries from influencing the logic or learning confidential information. When embedded within such secure hardware, a hidden key enables attestation—proof that genuine hardware is loaded with specified software. By turning to secure silicon, one can recreate similar trust setups to programmable cryptography by enforcing commitments or verifying computations, as well as unlocking new possibilities like binding digital actions to physical devices, enforcing policy compliance, or securing AI agents and robots operating

in the real world. As opposed to programmable cryptography, secure hardware is already practical for similar functionalities, but has weaker security guarantees. Several groups are actively working on improving those guarantees (eg. Ethereum Foundation, Trustless TEE Initiative) and improving its performance (eg. Nvidia, Apple, OpenAI), fuelled by demand from AI and cryptocurrency industries.

By looking at these two techniques in unison, we get to pick and choose: for lower-stake interactions i.e., ones representing little economic value or ones in the context of defense in depth, secure hardware can be used and experimented with today, bypassing the current impracticality of programmable cryptography. For higher-stakes applications that will either be lightweight (and therefore low overhead) or can tolerate high overhead, programmable cryptography can be used today.

Opportunity and risk at the intersection

There already exist exciting opportunities at the intersection of these trends, for instance secure coordination of self-driving cars or smart power grids. Likewise, the absence of a robust trust infrastructure already exposes us to risks, from catastrophic failures to simply stifling the transformative potential of these technologies.

As each trend advances – more capable AI, deeper integration with the physical world, and stronger security primitives – new secure interactions will emerge. At the same time, the urgency for the technologies this programme aims to develop will grow. Each new milestone in these trends will extend the impact of our efforts, and the leverage of the tools, research and capabilities we will fund.

How we think about the problem

We want to build a future where AI agents are our faithful representatives, continuously aligning to our preferences, able to go out into both digital and physical worlds and cheaply mobilise, negotiate, and verify on our behalf.

At a high level, there are three things we need to get to this future:

1) **Alignment**: agents need to be synchronised with our preferences and utility functions in order to be faithful representatives.
2) **Intelligence**: agents need to be capable enough for the task they are given.
3) **Coordination**: agents need to be able to coordinate with others despite competing objectives, information asymmetry and adversarial environments.

Intelligence and Alignment are significant undertakings the industry is focused on. We are focused on Coordination. We believe we can decouple Coordination from Intelligence and Alignment, to start making headway even with 'stupid' and 'misaligned' cyber-physical agents.

Within coordination, we're specifically interested in using security primitives like advanced cryptography and secure hardware as tools to enable (1) a scalable and machine-readable foundation for trust, (2) new kinds of secure interactions impossible to achieve with humans or traditional software.

Breaking down the problem into its component parts

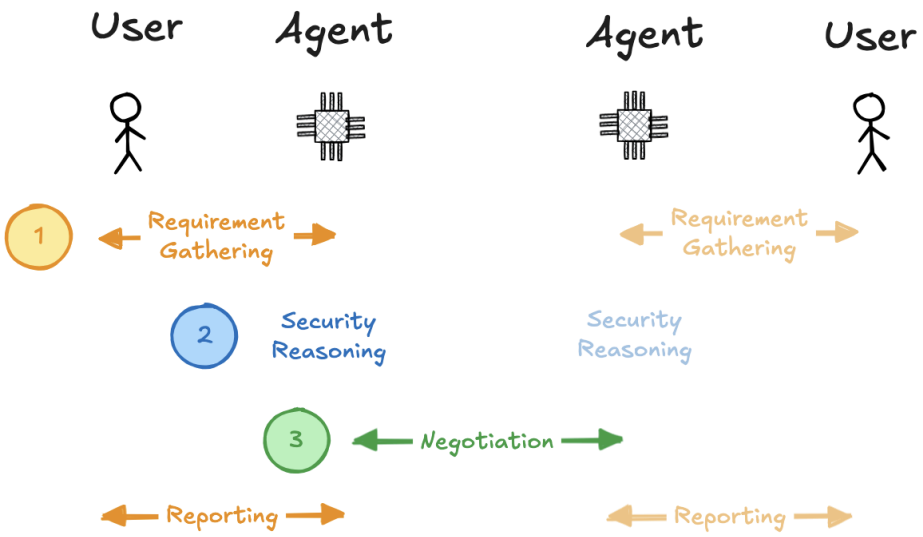

Fig 2: High-level core components

1. **User <> Agent**

   Capability: *Requirements gathering*
   An agent must be able to understand the security policies, the incentives of their user and the constraints of the environment they are operating.

   Can an agent gather security and privacy policies from the user? Can they help the user discover their policies, fill underspecifications and resolve inconsistencies? Can

they do so with minimal communication? Can they encode them in a formal spec? What is the user experience for doing this? How do we avoid loss of control from goal misspecification or goal misgeneralisation?

Capability: Reporting
An agent must be able to report back to their user the outcome of their interactions in a way that can be trusted, inspected and verified.

Can an agent prove the chain of their interactions? Can they report back and explain to their users their reasoning?

Note: While finding efficient ways for requirements gathering and reporting (e.g., having compilers to turn user requests into formal security constraints) is definitely in scope, we do not anticipate focusing our funding efforts on the broader problem of alignment. This is because others are doing great work on these problems, including the ARIA Safeguarded AI programme who we plan on closely collaborating with.

2. Security Reasoning

Capability: Security Reasoning
Once an agent has gathered requirements, it must be able to independently reason about security, evaluate options and execute.

Can they orchestrate a list of whitelisted operations? Can they identify secure available libraries? Can they come up with new protocols? Do security reasoning AI models require theorem provers? Can AI agents reason about physical security?

We've mapped this onto a rough spectrum of levels:

- **Level 1: Security Assistant** – Co-pilot like, support tool that augments a human operator when making security decisions, finds useful information and makes informed suggestions.
- **Level 2: Security Orchestrator** – Understands security goals, makes a plan of tasks to be done. Knows how to operate on a plan by tapping into external tools (e.g. whitelisted libraries) and knows how to reason about security for protocols that are already known. It does not design protocols.
- **Level 3: Security Engineer** – possesses a deeper understanding of security than an orchestrator. Can autonomously architect and implement complex security protocols on the fly.
- **Level 4: Security Researcher** – Like a security engineer, but capable of applying abstract security concepts to new domains. Can autonomously

discover novel cryptographic primitives, identify new classes of vulnerabilities, and formulate new cyber-physical security protocols.

Within this spectrum, we believe we are today at Level 1. The two levels we'd like to focus work on are *Level 2: Security Orchestrator* and *Level 4: AI Security Researcher*. We believe achieving Level 2 unlocks a large number of new functionalities and is relatively doable with today's technology. Level 4 on the other hand is harder to achieve and will likely need radically new approaches. We believe we should get started on both now.

3. **Negotiation**

Capability: Negotiation
After coming up with the right security approach, an agent must be able to interact with other parties (such as other agents), evaluate their proposals, independently negotiate and find compromises, potentially tuning their security approach to the other parties' security requirements as well.

To what extent can an agent reason about incentives? What are the interfaces for secure agent-to-agent communication? Can two agents interact with natural language safely or should their communication be constrained? Can they successfully convince others or securely evaluate others actions? How do agents jointly optimise on their utility functions?

The AI Advantage

As aforementioned, agents can engage in new secure interactions that would not be possible for humans or more traditional computer programs. Such interactions open up new market equilibria, new forms of coordination and ultimately new value creation. We've called this the 'AI Advantage'.

This includes:

+ Black-box access – being able to see the input-output behaviour of another agent without having access to its code
+ White-box access – being able to see the code of other agents and simulate their behaviour in interactions
+ Credible commitments – being able to verifiably bind oneself to a future course of action, often using cryptographic or secure hardware-based mechanisms, to ensure a promise is kept (e.g. verifiable memory erasure, change one's goal, reducing one's action space.)

+ Steganographic communication – being able to employ steganographic methods to conceal the true nature of interactions, be it communicative or otherwise, from oversight
+ Generative Cryptography – being able to write on-demand cryptographic protocols

What we expect to fund

We propose to split our funding efforts into three tracks:

1. **Arena** – adversarial testing grounds designed to empirically and scalably test AI systems capabilities in multi-agent coordination across digital and physical worlds.
2. **Tools** – interoperable coordination infrastructure usable by all to level the playing field in the arena, prevent redundancies, and steer innovation toward the most meaningful axes of progress.
3. **Fundamental Research** – flexible funding to create new fields of research, and build a reservoir of new knowledge that future iterations of this of Tools and Arena can draw upon.

Each track stands on its own and can independently result in artefacts that end up making the programme worth running. However, together they create compounding effects and a shared environment conducive to breakthroughs: Fundamental Research will produce the reservoir of new knowledge, Tools are informed by or implementations of the research, and the Arena is the live environment they are tested in.

This also creates an attractive proposition for participants of each track: empiricists get feedback and work with theorists, theorists get accountability for their ideas, and tool builders get direct feedback from both empiricists and theorists.

Track 1 – Arena

The Arena will host challenges open to all where the best teams will be awarded meaningful prizes. The challenges will be adversarial by design, with many including red and blue team dynamics. There are three kinds of challenges we plan on running, we welcome your suggestions!

1.1 – Component-level challenges

Each component-level challenge targets a scoped isolated capability, behaviour or feature. They are meant as ‘simple’ challenges, easy to ideate, host and enter. This allows us to

respond to community signals quickly, to build a critical mass of shared baseline knowledge early on, and to accumulate learnings faster on the best ways to run challenges.

Challenge suggestions:

- User <> Agent challenges
  - Requirements gathering challenges: gather the right user policy with minimal communication
- Security Reasoning challenges
  - Generative Cryptography challenges: e.g., challenges where the AI picks the right protocols for the right problems.
  - Cyber-physical challenges
    - Adversarial Sim2Real gap - Defend against physical attacks designed to exploit the subtle differences between an AI's training simulation and the real world
    - Hardware level prominence despite poisoned world models - Build trusted hardware to ground truth and override compromised AI powered world models
    - Zero-knowledge heterogeneous swarming - Enable a diverse swarm of robots to securely coordinate a mission without revealing their individual positions, capabilities, or observations to each other
    - Unforgeable receipts for robot actions - Create cryptographically secure logs for robot actions during complex multi-agent tasks
    - System-level consensus challenge for complex cyber-physical systems - Ensure a network of physical agents can agree on a single version of reality and a coordinated action, even when their sensors are noisy and some agents are malicious
- Negotiation challenges
  - Succinct negotiations over general utilities: the challenge and test agents are prompted with different utility function based on the same x parameters, and the test agent must interact with the challenge agent and make the best possible deal using <x/10 tokens

1.2 — End-to-end challenges

Building on top of component-level challenges, end-to-end challenges test compositional reasoning, e.g. can agents combine multiple security primitives to coordinate safely under real-world, or even exaggerated, constraints? We want challenges here that are useful for producing learnings, even if they aren't inherently impactful themselves.

An example of such a challenge is Anthropic's Vend, an experiment where an AI agent (Claudius) was put in charge of a physical vending machine, autonomously setting prices,

managing stock, designing and ordering new products on request, and communicating with customers, optimising against its P&L. The experiment has shown both coordination promise and classic failure modes: Claudius could adapt creatively and resist unsafe requests, but also hallucinated accounts, mispriced goods, and over-discounted items. The challenge itself is benign and fun, yet the learnings it has inspired for autonomous economic systems straddling digital and physical worlds have been significant.

Challenge suggestions:

+ A General Game Playing (GGP) style competition, in which agents are submitted before a variety of games (formal rule sets) are revealed, via a combination of hand-generated games and algorithmically generated rule-sets. The agentic systems compete against each other as well as traditional GGP symbolic algorithms.
+ High-traffic zone w/ competing drone swarms, each having a unique crash utilities and time discount
+ A multi-agent extension of GDPval that requires negotiation and includes a penalty for leaking "private" information
+ A social media site digital twin where agents compete to steer the discourse or stop misinformation
+ Cryptography-emergent challenges: security games where the optimal strategy involves engaging in a cryptographic/security protocol. Eg. Each agent holds a set of private strings, they earn a point if they find the intersection of their strings, but lose if any agent learns information beyond that. Other variants could embody principles from Zero-Knowledge proofs (ZK), Fully Homomorphic Encryption (FHE), Multi-Party Computation (MPC).
+ Secure-Hardware-enabled challenges: For example, a negotiation where two agents share private data through hardware that guarantees the information is erased if the deal falls through, allowing them to cooperate safely without fear of later exploitation

1.3 – ARIA Flagship Challenge

The ultimate end-to-end challenge intended as a North Star for the whole field. The ARIA Flagship Challenge will act as a symbolic milestone and a practical benchmark, the first large-scale demonstration that cyber-physical agentic coordination using security primitives can achieve real-world–useful outcomes under real adversarial and physical constraints.

Challenge suggestions (we plan on picking one, others could be scaled to fit in the previous category):

+ A fully autonomous corporation, that is profitable, and is selling a valuable cyber-physical product by coordinating with suppliers, autonomous factories and other entities

+ A peer-to-peer discovery challenge where one autonomous lab must find a new scientific method and securely teach it to an untrusted competitor for replication
+ A learning challenge where agents teach each other in untrusted environments (one agent discovers one thing, teaches it to another agent)
+ An autonomous assembly challenge to build a high-precision product by sourcing components from an untrusted, potentially malicious supply chain and deploying them into a high-stakes complex system
+ A collaborative containment challenge where untrusted 'government' and 'private' agents must find and neutralise a threat in an adversarially-compromised facility
+ An autonomous search-and-rescue challenge where competing robotic teams must collaboratively map a disaster site and triage victims, even if some agents are compromised and sharing false information
+ A collaborative manufacturing challenge where multi-agent robotic arms must self-organise to build an unseen novel object without a shared blueprint, learning the design from keys on the parts themselves
+ A bio-security challenge, autonomous BSL-3 lab handover, agent A teaches agent B a protocol, proves containment.

Track 2 — Tools

Funding shared tools is important to level the playing field in the Arena and ensure competitive energy is steered towards the most meaningful axes of progress.

We propose early tool-building efforts to be focused on standing up two 'model' agents: a Security Orchestrator and Security Researcher as laid out in Section. Each can become baseline tools in the Arena for people to improve and experiment with.

We anticipate funding R&D that explores what it means for an AI system to reason securely, and how such reasoning can be tested, verified, and improved.

We plan to fund answering foundational questions such as:

+ How can we build benchmarks to evaluate the security reasoning capabilities of AI systems at different levels (orchestrator, researcher)?
+ What are the learning strategies for cooperation and security among AI agents?

We also plan to fund different constructions of such reasoning engines, for example:

+ AI-assisted theorem proving approaches
    - How can AI-assisted theorem proving or formal verification accelerate secure protocol design?
    - Can we create formally verifiable cryptographic libraries suitable for agent orchestration?

+ Reinforcement learning approaches
    - How can reinforcement learning be used to teach agents security reasoning?
    - What kinds of datasets — of protocols, papers, interactions, and examples of both good and bad security — could be used to train such models?

Track 3 — Fundamental Research

Progress in the fundamental research track will not only reduce how much we're shooting in the dark over time, it will also provide theoretical grounding to our empirical approach in Track 1 and the tools built in Track 2. As an analogy: digital communication existed before Claude Shannon, but it was his Theory of Information that defined what information is. That formalisation transformed communication from empirical hacks into a rigorous, theory-driven discipline, enabling the long-distance, high-fidelity systems we rely on today.

We aspire to do the same for the fields relevant to our efforts here. The current research agenda we have in mind has three poles.

3.1 — Formal AI Security

Formal security definitions allow researchers to prove whether a system is secure under explicit assumptions, reason about what is possible (via feasibility and impossibility results, hierarchy of assumptions and guarantees), and provide building blocks for more complex protocols.

Although early work is taking place, we believe AI security stands where pre-cryptographic security once was, where we lack foundational definitions for concepts such as intelligence, alignment, robust communication etc. Formal AI Security could be a new discipline that brings the rigour of theoretical computer science to study intelligent systems.

We propose a few topics (for discussion!):

+ Foundational frameworks and definitions
    - Formal definitions, feasibility results, impossibility results and theoretical limits of AI safety and AI security reasoning.
    - Characterise the capabilities and limitations of agentic adversaries and explore novel security models such as: agents bounded by their level of intelligence (e.g., planning depth, sampling budget, world-model fidelity, size of the model) or whether parties have access to each other models (whitebox/blackbox)
+ AI communication security

    - Secure AI communication protocols: formal definitions, security games and provable defenses against prompt injection and adversarial manipulation (e.g. do we need more security features such as semantic integrity for AI communication?)
    - Safe emergent communication: understanding whether AI-to-AI languages should be free-form or constrained.
+ AI advantage
    - Designing new primitives unique to AI: agents can have more capabilities such as simulation access, credible commitments, cloning, self-modification, fine-tuning - are there more properties and unique security primitives for AI agents?
    - Designing games that demonstrate AI advantage: are there games that can identify if an action must have been performed by an agent? Are there protocols that would only be possible with agentic participants?
+ Generative Cryptography
    - Protocols for gathering user security requirements: designing models that are able to efficiently gather security requirements from users, languages for representing those and tools to verify if a proposed protocol satisfies these requirements
    - Using AI to write security proofs

3.2 – Agentic Game Theory

AI agents introduce new challenges as well as opportunities in game theory. For instance, identifying optimal strategies: machine learning revolutionised how humanity approaches computationally challenging problems. AlphaFold 2 won a Nobel Prize for "solving" the protein folding problem, which is NP-hard in the general case, not by violating complexity theory, but instead by showing that proteins of interest represent a narrow subspace of the problem where regularities and patterns can be leveraged. While this isn't specific to ML-based approaches (SAT and SMT solvers also approximate solutions to computationally challenging problems efficiently), we should expect that ML-based approaches may lead to revolutionary techniques for identification of optimal strategies.

This raises questions like:

+ What regularities might be leveraged for real-world games that can drive improvements in computational efficiency?
+ What are the dataset analogues of CASP that could be used to train such a model?
+ What would we do with an "AlphaStrategy" that could quickly identify nearly optimal strategies with the same performance improvements as AlphaFold 2?

+ Are there risks from creation of or public deployment of such an AlphaStrategy that should inform decisions around its creation (in the same way that an open-weight AlphaFold 2 could be used for bioweapon design)?

This topic is among a larger list of research questions we believe will be meaningful to look into.

3.3 – 'Nature' Cryptography

As agents interact with the physical world, there exists a latent opportunity to catalyse a new field of security that looks at using properties of nature to build security protocols, and verify physical and biological processes. This could extend the complexity and types of secure cyber-physical interactions agents can engage in, e.g., enabling autonomous engineering.

Some of this field already exists and has a sizable body of work: e.g., side-channel attacks, quantum cryptography, physically uncloneable functions. Yet some others are near non-existent at this point, such as protein cryptography or neuro-security. Crucially, there is no overarching community for people looking at the intersection of nature and security for embodied AI.

Through our discovery, we have found there exists a gap in the funding landscape for this kind of research, yet there is strong appetite from researchers to work on it. We believe there may be many low-hanging fruits to work on that become particularly relevant when combined with autonomous systems in the physical world.

You can find more papers on 'Nature crypto' in the resources section here.

How we expect to coordinate this effort

Safety

At the programme-level, we plan to:

+ Establish a Challenge Oversight Board (academia, industry, civil society) to adjudicate rule changes in challenges. *Who needs to be on this board to ensure it is trusted by all competitors – especially if international teams are involved?*
+ Open-by-default after 60 days with responsible release process (vuln embargoes, incident registry, red-team bounties)
+ Data rights & privacy: ban use-cases that could induce surveillance; require DP-style accounting for any human data and forbid biometric inferences in the arena. *Is there anything else we're missing here?*

+ Compute equity: provide baseline credits, hardware kits and grants so smaller teams can compete.

Aside from programme-level safety, we want to make sure safety and ethical concerns are infused to the core of the future we're building – this will be included in the shape of the challenges we run, the tools and research that we fund.

Some of the questions we're currently asking ourselves here include:

+ Making the real-world more verifiable and API-fying it for agents could invite new attack vectors and new surveillance programs that are breaches of our fundamental rights to privacy.
+ Accelerating the advent of autonomous systems without interpretability can pose significant risks.
+ Some individuals in AI Safety consider superintelligent agents to pose catastrophic risks, and instead argue for alternative systems that do not have execution power (the AI adviser paradigm rather than AI executor). We believe AI agents are coming and want to ensure the right trust infrastructure is laid out for them. These two views are not necessarily opposed but deserve more examination.
+ Teaching agents cryptography may allow unaligned agents or malicious actors to collude and cartelise against humans.
+ Early versions of these agents, if adopted widely, may lead to leakage of information.
+ If the capabilities we describe are built, they need to be accessible to all otherwise they will exacerbate inequality – how do we make sure this is rolled out to all? What is the role of compute here?
+ How does regulation come into play and how will agents interact with existing and new societal institutions? When do we engage relevant parties on this front? Who and how will system-wide rules aligned with society be decided?

We're keen to discuss with the community what you believe is important, and any of your suggestions for how to be thoughtful and engaged from the start in these issues and meaningfully incorporate thoughtfulness into the programme.

Open source, open weights

We see most of this work done to be open source, for the benefit of all and to accelerate progress. Naturally, contestants won't want to open source the work while they're participating in the challenges but we will ask all to open source their work after the fact (including open weights, open dataset for training/fine-tuning when relevant). We expect this to be bounded by safety concerns.

Service providers

We will strive to work with service providers to avoid duplicating work that has already been done by re-using existing infrastructure and tools. This also means we will happily give grants to existing projects to extend their work to fit the programme. *If this is you, please reach out!*

Cyber-physical arena

In order to lower the barrier to entry for individuals to participate in these games. We want to provide a high-fidelity Sim2Real environment that can be used anywhere in the world, as well as a dedicated physical environment where agents can be tested via trusted delegation. *Are there existing cyber-physical ranges we should partner with?*

Continuity

We want all three tracks to continue beyond the timing of an ARIA programme. We will ensure continuity by working with industry partners, and by making sure the artefacts of value are able to find economic sustainability.

How success will be measured

Our highest level measure of success is whether there are cyber-physical tasks where agents can coordinate as well as humans, or better than humans in untrusted settings (environment and/or counter-parties), under information asymmetry (e.g., private information) and optimising against competing objectives.

There are progressive milestones along the way that will tell us whether we are headed in the right direction.

+ **Milestone 1**: ascertaining a baseline of where we are today in terms of capabilities via component-level challenges and tools built
+ **Milestone 2**: being able to chart a roadmap from where we are today to where we could go via challenges results and fundamental research
+ **Milestone 3**: actively proving that via iterations, we are positively advancing in the roadmap by running multiple iterations of challenges and noticing improvements
+ **Milestone 4**: launching an end-to-end challenge
+ **Milestone 5**: launching the ARIA Flagship Challenge

Additionally, we expect artefacts in both Track 2 and 3 to be additional success proof points – tool adoption, meaningful papers that change how we see certain fields, and the building of new lasting research communities.

Why ARIA

Naturally, we need to ask ourselves why ARIA is best positioned to do this, and whether it is an opportunity that we are uniquely able to seize. We believe this to be true for three key reasons:

1) This programme will require bringing together experts across disparate fields such as AI security, multi-agent learning, complex systems, cybersecurity, game theory, distributed systems, technical AI governance, robotics, biosecurity, and more. This is the kind of task suited to ARIA's strengths, leveraging its pull among many different r&d communities (including many it is funding working in via other opportunity spaces).

2) The Arena can be neutral testing grounds, leveraging ARIA's convening power. This is something no single AI labs, or for-profit organisation would be able to run, but ARIA is uniquely positioned to run given its clear incentives.

3) Multi-agent settings in cyber-physical systems is the next frontier for AI and robotics, and is underserved at the moment as an area of focus relative to single-agent settings and human-alignment.

What we do not expect to fund

The problem of multi-agent security entails several moving pieces. Some of those pieces are important problems that are beyond the scope of our efforts in order to avoid stretching ourselves too thin. This doesn't mean we are not interested in these topics, rather we will look to external solutions/working with partners on them.

Those include:

+ **Identity for agents** – needed for recourse, and for traceability
+ **Commitment devices** – needed to enable classes of interactions that would not happen without it
+ **Communication channels** – likely needed if agents are constantly sending our preferences, or other data, to millions of other agents in real-time
+ **Socio-technical security defences** – how to align agents with multiple layers of society when objectives are competing
+ **Regulatory/Policy** – whether agents should be their own legal entity or tied to their legal owner, whether there should be regulation that complements architectural choices for agents.

What we are still trying to figure out

- What are the first sets of component-level challenges we should start with? Why?
- What about end-to-end challenges? and the ARIA Flagship Challenge?
- How do we ensure that challenges get smoothly and safely translated into real-world impact?
- What minimum viable Arena infrastructure is needed to have a low barrier to entry and channel competition energy towards the few vectors we are most interested in?
- Is there a risk that parts of this programme (e.g., Track 2) requires AI talent too expensive for our budget?
- If we're successful, there is a high chance value will be created fast here (as opposed to an R&D programme that produces a demo in five years). How do we ensure the UK can capture some of this value?
- What is one or a set of concrete, continuous measures of progress we can set for the programme?

ENGAGE

Our next step is to launch a funding opportunity derived or adapted from this programme formulation. Click here to register your interest, or to provide feedback that can help improve this programme thesis.

ACKNOWLEDGEMENTS

Written by Alex Obadia and Nicola Greco. Thanks to Evan Miyazono, Lewis Hammond, Ant Rowstron, Logan Graham, Daniel Freeman, Quintus Kilbourn, Edith-Clare Hall, Sarath Murugan, Harry Jenkins, Eder Medina, Orr Paradise, Davidad Dalrymple, Nora Amman, and all Trust Everything Everywhere Discovery workshop participants for their valuable contributions.

SOURCES

References cited, in chronological order.

- Coasean Bargaining At Scale, Seb Krier (Deepmind)
- The Coasean Singularity? Demand, Supply, and Market Design with AI Agents, Peyman Shahidi, Gili Rusak, Benjamin S. Manning, Andrey Fradkin, and John J. Horton
- Compute Forecast, Romeo Dean
- What is Embodied AI?, NVIDIA
- The Next $10 Trillion Opportunity: Why 'AI x Physical World' Is Where It's All Headed, Bilal Zuberi
- The Physical World is rate limiting, Ted Xiao
- Programmable Cryptography (Part 1), gubsheep (0xParc)
- SIEVE: Securing Information for Encrypted Verification and Evaluation, DARPA
- DPRIVE: Data Protection in Virtual Environments, DARPA
- SecureNumpy: Empowering Data Scientists with Secure Multi-Party Computation, The TikTok Privacy Innovation Team
- Opening up 'Zero-Knowledge Proof' technology to promote privacy in age assurance, Alan Stapelberg (Google)
- Extraction of Secrets from 40nm CMOS Gate Dielectric Breakdown Antifuses by FIB Passive Voltage Contrast, Andrew D. Zonenberg, Antony Moor, Daniel Slone, Lain Agan, and Mario Cop
- Hardness In Silicon, Quintus Kilbourn (Flashbots)
- Trustless TEE Overview June 2025, Quintus Kilbourn (Flashbots)
- NVIDIA Confidential Computing, NVIDIA
- Private Cloud Compute: A new frontier for AI privacy in the cloud, Apple
- Reimagining secure infrastructure for advanced AI, OpenAI
- Secure and secret cooperation in robotic swarms, Eduardo Castelló Ferrer, Thomas Hardjono, Alex 'Sandy' Pentland, and Marco Dorigo
- Privacy Reasoning in Ambiguous Contexts, Ren Yi, Octavian Suciu, Adria Gascon, Sarah Meiklejohn, Eugene Bagdasarian, and Marco Gruteser

- Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?, Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Soren Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King
- ARIA - Safeguarded AI Programme
- Recursive Joint Simulation in Games, Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer
- Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory, Andrew Critch, Michael Dennis, and Stuart Russell
- Conditional Recall, Christoph Schlegel and Xinyuan Sun
- Secret Collusion among AI Agents: Multi-Agent Deception via Steganography, Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt
- Project Vend: Can Claude run a small shop? (And why does that matter?), Anthropic
- General game playing
- Agents Rule of Two: A Practical Approach to AI Agent Security, Meta
- Models That Prove Their Own Correctness, Noga Amit, Shafi Goldwasser, Orr Paradise, and Guy N. Rothblum
- Foundations of Cooperative AI, Vincent Conitzer and Caspar Oesterheld
- Verification in physical systems enables autonomous engineering: from prototyping to manufacturing at scale, Eder Medina (Arcadia)
- Quantum Cryptography: Uncertainty in the Service of Privacy, Charles H. Bennett
- Physical One-Way Functions, Ravikanth Pappu, Ben Recht, Jason Taylor, and Neil Gershenfeld
- An Introduction to Protein Cryptography, Hayder Tirmazi and Tien Phuoc Tran,
- ARIA - Trust Everything, Everywhere Opportunity space resources

Some additional cool & relevant links:

- Trust Robots, Everywhere, Edith-Clare Hall (ARIA)
- NDAI Agreements, Matthew Stephenson, Andrew Miller, Xyn Sun, Bhargav Annem, and Rohan Parikh

- The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against Llm Jailbreaks and Prompt Injections, Milad Nasr, Nicholas Carlini, Chawin Sitawarin, Sander V. Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, Shuang Song, Harsh Chaudhari, Ilia Shumailov, Abhradeep Thakurta, Kai Yuanqing Xiao, Andreas Terzis, and Florian Tramèr
- Benchmarking for Breakthroughs, Seb Krier and Zhengdong Wang
- Don't lie to your friends: Learning what you know from collaborative self-play, Jacob Eisenstein, Reza Aghajani, Adam Fisch, Dheeru Dua, Fantine Huot, Mirella Lapata, Vicky Zayats, and Jonathan Berant
- Cyber Competitions, Anthropic Frontier Red Team
- Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents, Christian Schroeder de Witt (University of Oxford)
- AIxCC Darpa Cyber Challenge
- Infrastructure for AI Agents, Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung
- Virtual Agent Economies, Nenad Tomasev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero
- Learning Collusion in Episodic, Inventory-Constrained Markets, Paul Friedrich, Barna Pásztor, and Giorgia Ramponi
- AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents, Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr
- Language Models Can Reduce Asymmetry in Information Markets, Nasim Rahaman, Martin Weiss, Manuel Wurthrich, Yoshua Bengio, Li Erran Li, Chris Pal, Bernhard Schölkopf
- Mechanism Design for Large Language Models, Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo
- TTEE: Marrying Cryptography and Physics, Quintus Kilbourn (Flashbots)
- Cryptographic Sensing, Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai
- Computer-inspired Quantum Experiments, Mario Krenn, Manuel Erhard, and Anton Zeilinger

For those interested in diving deeper into resources, check out https://www.aria.org.uk/opportunity-spaces/trust-everything-everywhere#resources

APPENDIX

Prompts to use this document with

We welcome you feeding this document to Gemini, ChatGPT or other tools in order to chat with it! Here are some prompts you may find useful:

+ If this programme is successful, what happens in 2035? Continue the sentence 'The year is 2035...', painting a vivid picture of what the world could look like in 2035, anchoring it in a real use-case.
+ Suggest challenges for each type, spec them out fully and explain how someone would participate.
+ If this programme is successful, what would a second programme in the Trust Everything Everywhere Opportunity Space look like?
+ Tell us your own prompts!