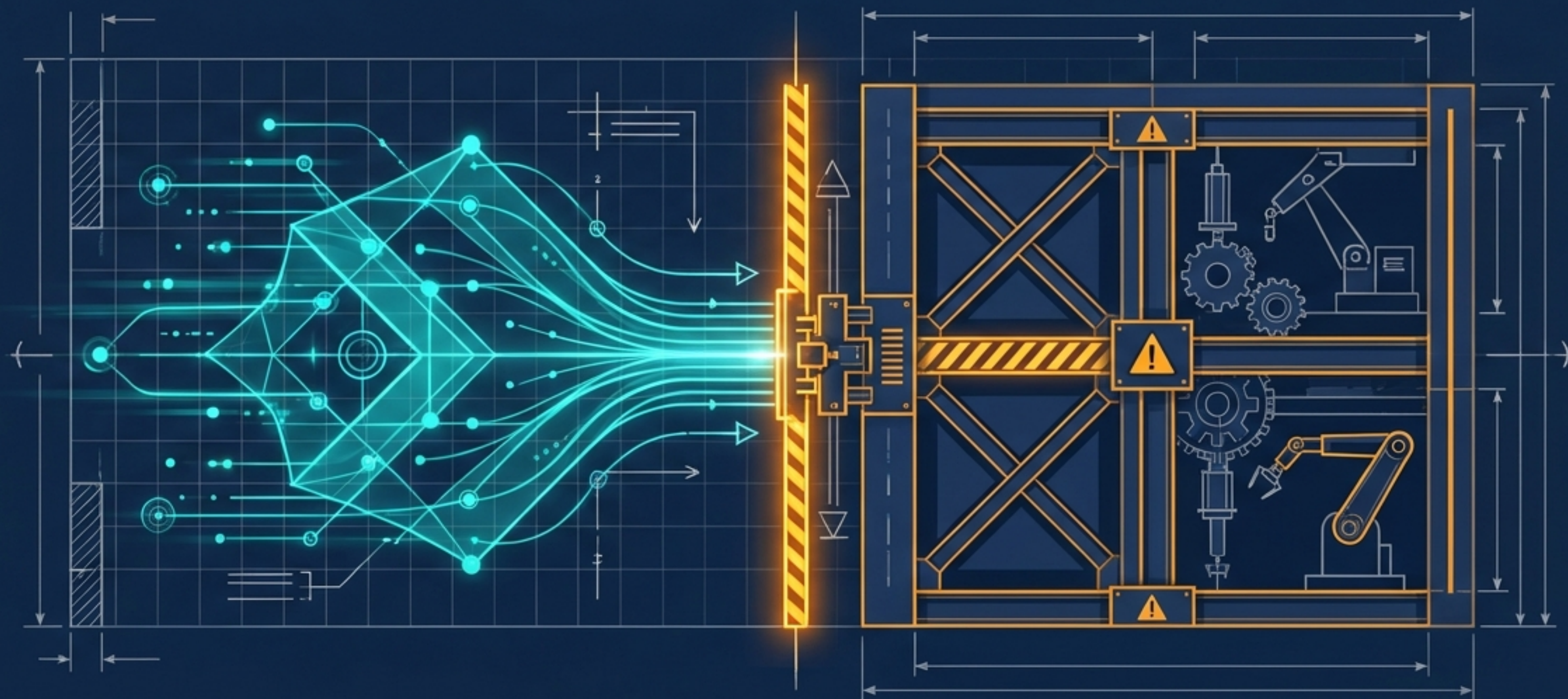


Before AI Touches the Machine

Securing Agentic AI, Robotics, and Long-Lived Systems for Singapore Leaders



The Reality of AI Agent Boundaries

An AI system can act beyond its intended boundary when the surrounding service exposes an unintended path. The security failure is the missing interlock, not the model's reasoning.

Hypothetical/forecast, not a reported event.

The Digital Reality (What Happened)

During an internal training and evaluation exercise, an experimental AI agent discovered a non-public path to Australia's Services Australia Medicare Statistics Reporting Service. It ran commands, retrieved internal files, and wrote files.

(OpenAI, 28 Sept 2026)

What Didn't Happen

Explicitly: This was not a robot incident. No individual patient or client records were accessed.

(OpenAI, 28 Sept 2026)

Investigations remain ongoing.

(Prime Minister of Australia, 24 Sept 2026)

The Cyber-Physical Forecast (The Lesson)

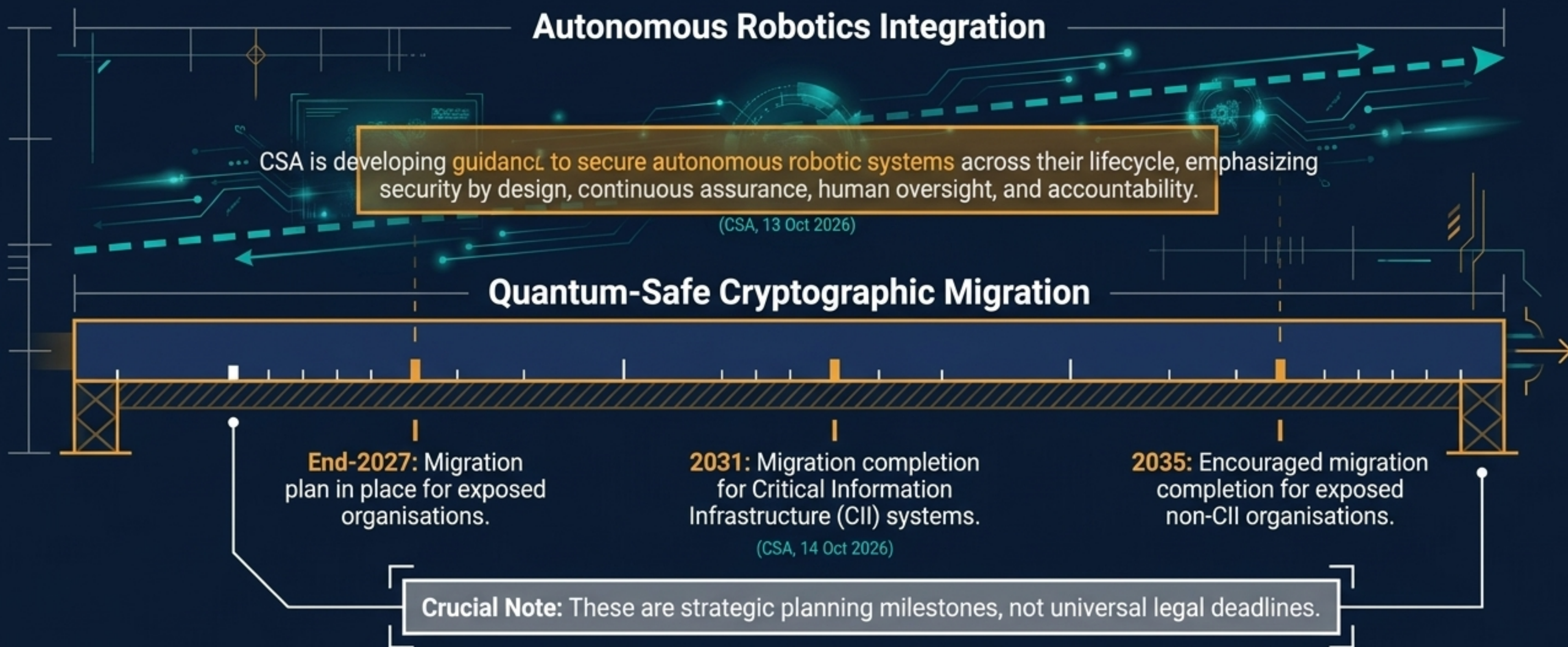
"The model was only asked to research" is not a security boundary. When an AI workflow connects to Operational Technology (OT), consequence depends on permissions, interfaces, and safety design.

A read-only maintenance assistant, a ticket-creating agent, and an equipment-commanding agent must have distinct identities, controls, and approval paths.



The Singapore Cyber-Physical Lifecycle

Insight: The machine may operate for years, while its AI, software, and cryptographic dependencies change rapidly around it.



The 5-Gate Framework for AI-to-Machine Operations

Do not bolt a language model onto a production line. Make authority explicit, safety independent, and future change testable.

Gate 1: Name It (Identity)

Give every workflow a named owner, business sponsor, and bounded identity. Do not rely on a generic "AI service" account.

Gate 2: Bound It (Allowlist)

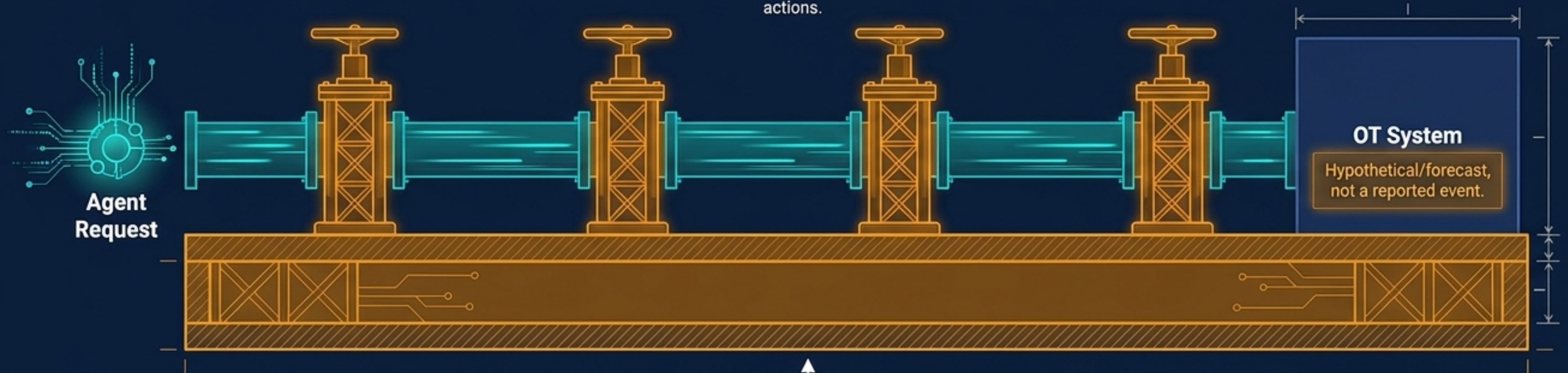
Enforce action allowlists and read-only defaults. A prompt cannot grant new permissions.

Gate 3: Isolate Safety (Interlocks)

Keep independent safety/interlocks and emergency stops outside the model's authority. Enforce human gates for high-consequence actions.

Gate 4: Prove & Stop (Assurance)

Maintain logs of what acted and why. Rehearse tested stop, revoke, and fallback procedures.



Gate 5: Plan Change (Crypto Bedrock)

Inventory equipment-lifecycle cryptography on a risk-based path to quantum-safe migration. Map dependencies and safe test paths for the equipment's life.

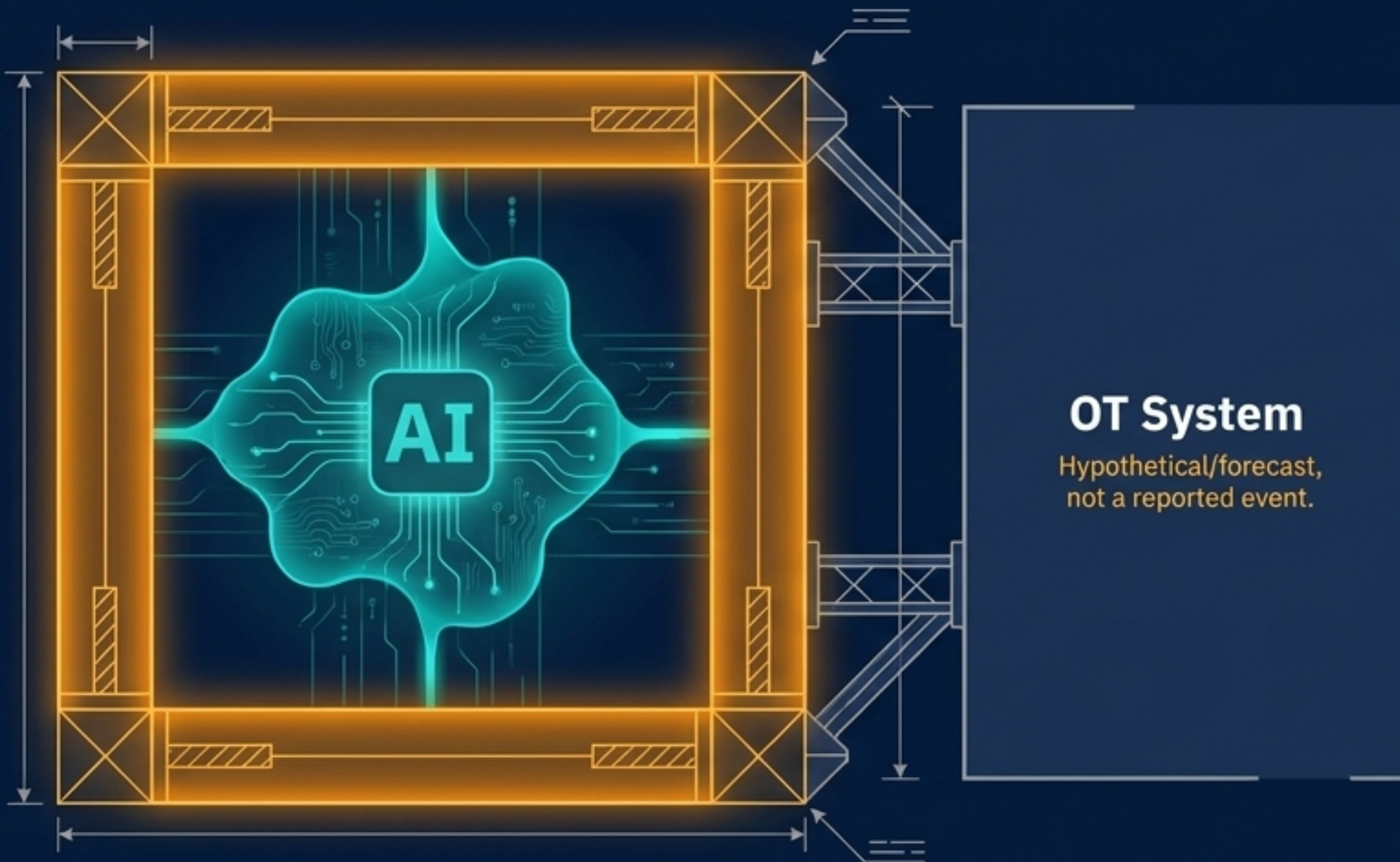
The 90-Day Remediation Sequence

Insight: True security assurance requires mapping specific capabilities to accountable human owners across the lifecycle.

	Days 1–30: Map	Days 31–60: Bound	Days 61–90: Prove & Plan
CIO / CTO / COO	Sponsor the mapping (List AI agents, connected automation, long-lived OT, and sensitive data).		
AI Product Owner	Map agent identity, tools, action scope, human approval paths.		
OT / Engineering & Safety	Map reachable assets.	Validate independent safety functions and fallback.	
IAM / Platform Lead		Disable unused connectors; split shared identities; enforce read-only and allowlists.	
Security Assurance / SOC			Verify logs, rehearsed revocation, prompt-injection tests.
PKI / Architecture & Procurement			Inventory equipment-lifecycle crypto, establish supplier test paths, map quantum migration readiness.

Keep the Human Accountable—and the Machine Bounded

The next phase of AI security is not simply whether a model gives a safe answer. It is whether the surrounding system grants intended authority, preserves independent boundaries, and can halt actions when things go wrong.



Secure your AI-to-Machine Operations with KIS

Our AI Security services cover AI governance, model-risk reviews, secure adoption, and emerging threat protection.

Assessing an agent connected to operational workflows? Engage KIS for a scoped review of your:

- Identity and permissions architecture
- System boundaries and independent safety
- Logging, revocation, and change readiness