High Assurance AI Technical Roadmap

Table of Contents

What is CommonAI CIC? **2**
How Programmes Work 2
The Technical Roadmap 2
High Assurance AI **3**
Vision 3
Current Members 4
Areas of Interest 5
Model Fabric Workstream **6**
Key purpose 6
Concept 6
Technical Priorities 7
Panopticon Workstream **9**
Key Purpose 9
Concept 9
Technical Priorities 9
Auto-formalisation Workstream **11**
Key Purpose 11
Concept 11
Technical Priorities 11
Workstreams in development **15**
Multi-agent systems 15
Key Purpose 15
Cross Section Areas **17**
Domain definition 17
Synthetic data 18
Security 18
Observability 19

What is CommonAI CIC?

CommonAI C.I.C. is a not-for-profit Community Interest Company promoting safe and responsible AI and machine learning technologies, and the libraries and data they require. It enables startups, larger companies, public sector bodies and universities to solve shared technical challenges and develop foundational AI for their own products and services, which may later be open-sourced, after an embargo period, for wider benefit.

How Programmes Work

Organisations join as members of one or more programmes, each with a specific technical focus. Current programmes are High Assurance AI and the Scaling Inference Lab, with more planned.

Members build foundational AI and ML technologies, in collaboration with other member organisations and CommonAI's engineers. In return they get early access to the resulting IP for their own product, while keeping full ownership of their work. They also benefit from involvement in programme governance, access to member-only resources (e.g. discounted compute) and a pathway to production. In return, members commit engineering time to the Technical Roadmap and pay an annual fee that is ring-fenced for their chosen programmes. For more detail on membership see our Member Agreement.

The Technical Roadmap is delivered by members and CommonAI engineers collaborating in **Working Groups**. In the High Assurance AI programme these Working Groups follow the key technical workstreams: **Model Fabric**, **Panopticon**, and **Auto-Formalisation**.

IP developed in the programme, including access to programme repositories, will generally be held under an embargo period of 12 months. During that time only members can use it; after this it will typically be released to the open source. For more details, see our IP Licence & Policy and Governance Framework.

The Technical Roadmap

Each programme has a **Technical Roadmap**, set under CommonAI's Governance Framework, that outlines a current work-plan for the Programme, including high-level targets, major milestones and overall prioritisation of target features. It is a living document that evolves as the programme develops. This is the public version of the roadmap; an internal version goes into greater technical depth and sets out the project plan; this version is available to members. For more information you can reach out to info@commonai.org or visit our website.

High Assurance AI

Vision

AI holds enormous promise in sectors where compliance, correctness and safety are of business critical importance. The deployment of AI solutions (including LLMs) requires systems that operate safely, reliably and at scale. **The High Assurance AI Programme will build the technical foundations that make safe, compliant AI deployment achievable in critical and regulated industries.**

High Assurance refers to systems, products, or services with extremely high levels of trustworthiness, security, and reliability, designed to resist sophisticated attacks and guarantee function as intended, often involving rigorous evaluation, formal methods, and specialised development for critical government, defence, and national infrastructure.

Barclays is the programme anchor, with initial focus on solving strategic use cases and industry-wide challenges in financial services.

Current Members

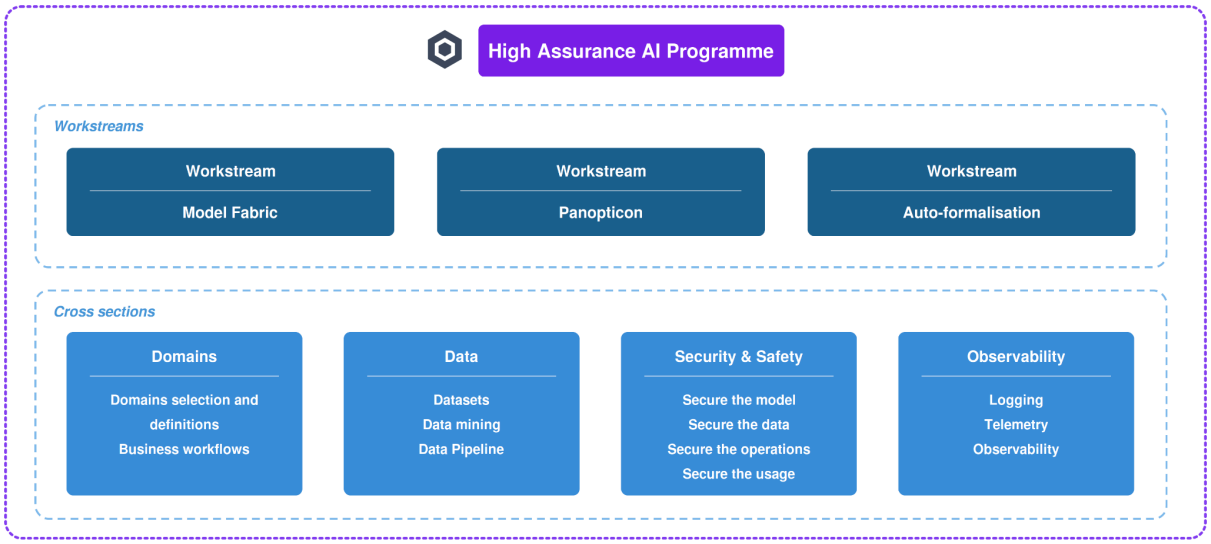


- **Barclays:** anchor to the High Assurance programme, providing strategic direction and applying the technical foundations within financial services.
- **Cambridge Inference:** develops high-assurance AI and optimisation systems for regulated environments. Their work spans high-performance Bayesian optimisation for banks, AI control and monitoring infrastructure, frontier AI safety research, and UK AI policy. For the High Assurance programme they are developing Panopticon, a monitoring and control layer. Panopticon sits between users and AI systems. deploying cutting-edge control protocols to audit interactions, sandbox code execution, and catch risks in real time.
- **Meridian Cambridge:** independent research and incubation hub focused on AI safety, biosecurity, frontier-risk policy, and institutional design. They are currently collaborating with the Cambridge Inference team in the High Assurance programme, informing the design and testing of novel control protocols.
- **Proof Relevant:** Specialises in formal methods, mathematical reasoning and proof-based techniques for verifying complex software and AI systems. Within the High Assurance Programme, they bring expertise in making system behaviour precise, auditable and deployable with high assurance. This capability is essential where AI operates in regulated, high-value environments that demand correctness, safety and compliance. Proof Relevant leads the Auto-formalisation workstream.
- **Aurium Labs:** transform specialised workflows and fragmented records into controlled synthetic training data which are sufficiently representative for model training and testing. They are a verticalized foundry optimising data, models and deployment towards domain specific use-cases. Within the High Assurance AI programme, they are working across data fidelity, synthetic data, evals and training methods.

We expect to onboard more members soon, and welcome applications from companies large and small, non-profits, and academic and governmental bodies.

Contact us at info@commonai.org to find out more.

Areas of Interest



The programme currently runs three workstreams: Model Fabric, Panopticon, and Auto-Formalisation. More, including Multi-Agent-Systems, are in development and will be added as the programme grows.

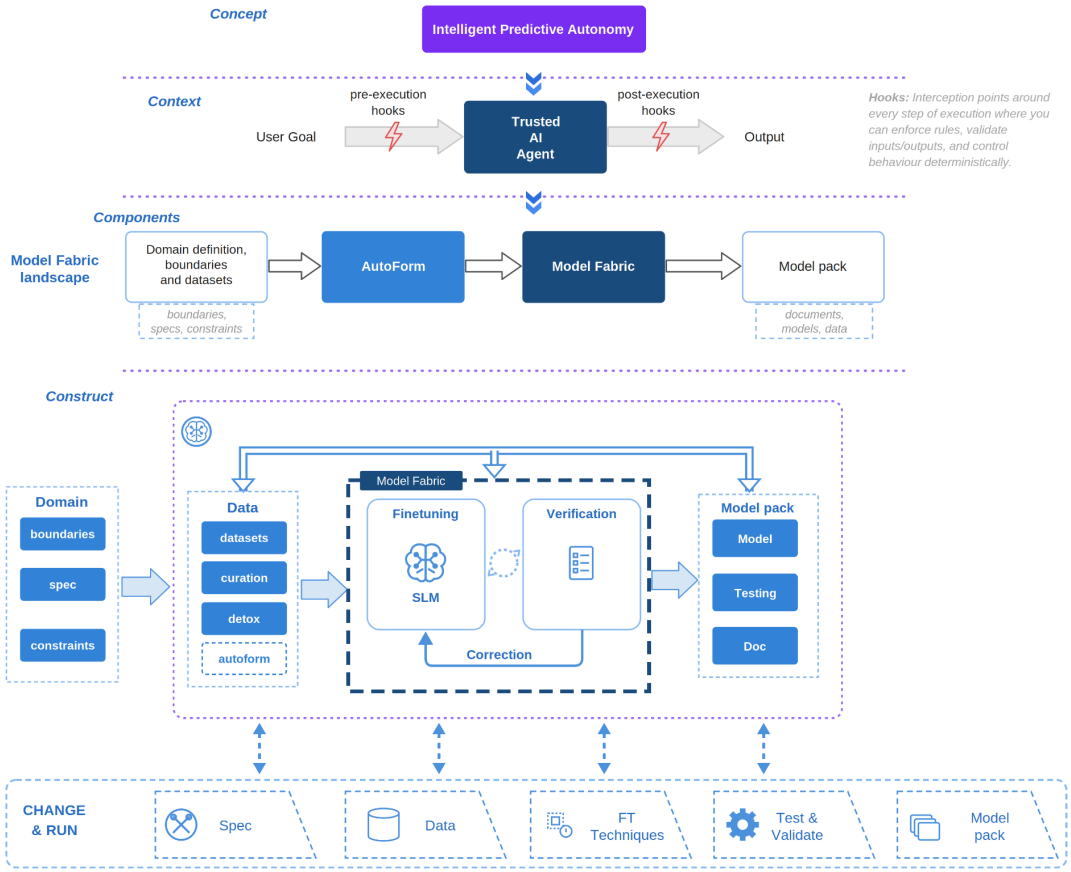
Model Fabric Workstream

Key purpose

Model Fabric is a framework that enables teams to rapidly create, train, evaluate and operate domain-specific small language models (SLMs) through a governed, repeatable process. Its current objective is to scale SLM development approaches within financial institutions (with Barclays as an early partner) , moving away from building individual models as one-off initiatives. This approach aligns with the strategy of Barclays and other regulated organisations, where domain specialisation, bounded contexts, governance and operational confidence are essential.

Concept

Our reference architecture positions the Model Fabric as a foundational component within the broader vision of establishing a governed, automated platform for enterprise-scale, domain-specialized agents. It serves as the core intelligence layer for agents operating within broader multi-agent ecosystems.



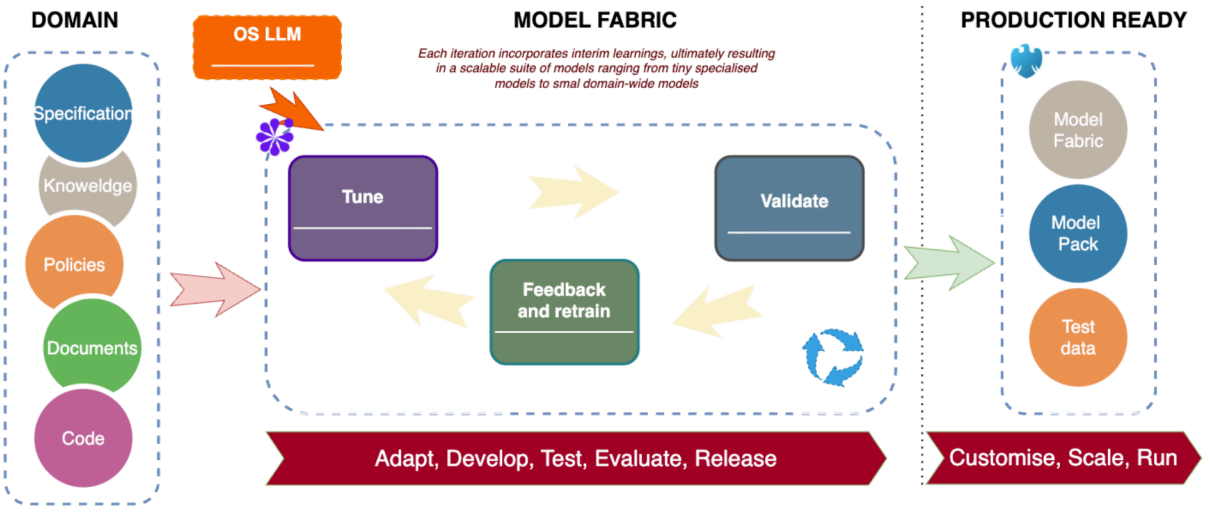
Technical Priorities

Domain Modelling

In this context, a domain represents the specific area of knowledge or set of tasks for which the system is responsible. Modelling a domain means formally defining what is in and out of scope for the system. This mirrors the bounded context in the Model Fabric architecture, a Domain-Driven Design principle where systems are built around clearly separated subdomains. For agentic AI, this means using specialised agents, knowledge bases or prompts for different areas, rather than expecting one agent to know everything.

Model Fabric Flow

The goal is to deliver a production-ready Model Fabric framework for rapidly creating and operating small language models (SLMs) across any business domain. By leveraging a common architecture, standardized tooling, reusable components, and governed delivery patterns, the framework accelerates time-to-market, maximizes model reuse, and maintains enterprise-wide consistency and compliance.



The key stages are listed below:

Model Selection

The Open-Source LLM selection stage identifies the most appropriate foundation model and adaptation strategy for the target use case. Selection should consider domain requirements, model capabilities, deployment constraints, and efficiency considerations to ensure the chosen model is fit for purpose.

SLM Training

The Training stage adapts the selected foundation model to the target domain through approved training and optimisation techniques. This includes parameter-efficient fine-tuning approaches, preference-based optimisation, distillation techniques, and output constraint mechanisms such as schemas, DSLs, and ontologies. The objective is to create a specialised SLM that is both effective within its domain and capable of producing structured, consistent outputs.

Evaluation & Retrain

The Evaluation stage assesses whether the trained model meets functional and domain requirements. Evaluation activities provide evidence of model quality and trustworthiness by measuring factors such as domain accuracy, benchmark performance, hallucination rates, robustness, and safety characteristics.

Benchmarking and Validation Frameworks

Establishes consistent methods for measuring fitness for purpose, covering domain-specific benchmarks, accuracy, hallucination, robustness and safety, alongside retraining and model versioning approaches.

Model Pack

The Model Pack stage captures reusable artefacts generated during the SLM development process. These artefacts may include model cards, dataset cards, benchmark suites, implementation templates, and recommended training recipes. The purpose is to package lessons learned and proven approaches into a reusable capability that can accelerate adoption during the production-ready phase. These artefacts will be published and made available, subject to the licensing arrangements defined in CommonAI's membership terms, through authorised distribution channels, following the embargo period.

Implementation

Provides implementation patterns, reference architectures and reusable pipelines, creating a standardised path from business idea to production-ready domain SLM.

Panopticon Workstream

Key Purpose

Panopticon is an AI control platform that aims to replace the industry's current, fragmented set of tools with a unified control layer providing comprehensive governance, security, and auditing for enterprise AI applications. Already in advanced development and built in Rust for maximum performance and security, Panopticon will provide a cloud-agnostic platform with sophisticated control protocols, managed sandboxed execution, and robust auditing capabilities.

Concept

Panopticon mitigates risks that arise in three core areas:

- Inputs: e.g. prompt filtering, preventing misuse
- Model state: e.g. whitebox monitoring, reasoning validation
- Outputs: e.g. sandboxing, blackbox monitoring, audit and replay

CommonAI is working with Cambridge Inference to deliver Panopticon's core components. These will be made available without additional charge to programme members during an initial exclusive period, and subsequently released as open source. Cambridge Inference also plans to offer paid enterprise extensions, while members remain free to build their own versions at no cost.

Technical Priorities

Core platform (Delivered):

The current working platform, available to members under embargo, monitors and constrains actions in real time, supports a combination of self-hosted models and external providers, and maintains a full audit trail of all interactions. Additional audit features include working-state rollback and branching trajectories for course-correction and scenario analysis. Panopticon provides a flexible framework for constructing control protocols, supporting automated triggering and sync/async scheduling, together with secure isolated tool execution. An initial set of oversight protocols are included, validated against published research within a comprehensive evaluation framework, alongside a demonstration interface.

Advanced protocols (in progress):

Support for further major AI providers, plus more advanced controls that allow uncertain AI responses to be checked and regenerated, benchmarked against published results. Work is beginning on monitoring AI reasoning and evaluating prompt quality.

V1.0 release

Release of Panopticon v1.0, including control agents, conversations spanning multiple AI providers, a production interface for small teams, completed human-oversight controls with benchmarks, and full documentation. Further direction will be decided based on findings to date, with options including developer tooling or methods that examine a model's internal workings.

Auto-formalisation Workstream

Key Purpose

The auto-formalisation workstream focuses on converting natural-language policies, regulations, standards and business rules into precise, machine-checkable rules. This allows critical decisions to be calculated consistently and verifiably, with full audit trails, rather than relying on the probabilistic output of an AI model. We see this as a key step towards high-assurance AI, accelerating trustworthy deployment in regulated sectors, particularly financial services, with Barclays as an early partner.

The auto-formalisation work is driven by Proof Relevant. This section sets out the working group's initial deliverables within the CommonAI High Assurance AI Programme. Collaboration in the workstream will focus on regulatory priorities, benchmark design, evaluation and business integration.

Concept

The idea of Proof Relevant *pr.solve* is to describe a policy document explicitly as a set of rules that a computer can check. It accommodates complex regulations in such a way that answers are still explainable. A clause can often be interpreted in more than one way, and then we flag the ambiguity, rather than choosing one reading. It can also show which facts of a case can be changed without changing the answer, and which cannot.

Technical Priorities

Platform and governance

Establishing a shared platform will provide a common framework for collaboration. This includes agreeing secure, consistent access to a range of AI models (including leading commercial, smaller and self-hosted models), scoping integration with relevant platforms and AI coding tools.

Focus

The potential applicability of Auto-formalisation methods is broad, and early focus areas will therefore be chosen carefully to ensure initial work is well-targeted:

Initial auto-formalisation targets

Consider business need and tractability to determine most valuable initial auto-formalisation targets. We are currently assessing internal (business processes), regulatory (compliance) and customer (service, sales) domains.

Modalities

Determine the relative importance of the distinct modalities in which auto-formalisation methods could apply, such as decision audit (did this historical process adhere to guidelines?), real-time decision making (accurate live interactions, via chat or API), code review (does this internal tool respect regulatory requirements?), and strategic planning (is this new product design consistent with the relevant policy documents?).

Assurance targets

Define assurance targets and success criteria for initial systems, including the level of correctness and error tolerance, extent of human oversight required, and overall coverage required for deployment in business contexts.

Policy documents

Within these targets, identify specific policy documents as the focus of initial work.

Datasets

High-quality datasets will be essential to validate capabilities.

Real world datasets

Identify real-world datasets for guiding and evaluating auto-formalisation system development. These could include chat transcripts, audit/decision logs, and software code.

Public benchmarks

Assess the value and coherence of existing public benchmarks such as taubench-2 retail and airline, and taubench-3 banking. Preliminary formalisation has already surfaced many inconsistencies in taubench policies, correcting these (and developing our own public benchmarks) will be important early deliverables.

Synthetic datasets

Generate high-quality synthetic datasets covering relevant business areas. Various approaches currently under review include few-shot generation by a language model, symbolic generation via a simulator, or retro-generation from a formalisation (a Proof Relevant pipeline output).

Novel benchmarks

Create, evaluate and if approved publish novel benchmarks for the challenge domain the autoform group is addressing: hard cases drawn from real regulation, on which current frontier models still make errors.

Evaluation

Create approaches for auto-formalisation techniques to be effectively evaluated, with a focus on maximising business value for the regulated financial sector.

Error categories

Analyze performance across three error categories: formalisation (flawed logical encoding), deduction (flawed logical reasoning) and interaction (flawed interfacing with the logical artefact). Together with genuine ambiguity, these are the principal sources of uncertainty in the pipeline, and evaluation should attribute failures to their source.

Faithfulness to source policy

Develop methods to establish that a generated formal artefact faithfully represents the source policy, including auditability/traceability from formal rules to clauses in source policy documents, and systematic testing (including boundary cases).

Elicitation

Consider whether the auto-formalisation system can perform necessary elicitation, where supplied facts are insufficient to determine validity and further questions must be asked.

Ambiguity analysis

Evaluate performance in the presence of genuine ambiguity. Consider whether this should trigger human-in-the-loop and/or refinement of the logical model, and under what conditions.

Business integration

Identify and scope opportunities for real-world trials that will help quantify the transformative impact that auto-formalisation methods can have in high-value LLM-powered business workflows (e.g. converting specific natural language policies into API implementations).

Strategic outcome: Machine-checkable rules derived from selected policy documents, supported by evaluation methods and audit trails, ready for trial in business workflows.

Workstreams in development

Multi-agent systems

Key Purpose

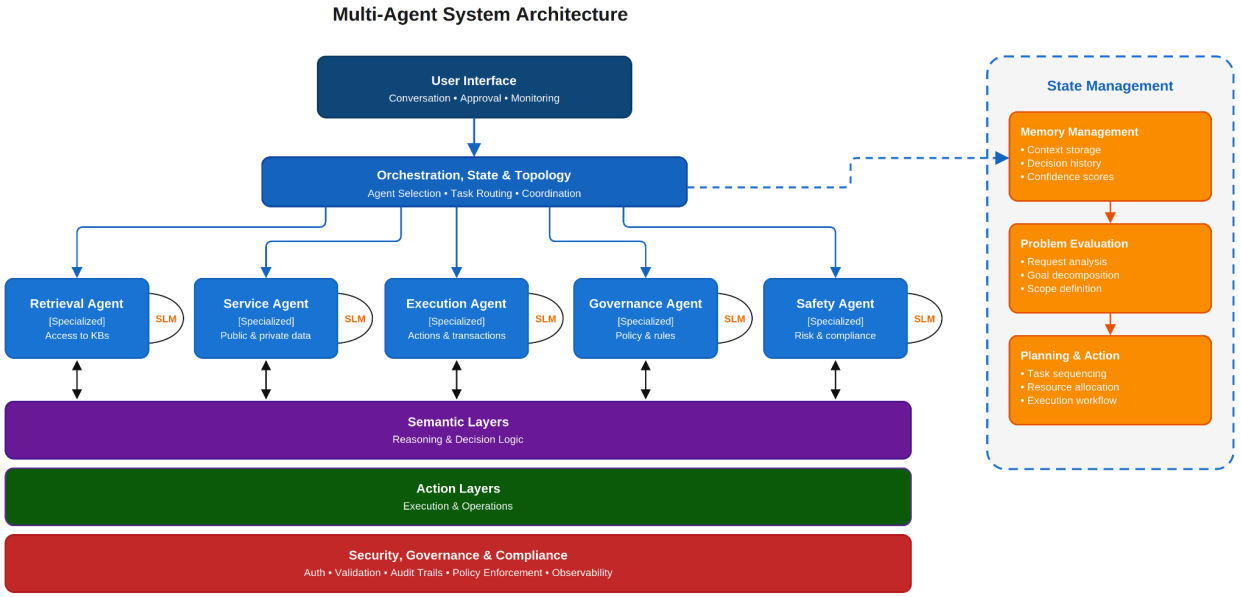
Providing the standards, governance, monitoring and traceability needed to deploy multi-agent AI in a trustworthy, compliant way.

Note: This section is at an early stage. The concept and architecture below are illustrative and will be refined as the programme progresses.

Concept

Multi-agent systems use a team of specialised AI agents working together on complex business problems. In regulated sectors, this coordination must be predictable, transparent and verifiable. Each agent operates within set boundaries covering its domain, data access and permitted actions, and communicates through a governed coordination layer, so every decision is recorded and can be traced back to the agents and data behind it.

Example architecture



Users interact through an interface supporting conversation, approvals and monitoring. A coordination layer breaks requests into tasks, routes them to the right agents, combines the results and checks actions against governance rules before they are carried out.

Specialised agents, each powered by a small domain-specific model, handle tasks such as retrieving information, managing data, carrying out transactions, applying policies and assessing risk. Shared layers support memory and planning, reasoning and execution, all underpinned by security, governance and compliance controls that maintain audit trails, route high-risk or ambiguous decisions to human review, and automatically intervene if unsafe behaviour is detected.

Strategic alignment

Multi-agent systems bring together work from across the programme: agents use models built through Model Fabric, enforce policy rules from Auto-formalisation, apply Panopticon's controls, follow boundaries set in Domain definition, and feed activity data into Observability. The result is a system where every action is logged and traceable, high-stakes decisions reach human review, and operations continue safely if individual components fail.

Cross Section Areas

The following sections apply across all work streams As the programme progresses, they will be developed further, and work streams may be formed to take some of them forward.

Domain definition

Key purpose: Creates a set of business concepts, rules, processes, data and behaviors related to a specific area selected by the programme.

Concept

Ensures that business objectives, domain boundaries, success criteria and data ownership are clearly defined and understood.

Domain modelling: Language models stand out for how they model the world reflected in their training data, but enterprise systems need consistent, trustworthy outputs, so this unpredictability must be controlled. By clearly defining what a model can do and the information it has access to, many failure modes can be avoided. Although model outputs remain probabilistic, a well-defined scope is expected to reduce variability to a practically acceptable level.

A clearly bounded scope also makes testing more practical. Rather than testing a general model against every possible input, targeted tests can cover the known scenarios within the domain, measuring accuracy, consistency, safety and more. Once a model reliably meets these criteria, it can be deployed with confidence. In high-stakes settings such as finance, a model could also be set to answer only when sufficiently confident, otherwise referring the question to a person or another system.

Constraints: Restricting how a model responds, for example by requiring answers in a fixed format or a simple, purpose-built language, acts as a guardrail that keeps it within valid logic. The model's task shifts from generating free-form answers to filling in a defined structure, making its outputs more precise, consistent and trustworthy.

Strategic outcome: A clear definition of each business domain, setting out its information boundaries, decision rights, controls and success measures.

Synthetic data

Key purpose: Creates a shared foundation that reduces duplication and accelerates research and development across the programme.

Concept

Data sourcing, preparation and curation: Establishing repeatable processes for acquiring, cleaning, validating and versioning training data, drawing on both public structured and unstructured sources.

Synthetic data generation: Defining approaches for generating realistic domain data where needed, including expanding existing datasets, creating privacy-preserving data and generating rare or unusual scenarios.

Both approaches remain subject to further research and validation.

Strategic outcome: High-quality, reusable datasets that meet defined standards for accuracy, completeness, consistency and compliance.

Security

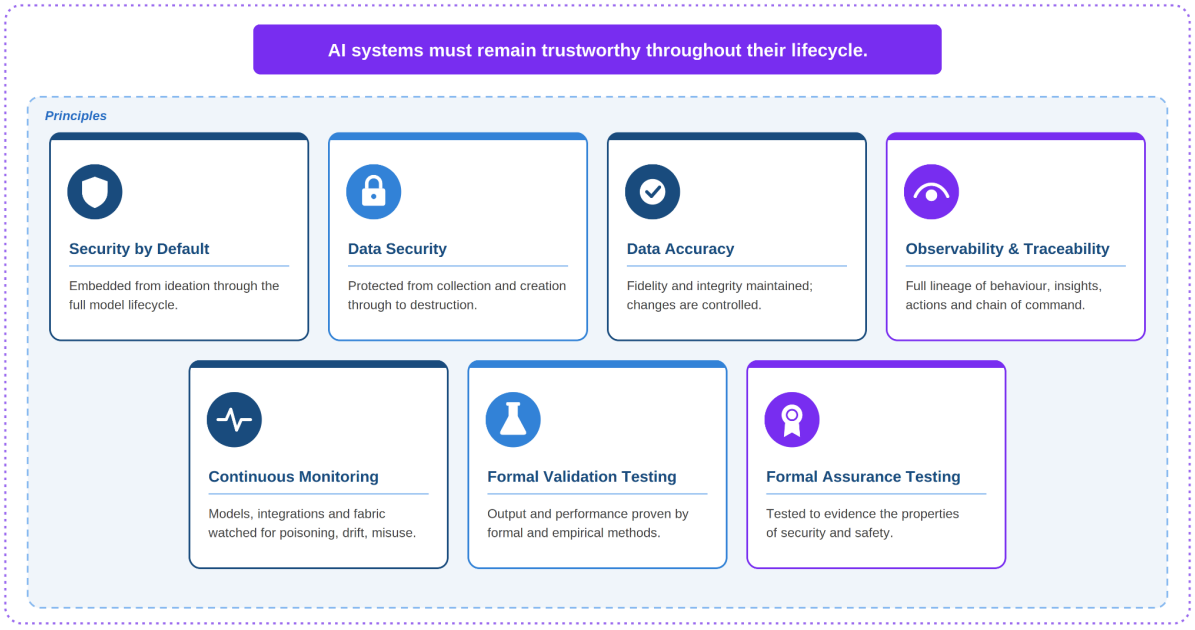
Key purpose: Underpins every other area of the programme.

Concept

Securing AI requires safe and scalable adoption through a holistic security approach that strengthens model resilience, protects sensitive data, secures deployment and integration architectures, and ensures governed, authorised, and accountable use of AI capabilities. High level principles include:

- **Secure the Model** - Ensure the model is robust to different types of LLM 'attacks' (including as they evolve).
- **Secure the Data** - Ensure robust data controls are in place for any deployed LLM.
- **Secure the Operation** - Ensure the design and deployment of the model are secure (this includes the architecture pattern, deployment pattern and integrations).
- **Secure the Usage** - Ensure appropriate governance controls are in place (e.g. only authorized and expected use is made of any deployed LLM).

Principles



Strategic outcome: Trusted, secure and reliable AI capability.

Observability

Key purpose: Provides operating standards, design patterns and control frameworks that turn system signals into automated, governed action.

Concept

Observability transition from static dashboarding into an active intelligence engine - providing immediate, evidence-backed root cause identification to accelerate incident resolution.

Principles

- **Transparency and Explainability** - AI systems must provide clear visibility into their decision-making processes. Transparency should extend across data sources, model logic, configuration settings, and operational workflows.
- **Continuous Monitoring** - Implement real-time monitoring across the entire AI lifecycle, from data ingestion and model training through inference, integration, and output delivery. Monitoring should capture operational health, performance metrics, resource utilisation, latency, accuracy, and business outcomes.

- **Domain-Aware Evaluation** - Beyond technical metrics, leverage domain expertise to assess whether AI outputs meet business, regulatory, ethical, and quality expectations.
- **Proactive Detection** - Enable early identification of performance degradation, data drift, model drift, bias, anomalies, and unexpected behaviours before they impact end users.
- **End-to-End Traceability** - Maintain complete audit trails from input data through model processing to final outputs, supporting both debugging and compliance requirements. Traceability should provide visibility into data lineage, model versions, prompts, configurations, decisions, and downstream actions.
- **Reliability and Resilience** - AI systems should be designed to operate consistently under varying conditions while maintaining acceptable levels of performance, availability, and security.

Strategic outcome: Transparent, explainable, measurable and continuously improvable AI ecosystem.