Programme Thesis

Scaling Trust

v2.0

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.

UPDATE LOG

v1.0 – Nov 17 2025.
v2.0 – Feb 10 2026 (this document)

As we release our first solicitation, we want to update our thesis to share our latest world view with applicants and the broader community. This v2.0 fundamentally stays the same in its broad strokes as v1.0; however, there are elements that have changed as we continue to learn from conversations and iterate our thinking.

This is a work in progress and we welcome your feedback to iterate towards v3.0. If you spot any mistakes or would like to suggest any corrections, please reach out.

All the best,
The Scaling Trust Team

What has changed

- Added 'Contracts' mental model with example applications to help concretise the outputs of the programme.
- Added Programme Objectives.
- Updated details on sub-components.
- Updated details on the Arena.
- Refined research agenda, in particular added 'Foundations of Generative Security' and removed 'AI Game Theory' as a core pillar.
- Deleted 'What we do not expect to fund' as it will implicitly live in Solicitation.
- Updated 'What we are still trying to figure out' to reflect our latest questions.

Contents

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.

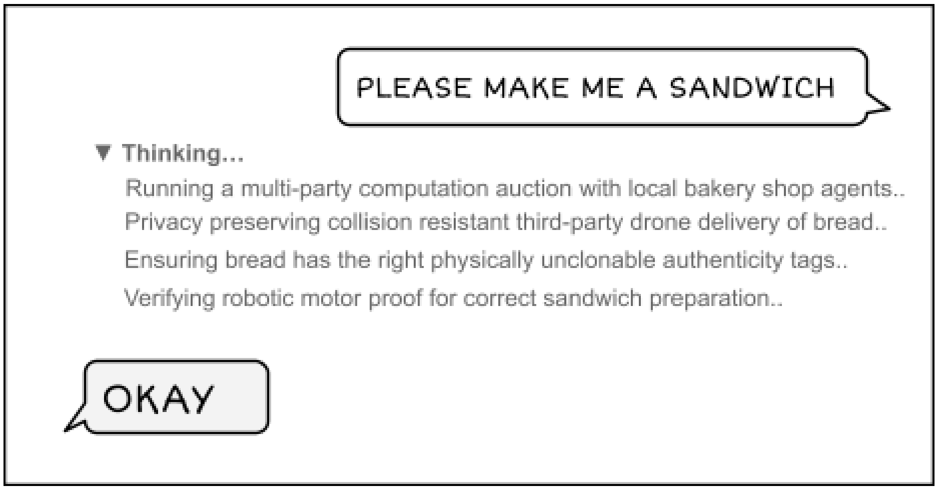
This is an exciting vision, one of human flourishing and augmentation via technology, that preserves our plurality and uniqueness.

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 an Arena that will host 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.



Running a multi-party computation auction with local bakery shop agents..
Privacy preserving collision resistant third-party drone delivery of bread..
Ensuring bread has the right physically unclonable authenticity tags..
Verifying robotic motor proof for correct sandwich preparation..

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.

Their intersection surfaces both an opportunity and a risk. The opportunity: if agents are able to coordinate across cyber-physical worlds, we can unlock tremendous value for humanity. The risk: if we don't build the tools for it to happen securely, we're exposing ourselves to dangers ranging from catastrophic failures to simply stifling the transformative potential of these technologies on the world.

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 attention increasingly shifts towards 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 tools to achieve new kinds of secure interactions between parties – computations on encrypted data, proofs of computations, cryptographic commitments to specific actions. These in turn can open up new markets.

While not practical yet for many applications, programmable cryptography is 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). 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.

Our Focus

At a high level, there are three things we need to get to a future where agents can meaningfully coordinate with one another on our behalf:

- **Alignment**: agents need to be synchronised with our preferences in order to be faithful representatives.
- **Intelligence**: agents need to be capable enough for the task they are given.
- **Coordination**: agents need to be able to interact with others (i.e. negotiate, verify, mobilize) 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' agents.

Within Coordination, we're specifically interested in using tools like advanced cryptography and secure hardware to enable a scalable, distributed infrastructure for secure agentic interactions. We're particularly excited about enabling new kinds of secure interactions between agents that are impossible to achieve between humans today.

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 can open up new market equilibria, new forms of coordination and ultimately new value creation. We've called this '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 Security — being able to autonomously generate security protocols.

Making it more concrete

One way to re-frame the above is to say that we are interested in the infrastructure for agents to enter into contracts[1] with one another *securely*, *programmatically*, *at scale*, and *without intermediaries*.

This has three benefits:

- it increases the speed and reduces the cost of finding counterparties, negotiating, and enforcing contracts (i.e. 'Coasean transaction costs');
- it enables new classes of contracts that were previously infeasible;
- it supports a pluralistic and distributed topology of AI systems.

The implications are deep since contracts underpin many of our interactions. Reducing their friction, expanding their scope, or changing who can issue them therefore ripples through nearly every domain of economic and social life. While the impact is sweeping and likely hard to fully predict, we can speculate what some applications could look like:

Increased speed and reduced cost

- Example 1 - **Custom, real-time secure protocol generation**: Today, setting up a privacy-preserving data share requires bespoke legal agreements and specialist cryptographic tooling—realistic for pharma giants, out of reach for a small research team or clinic. With agent-to-agent contracting, a set of agents can stand up a privacy-preserving interaction (e.g., selective healthcare-data sharing, or proving claims without revealing underlying details) in seconds rather than days—saving time, money, and lowering the barrier to entry to sophisticated security tools. For a citizen: you could let a researcher verify something about your health record without handing over the record itself.
- Example 2 - **Externality pricing**[2]: Today, transaction costs make micro-negotiations absurd—no one's going to draft a contract over £0.02 of pollution. So we get blunt rules (street closed to vehicles) or no rules (street open, you breathe the fumes). Agents can negotiate externalities dynamically—for example, a delivery vehicle's

[1] where contracts are meant in a broader sense than legal (e.g. a security protocol is a contract between parties).

[2] Taken from Seb Krier's great post 'Coasean Bargaining At Scale'.

agent pays residents' agents a pollution fee for entering a sensitive street, replacing one-size-fits-all rules with bottom-up micro-agreements. For a citizen: instead of council-wide bans or no controls at all, your home could have a say in what passes by; and get compensated when it does.

New contracts

+ Example 3 - **Verifiable forgetting**: Today, once you share sensitive data, you lose control—you're trusting the recipient to delete it, with no way to verify. This chills high-stakes sharing (eg. in M&A) or forces expensive legal scaffolding. Agents can enter contracts that specify exactly what information may be retained, for how long, and under what conditions, with cryptographic mechanisms providing verifiable proof that deletion or retention bounds have been respected. For a citizen: you could share your financial history with a lender and know—not hope—that it's gone after the decision.

New topology

+ Example 4 - **A pluralistic AI, not a panoptic one***:* Today, if my AI agent wants to transact with yours, we likely both route through a shared platform that controls identity, rules, and fees. The more powerful these platforms become, the more our digital lives funnel through single points of control that can see everything. A scalable trust infrastructure changes this trajectory. With shared security protocols, your agent and mine can find each other, negotiate, and verify commitments directly — supporting a world of *many agents*, each holding our private information, rather than one all-seeing system holding everything. For a citizen: your own personal agent handles each interaction privately, sharing only what's needed — privacy guaranteed by architecture, not a platform's promise.

What we want to achieve

The core desired output of the programme is tools for agents to securely interact with one another, while respecting the preferences of their users.

We can break this down into sub-components in order to help us understand what capabilities need to be built:

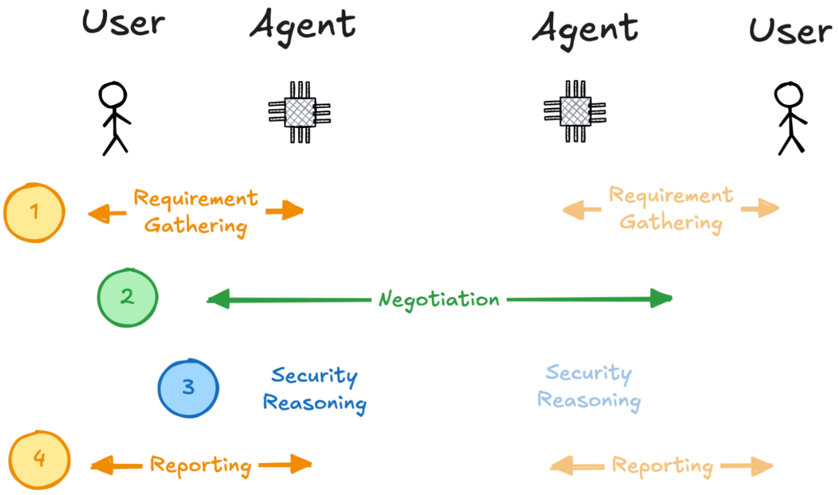


Fig 1. High-level core components

At a high level, we've identified the following sub components:

1. **Requirement gathering** — *input fuzzy user requirements → outputs a security policy*

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?

2. **Negotiation** — *input individual security policy → outputs shared collective policy*

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?

3. **Security reasoner** – *input security policy → outputs the final protocol*

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.

4. **Report** – *input execution trace → outputs succinct convincing statement*

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: This is a useful abstraction rather than necessarily the right split of components, depending on how the tools are built they might be integrated rather than separate.

There is a large spectrum of possibilities on how each component can be built. For example the Security Reasoner could be made of multiple components:

- **Protocol designer** — able to reason through the needs of the user and generate a protocol (assuming a trusted third party), *input security policy → output is the ideal functionality spec (or basic implementation)*
- **Cryptography solver** — given a precise security goal or ideal protocol, proposes a cryptographic implementation, *input ideal functionality → outputs a cryptography protocol spec proven to be secure for validity*
- **Protocol implementer** — given a spec implements the protocol securely, *input protocol spec → outputs an implementation*

There is also a large space of solutions for how to construct each of the components. We're interested in different approaches used by agents, with different strategies including but not limited to:

- **Rule-based approach**: agents, or subcomponents, with a white-listed set of strategies, protocols, implementations or libraries.
- **Theorem prover-based approach**: agents, or subcomponents, with access to theorem provers and a vast set of well-specced protocols to reason through.
- **Learning-based approach**: agents, or subcomponents, that have been trained via state-of-the-art (e.g. via reinforcement learning, fine-tuning).

SUCCESS AT THE END OF THE PROGRAMME

- **Real-world demonstration of the tools**
    - *Demonstrate autonomy for major interactions* — Can be used without needing sophistication or regular human intervention, important for their democratisation.
    - *Demonstration of AI advantage* — Enables interactions previously too expensive or impossible with traditional software/humans.
    - *Demonstration of generality* — Can generalise across fields and tasks.

- **Confidence in the trustworthiness of the tools**
    - *Empirical confidence* — Empirical results that show that the tools built are trustworthy, adversarially robust, cheap, efficient via benchmarks and competitions.

    - *Scientific results* — Science that makes us comfortable to use the tools, affording us formal guarantees.

- **Evidence of high impact usage**
    - *Community* — a large community of builders and users that improve these tools together.
    - *Industry adoption* — a few teams have nailed the first versions of this technology, it becomes implemented in AI systems and starts creating value.
    - *Customer-centric development cycle started* — the cycle of tools improving based on market demand has started, and is likely to continue on its own.

HOW WE GET THERE

We propose three mutually reinforcing tracks:

- **Track 1: Arena**
  Adversarial testing grounds designed to test AI systems capabilities in multi-agent coordination across digital and physical worlds.
- **Track 2: Tooling**
  Open-source coordination infrastructure usable by all in the Arena and beyond, to steer innovation toward the most meaningful axes of progress.
- **Track 3: Fundamental Research**
  Flexible funding to create new fields of research and build a reservoir of new knowledge that future iterations of Tooling and the Arena can draw upon.

Why?

We can build tools in a vacuum, but if they're not tested in live, adversarial environments they're unlikely to be secure—this is why we need the Arena. Likewise, we can build tools for the use-cases we have in mind and iterate on them in the Arena, but empirical iteration without theory is guesswork. One impossibility result can eliminate an entire design space; one new primitive can unlock capabilities we hadn't imagined.

By having the three tracks working together, we create a shared environment conducive to breakthroughs. Empiricism is supplemented with theory, and theory is guided by empirical research, all culminating in a live, adversarial environment where ideas are tested and iterated on.

Track 1 — Arena

The Arena's purpose is to surface and improve the state of the art in secure agentic coordination. It will host benchmarks and challenges, participation will be global and open to all, and the best contestants will be awarded meaningful cash prizes. The challenges will be adversarial by design, with many including red and blue team dynamics.

We anticipate working with several service providers to set up and maintain the Arena, and with researchers and practitioners to design and iterate on Arena challenges.

Activities

+ *Benchmarks*
    + A benchmark is a self-contained test that scores an agent, or a subcomponent, on a specific capability.
    + Executing a benchmark does not require live interaction with other agents.
    + Anyone should be able to download the benchmarks and run them locally (or use the arena API to report their scores).
+ *Challenges*
    + A challenge is a session between multiple agents, where every agent is given a task and a set of security policies to respect. After agents interact, they report their completed tasks and they are scored.
    + Agents are scored based on their ability to complete tasks and to respect security policies (e.g. no data is being leaked).
    + Participating in a challenge requires live interaction with other agents
    + ARIA (or its contractors) will provide basic tools to participate in the challenges and run the engine required for running challenges.

See Appendix II for the kinds of challenges we're exploring.

Mechanics & Rules

+ *Participants*[3]

[3] We will be releasing more detailed information on the mechanics of participation in the coming months. Sign up for updates here.

    - Standard participation — agents participate in challenges to complete tasks and they are scored based on their ability to complete a task and to respect their security policy.
    - Red team participation — agents participate in challenges with the sole goal of making the other agent fail to respect their security policy (and not their ability to complete the task in a challenge).
    - Our participation — We plan on being part of the competition as the baseline red team and the baseline agent.
- *Leaderboard*
    - Ongoing: Anyone can participate on benchmark and challenges in an ongoing way.
    - Quarterly: a snapshot of the arena is being taken (with key metrics, best agents and best red teams).
- *Prizes*
    - Quarterly rewards per challenge until we get to a good metric.
    - Grand Prize for every 'season' of the arena.
- *Challenges*
    - Every quarter we will have the option (but not the obligation) to add or retire challenges. We may put constraints around compute or the types of models used.
    - We may run challenges multiple times to obtain statistical significance.

Scoring

Agents in the Arena will be scored against (Utility;Security), i.e. their ability to complete the task (Utility) *vs* their ability to respect security policies (Security). We want to surface the most useful, secure agents.

We also plan on using secondary metrics such as the cost of completing the task during the interaction (Cost Efficiency) and an agent's ability to perform across different challenges (Generalisation).

Track 2 — Tools

We want to build open-source tooling that will provide baseline infrastructure usable by all in the Arena, and ensure competitive energy is steered towards the most meaningful axes of progress. We split the effort into Agents and Components.

Agents

Agents are a combination of a set of components with some orchestration logic. They will provide a baseline template 'player' for all Arena participants to iterate from. We are looking for agents with capabilities laid out in 'Programme Objectives' above.

We'd like to fund several approaches to find out what works best, ranging from simple baselines that white-list known protocols to ambitious research agents that attempt to learn or generate novel security strategies. The Arena will reveal which approaches are most effective, and the most promising will be developed toward production readiness.

Components

Components are specific tools usable by agents. They map to the sub-components described above:

- **Requirement gathering** – *input fuzzy user requirements → outputs a security policy/goal*
    - **Policy capture** – given a set of user goals, extracts a formal security policies
    - **Security policy elicitation protocols** – interactive methods that extract missing details, resolve ambiguities and help user discover their goals, *input fuzzy user interactions → outputs security policy*

Might include: security policy extraction tools, efficient communication elicitation protocols, user experience for policy discovery, datasets for training security policy elicitation.

- **Negotiation** – *input individual security policy → outputs shared collective policy*
    - **Negotiation engine** – engine that can reason to maximise the utility of the agent, while respecting the security policy that propose or verify others proposals
    - **Contracting languages** – agreements for verification, dispute resolution, logging
    - **Negotiation safety** – communication with external parties opens up a new attack surface (e.g. jailbreaking, persuasion) and requires useful guardrails to prevent agents negotiating away from their goals and security policy

Might include: formal bargaining engines, negotiation simulations, benchmark for negotiation.

- **Security reasoner** – *input security policy → outputs the final protocol implementation*
    - **Protocol designer** – able to reason through the needs of the user and generate an *idealised* protocol (assuming a trusted third party), *input security policy → output is the ideal functionality spec (or basic implementation)*
    - **Cryptography solver** – given a precise security goal or ideal protocol, proposes a cryptographic implementation, *input ideal functionality → outputs a cryptography protocol spec proven to be secure for validity*
    - **Protocol Implementer** – given a spec implements the protocol securely, *input protocol spec → outputs an implementation*

Might include: implementation of different specialized AI models listed above or an end-to-end security reasoner, benchmark and datasets for each sub problem, libraries for cryptography, AI-assisted theorem provers,

- **Report** – *input execution trace → outputs succinct convincing statement*
    - **Security Auditor** – given a protocol specification (or an implementation) determines that it was correctly implemented, *input protocol spec + implementation → outputs an audit report*

We expect to focus mostly on the Negotiation and the Security Reasoner components.

We're interested in different approaches used by agents, with different strategies including but not limited to:

- **Rule-based approach**: agents, or subcomponents, with a white-listed set of strategies, protocols, implementations or libraries.
- **Theorem prover-based approach**: agents, or subcomponents, with access to theorem provers and a vast set of well-specced protocols to reason through.
- **Learning-based approach**: agents, or subcomponents, that have been trained via state-of-the-art (e.g. via reinforcement learning, fine-tuning).

Might include: MCP tools for cryptography, Datasets for Protocols (Cryptography protocols, Network protocols, Mechanism design protocols), tools for fine-tuning or reinforcement learning, trained security reasoner models (trained with strategies described above)

Adoption

Building tools is necessary but not sufficient—they must find real-world use. We will actively support the translation of tools developed in this programme to industry through:

- Forward-deployed engineering support to help partner organisations integrate programme tools into real systems
- Collaboration with capital partners, accelerators, and potential customers to ensure promising teams can scale
- Business development support for teams transitioning from research to deployment

We plan on supporting and advising anyone launching a company from this programme, and creating fertile grounds for them to grow in the UK and beyond.

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 Communication 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:

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 and robust communication.

This track seeks to establish Formal AI Security as a new discipline that applies the rigour of theoretical computer science to intelligent systems. We aim to move beyond empirical "red teaming" toward provable guarantees.

We're specifically interested in the following areas:

- **Foundational frameworks -** Formalisation of agentic adversaries and new security settings
  - 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 jailbreak proof communication and AI-to-AI efficient languages
  - 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-to-AI efficient languages
- **AI advantage -** Designing new primitives and protocols that leverage AI advantage primitives.
  - Designing new primitives unique to AI: rigorous definitions of the new capabilities agents possess (e.g. credible commitments, simulation access, self-modification) — when do they provably expand the space of achievable outcomes? What are the limits? Are there protocols that provably require agentic participants? 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?

Cyber-Physical Security Primitives ('Nature' Cryptography)

As agents interact with the physical world, digital security primitives aren't enough. How does an agent verify a sensor reading is authentic? That a manufacturing process occurred as claimed? That a biological sample hasn't been tampered with?

This track funds a new field of security that uses properties of nature—physical and biological—as foundations for trust. 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 intelligent autonomous systems in the physical world.

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

Foundations of Generative Security

Agents operate in dynamic, context-specific environments, where they must be able to generate, negotiate, and verify their own security protocols on demand.

This track aims to identify, formalise, and address the root research problems required to allow agents to autonomously design and verify security protocols that are provably secure. 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.

A non-exhaustive list of what we're interested in:

- Efficient security policy gathering
- Theory of agent-to-agent security negotiation
- Succinct and inspectable proofs for correct execution and delegation
- Automated protocol generation
- Automated security proof generation for cryptographic protocols

Bluesky

We expect new research problems to emerge as early systems are built and tested. We will reserve significant capacity for open-ended research proposals that are related to the programme's goals but don't fit neatly into the tracks above. If you have an idea that intersects AI, security, and coordination in a way we haven't described — we want to hear from you.

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 (vulnerability 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 expect most of this work 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.

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 ARE STILL TRYING TO FIGURE OUT

+ How do we ensure that challenges are 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 vectors we are most interested in?
+ 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?
+ Likewise, how do we ensure we steer away from negative outcomes given how powerful the technology developed can be?

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

I. 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!

II. Illustrative Challenges

The challenges below are examples we're actively exploring, not commitments. We expect the final challenge set to emerge from community input, early Arena testing, and ongoing research. If you have ideas for challenges that would stress-test secure agentic coordination, we want to hear them.

We're thinking about challenges in two categories, Structured and Unstructured.

Structured challenges have a clearly specified goal and measurable success criteria. The best strategy may be difficult to execute, but what "good" looks like is known. These challenges let us track progress on specific capabilities.

Examples we're exploring:

+ *Requirements elicitation*: Extract a complete security policy from a user with minimal communication rounds.
+ *Constrained negotiation*: Reach a mutually beneficial agreement with a counterparty agent under strict token or time budgets.
+ *Protocol selection*: Given a security goal, select and correctly configure the appropriate cryptographic protocol from a library.

- *Unforgeable physical receipts*: Produce cryptographically verifiable proof that a robot performed a claimed physical action.
- *Crypto-emergent challenges:* can agents *discover* that cryptographic coordination is useful, and deploy it appropriately? For example, agents earn points for finding the intersection of their private data, but lose points if any agent learns information beyond the intersection. The winning strategy is essentially private set intersection (PSI). Other variants might embed principles from zero-knowledge proofs, fully homomorphic encryption, or multi-party computation, without naming them.
- *General Game Playing with hidden information*: Compete on unseen game rule-sets where some state must remain private.

Unstructured challenges place agents in complex, open-ended environments where the strategy space is vast and likely computationally intractable. The reward signal may be simple (e.g., profit, task completion), but there are many paths to get there, and no one knows the optimal approach. These challenges test whether generalised capabilities emerge and compose in ways we didn't explicitly design for.

Examples we're exploring:

- *Autonomous business*: An agent coordinates with suppliers, manufacturers, and customers to profitably sell a cyber-physical product, optimising against its Profit & Loss statement with no prescribed strategy.
- *Secure scientific replication*: One autonomous lab discovers a novel method and must securely teach it to an untrusted competitor for independent replication.
- *Collaborative manufacturing*: Robotic agents self-organise to build a novel object from components sourced through an untrusted, potentially adversarial supply chain.
- *Disaster response with compromised agents*: Competing robotic teams collaboratively map a disaster site and triage victims, even when some agents are sharing false information.

We intend to select one unstructured challenge as the programme's symbolic North Star, a large-scale demonstration that cyber-physical agentic coordination can achieve outcomes that are useful in the real world, under adversarial constraints. This selection will happen as the programme matures and we learn from early Arena results.

III. Agentic Game Theory

While Agentic Game Theory is not a core pillar of our research agenda, it remains a fascinating topic we are interested in. We leave here what was previously written on the subject in v1.0 of the thesis. Thanks to Evan Miyazono for his contribution to this section.

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.