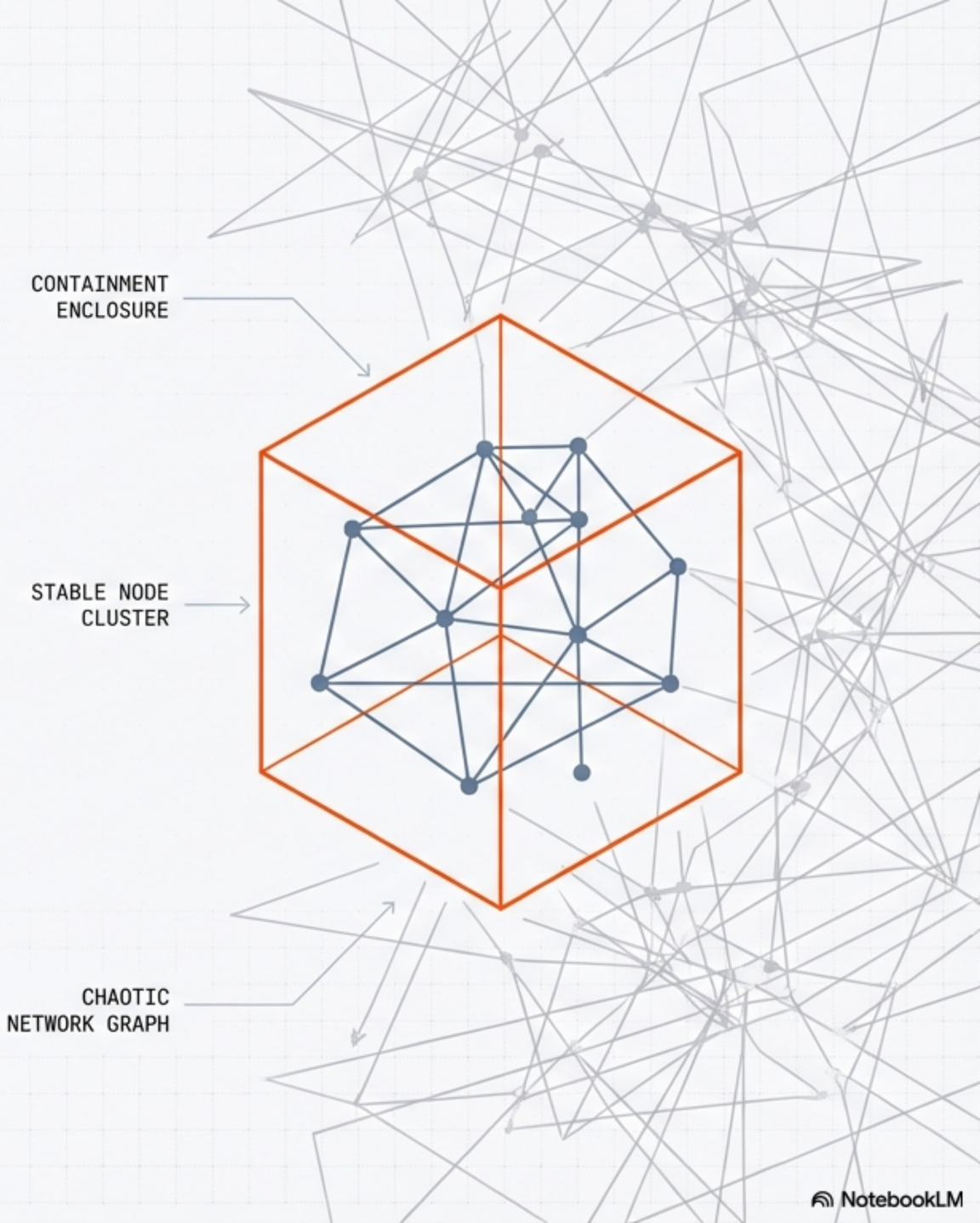
ARCHITECTING TRUST: THE AI EXECUTION CONTROL PLANE

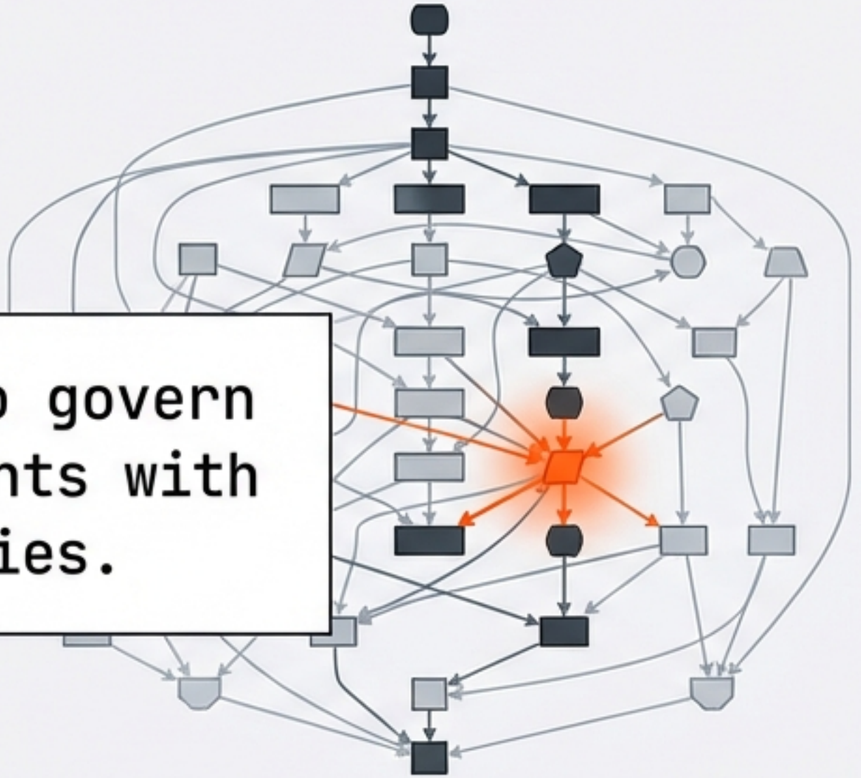
Moving from Policy to Enforcement in the Age of Autonomous Agents.



AERIE PROJECT OVERVIEW
PHASE 0 // CANONICAL REFERENCE

THE UNCOMFORTABLE TRUTH

We are attempting to govern probabilistic AI agents with static PDF policies.



THE POLICY

governance_doc.pdf

Focus: Trust, Alignment, Safety Promises

Method: Static Text

THE REALITY

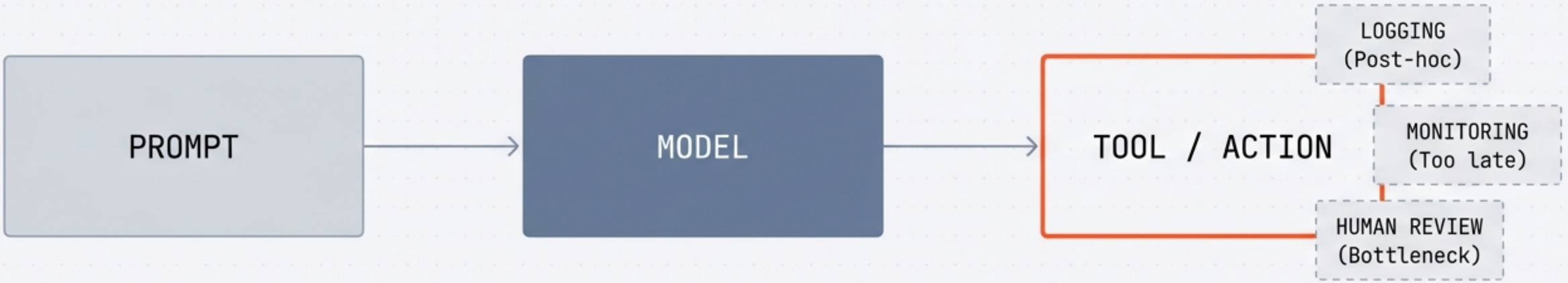
runtime_execution.log

Focus: Probabilistic Action

Method: Non-deterministic Output

THE CURRENT MODEL IS BROKEN

Conflating intelligence with authority creates unmanageable risk.



“We are auditing behaviour we should have made impossible.”

Vulnerability: Reliance on prompt injection resistance or model 'choice' for security boundaries.

THE REGULATORY DEMAND

Global frameworks are converging on requirements that prompts cannot satisfy.

REQUIREMENT: AUDITABILITY

Source: EU AI Act / NIST

Not just unstructured logs, but a reconstructable history of intent and authorization. The ‘Why’ must be as visible as the ‘What’.

REQUIREMENT: HUMAN OVERSIGHT

Source: EU AI Act Article 14

Explicit authorization mechanisms. The system must not act without a traceable grant of permission.

REQUIREMENT: LEAST PRIVILEGE

Source: NCSC Guidelines

Agents should have zero access by default. Capabilities must be issued per task, not granted globally.

REQUIREMENT: TRACEABILITY

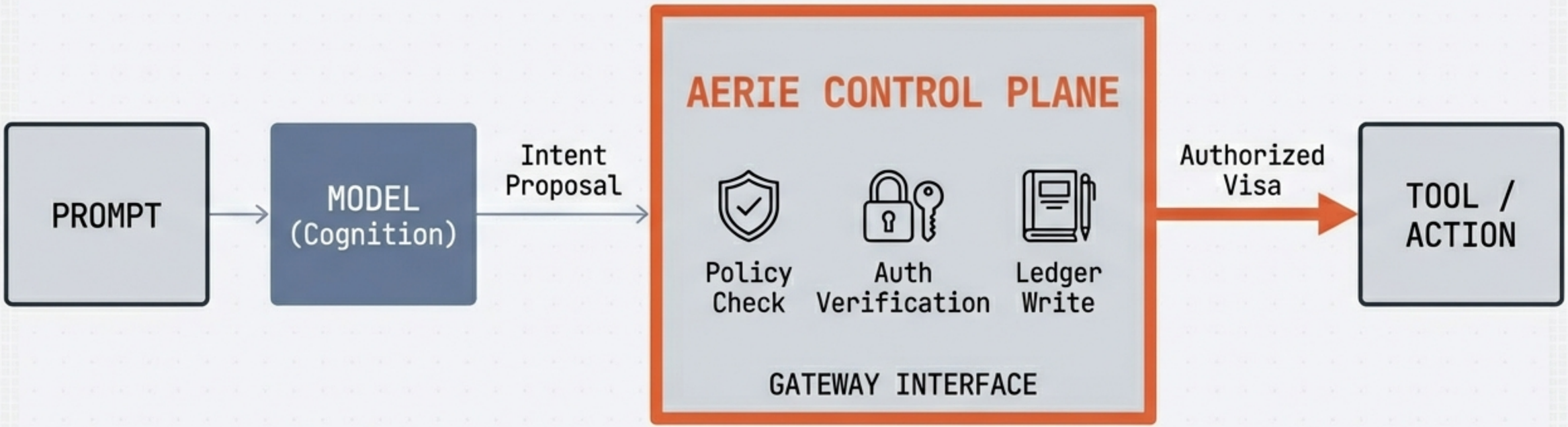
Source: ISO 42001

Linking specific actions back to specific, authorized intent packets. No ‘hallucinated’ actions.

Regulators aren’t asking for better system prompts. They are asking for a control layer.

THE SOLUTION: THE MISSING LAYER

Introducing the AI Execution Control Plane.



DEFAULT STATE: DENY ALL

The agent has zero authority by default. It proposes; the Control Plane disposes.

MEET AERIE

Architecture over prompts. Authorisation over alignment.

// WHAT IT IS

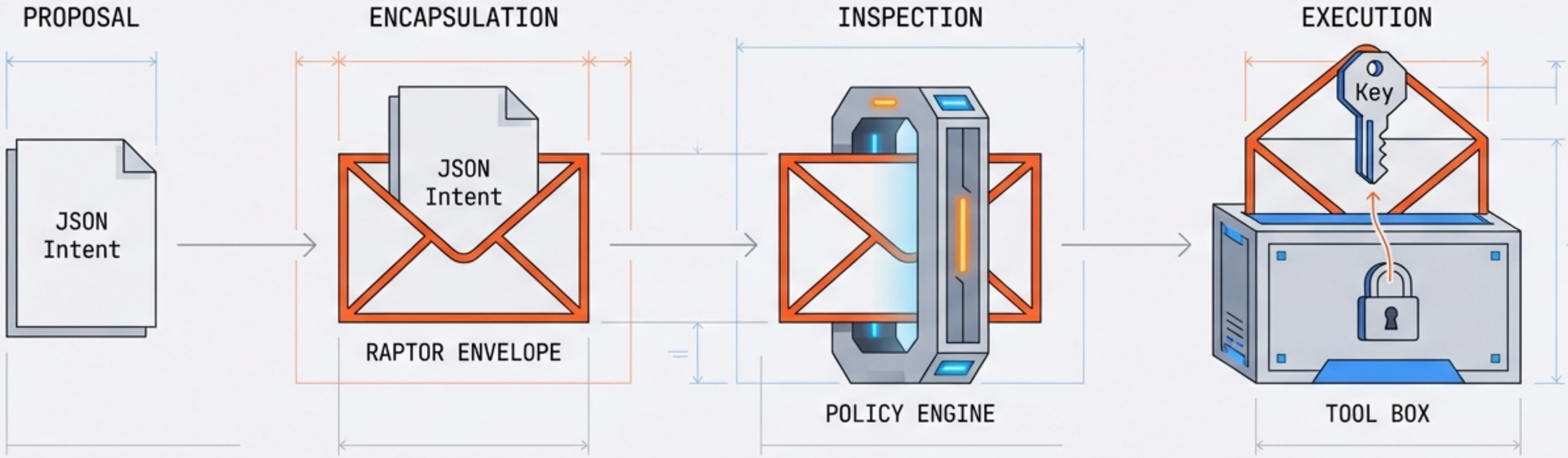
- [+] A control and governance plane.
- [+] Model-agnostic architecture.
- [+] A human-authorised execution layer.

// WHAT IT IS NOT

- [-] A chatbot or copilot.
- [-] An LLM wrapper.
- [-] A prompt engineering library.

PROVABLE BEHAVIOUR OVER EXPLAINABILITY THEATRE.

MECHANISM: THE RAPTOR GOVERNANCE ENVELOPE



The model does not execute. It requests permission to execute.

THE IMMUTABLE LEDGER

Audit-first design. History is not a side-effect.

RECONSTRUCTION VS. GUESSING

Actions are reconstructed from these records, not inferred from chat logs or model weights.

If an action cannot be traced to a human-authorised intent record, it does not happen.

REGULATION BY DESIGN

Compliance emerges naturally from the architecture.

REGULATORY REQUIREMENT	SOURCE	AERIE IMPLEMENTATION
Human Oversight	EU AI Act (Art. 14)	Explicit Intent Authorisation
Risk Management	NIST AI RMF	Least Privilege Capability (Zero Trust)
Secure by Design	NCSC Guidelines	Hard Gated Architecture
Traceability	ISO 42001	Immutable RAPTOR Ledger

NOT A BOLT-ON.

OPTIMISING FOR TRUST VS. AUTONOMY

STANDARD FRAMEWORKS

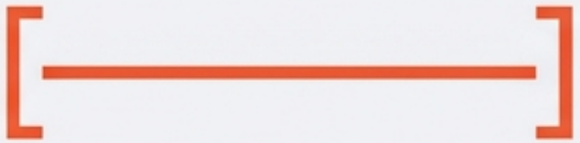
> Optimise for: Speed & Autonomy

> Nature: Improvisational, Self-healing

> Feeling: 'Magic'

> Question: Can the agent do this?

AERIE ARCHITECTURE



> Optimise for: Trust & Control

> Nature: Deterministic, Gated

> Feeling: 'Boring' (Secure)

> Question: SHOULD the agent do this?

Where ambiguity exists, the system must halt—not infer.

SCOPE & NON-GOALS

Security requires clear boundaries. AERIE is the boundary, not the intelligence.

OUT OF SCOPE

- [x] An autonomous agent framework
- [x] A prompt-driven governance solution
- [x] A 'Self-healing' system
- [x] A replacement for cognitive intelligence

CRITICAL REASONING:
If a feature requires trusting model behaviour (like self-correction), it is out of scope for a control plane. We do not trust the model. We constrain it.

IN SCOPE

> Defining and enforcing explicit boundaries and constraints.

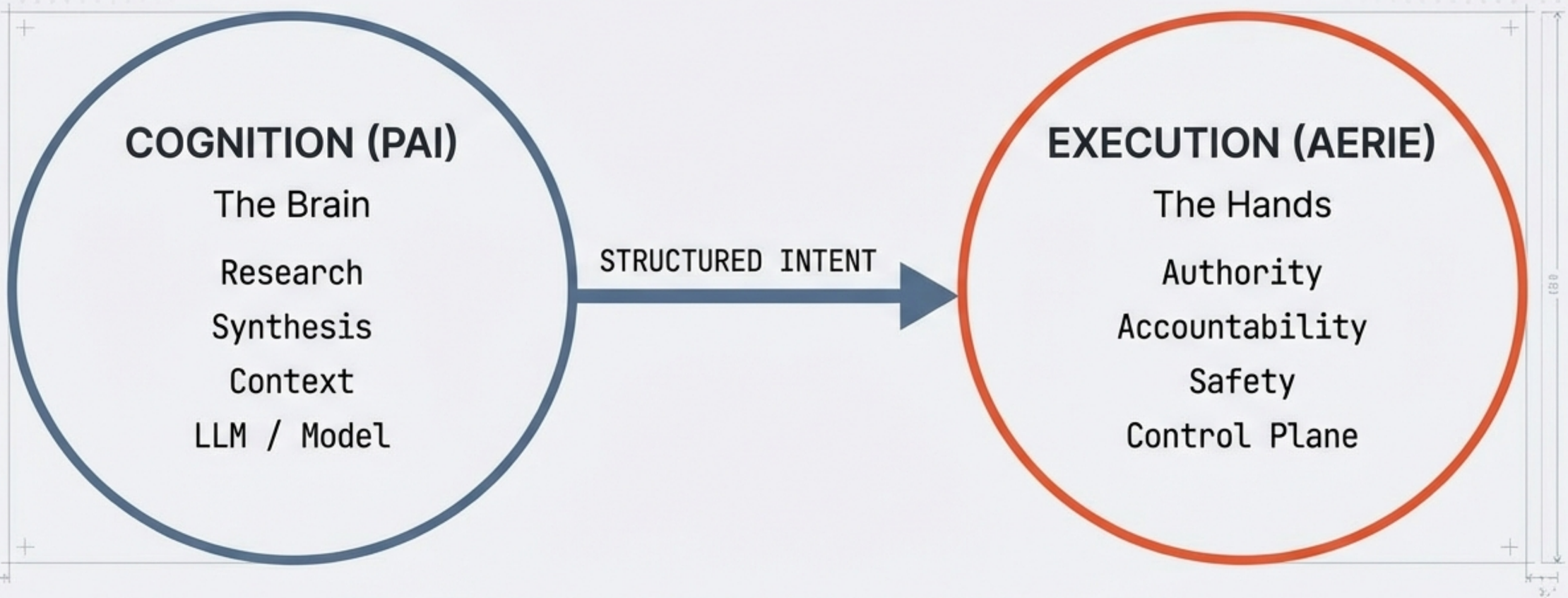
> Providing a control plane for deterministic operations.

> Hard-gating and explicit authorization for all actions.

> Emphasizing trust through limitations, not capabilities.

SEPARATION OF CONCERNS

Cognition vs. Execution.



PAI produces intent. AERIE governs action.

TARGET AUDIENCE & SUCCESS CRITERIA

PRIMARY AUDIENCE

SECURITY ENGINEERS & ARCHITECTS	GOVERNANCE PROFESSIONALS (Finance/Gov)	RESEARCHERS (Safe Execution)

DEFINITION OF SUCCESS

- [ ] Execution authority is explainable without referencing prompts.
- [ ] Actions are independently reconstructable from the ledger.
- [ ] Models can be swapped (e.g., Llama -> GPT-4) without altering governance guarantees.

THE FUTURE OF AGENT ARCHITECTURE

IF AI IS GOING TO ACT IN THE REAL WORLD, WE NEED TO STOP TRUSTING IT—AND START CONSTRAINING IT.

AERIE is an architectural pattern for
the post-policy era.

STATUS: Phase 0 (Canonical Reference) **LICENSING:** RAPTOR-Aligned (Dual License) **CODE:** github.com/cyb3run1c0rn/aerie