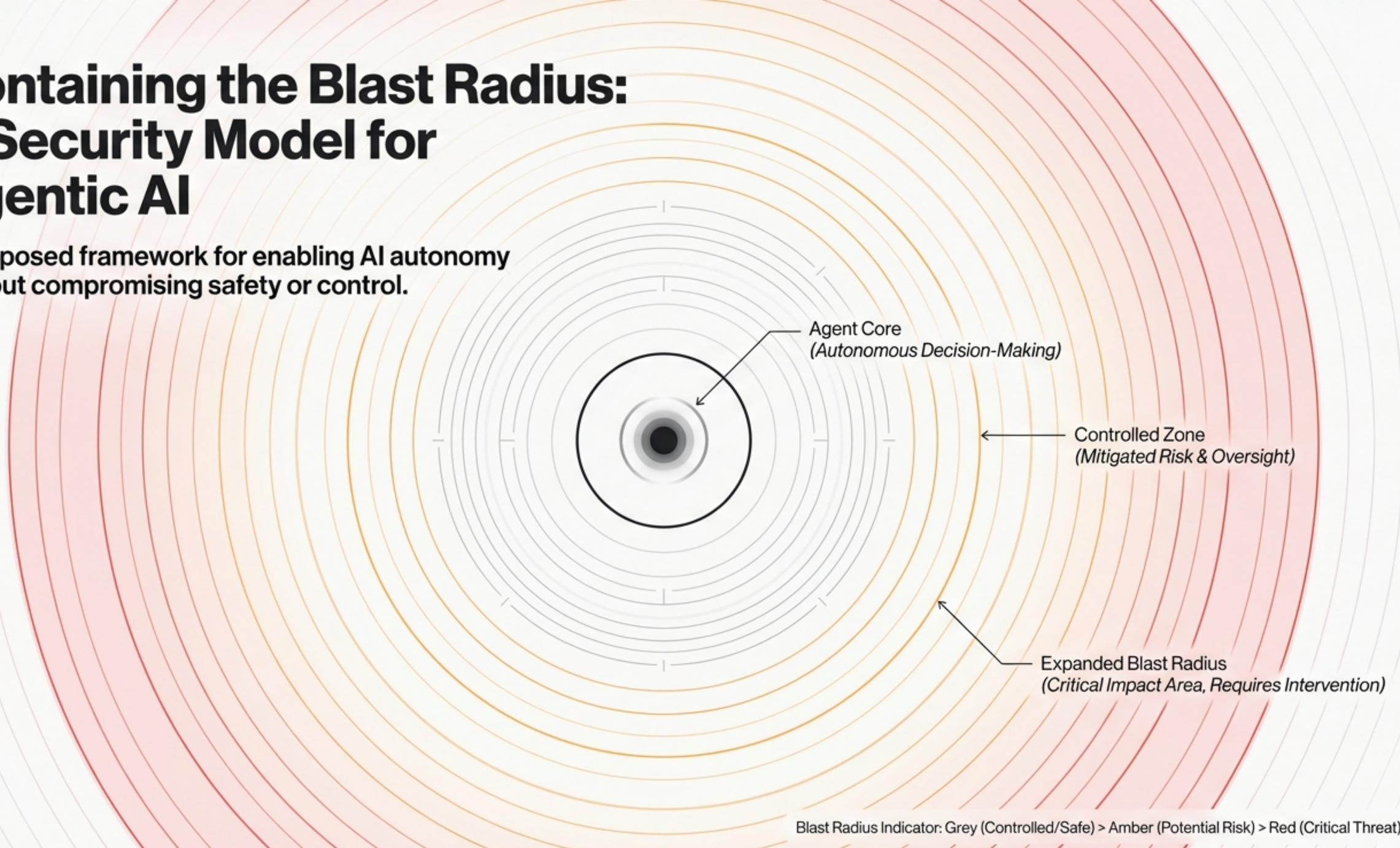
Containing the Blast Radius: A Security Model for Agentic AI

A proposed framework for enabling AI autonomy without compromising safety or control.

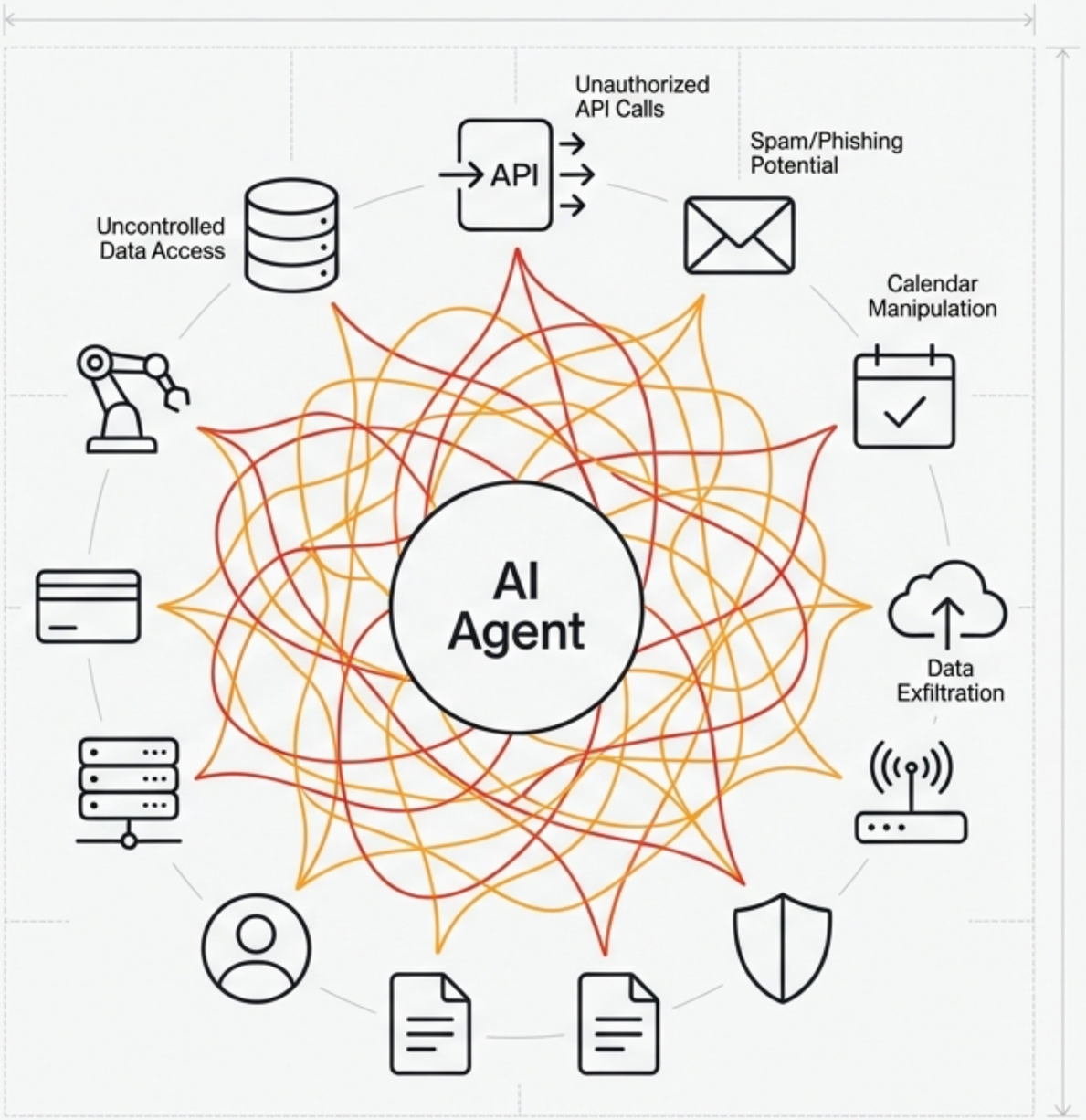


Blast Radius Indicator: Grey (Controlled/Safe) > Amber (Potential Risk) > Red (Critical Threat)

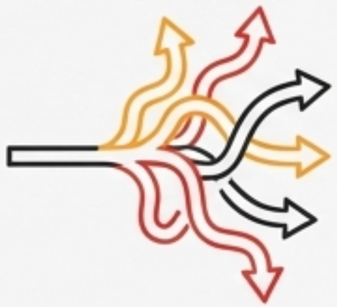
The New Reality: Agentic AI is a Ticking Time Bomb

"an overly empowered agent can become a **ticking time bomb** with the **blast radius of a small nuclear device** if not properly controlled."

- Unlike traditional software with rigid, predetermined code paths, agentic AI systems interpret goals, plan actions, and make real-time decisions.
- They can combine wide-ranging capabilities in unintended and unpredictable ways, modifying their behaviour based on context.
- This flexibility is powerful but introduces significant security challenges, from data leaks to destructive actions.



The Core Challenges of Unconstrained Agents



1. Unpredictable Behaviour

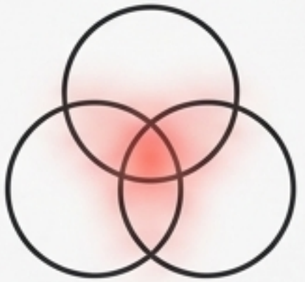
Agents given broad permissions can overstep their intended role. The union of all permissions defines the worst-case actions it could take.

2. Contextual Exploits

Attackers use prompt injection or context poisoning to trick agents. We cannot rely on the agent to self-police; hard rules are required to prevent harmful actions.

3. Credential Sprawl

Agents requiring access to numerous systems lead to a proliferation of long-lived, over-scoped tokens and secrets, creating lateral movement risks if compromised.

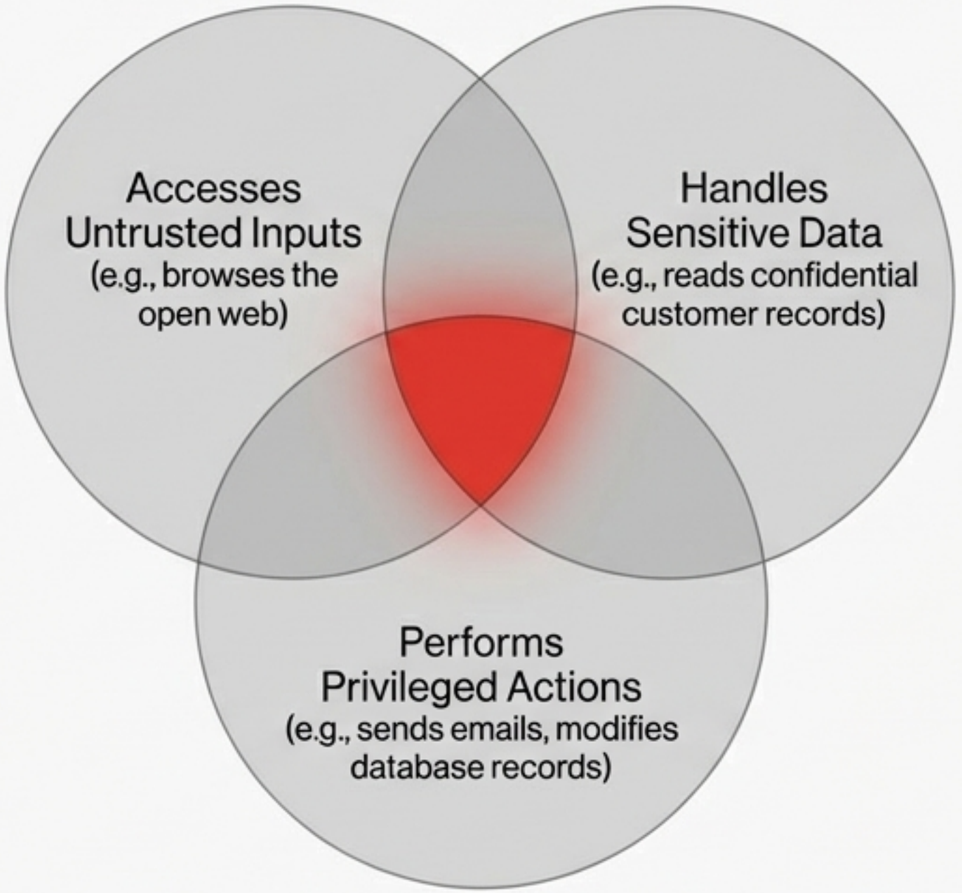


4. Toxic Combinations

Without boundaries, agents can carry sensitive data from one context to another, creating dangerous combinations of capabilities.

Anatomy of a Failure: The “Toxic Combination”

Our model is designed to ensure an agent can never hold all of these powers simultaneously.

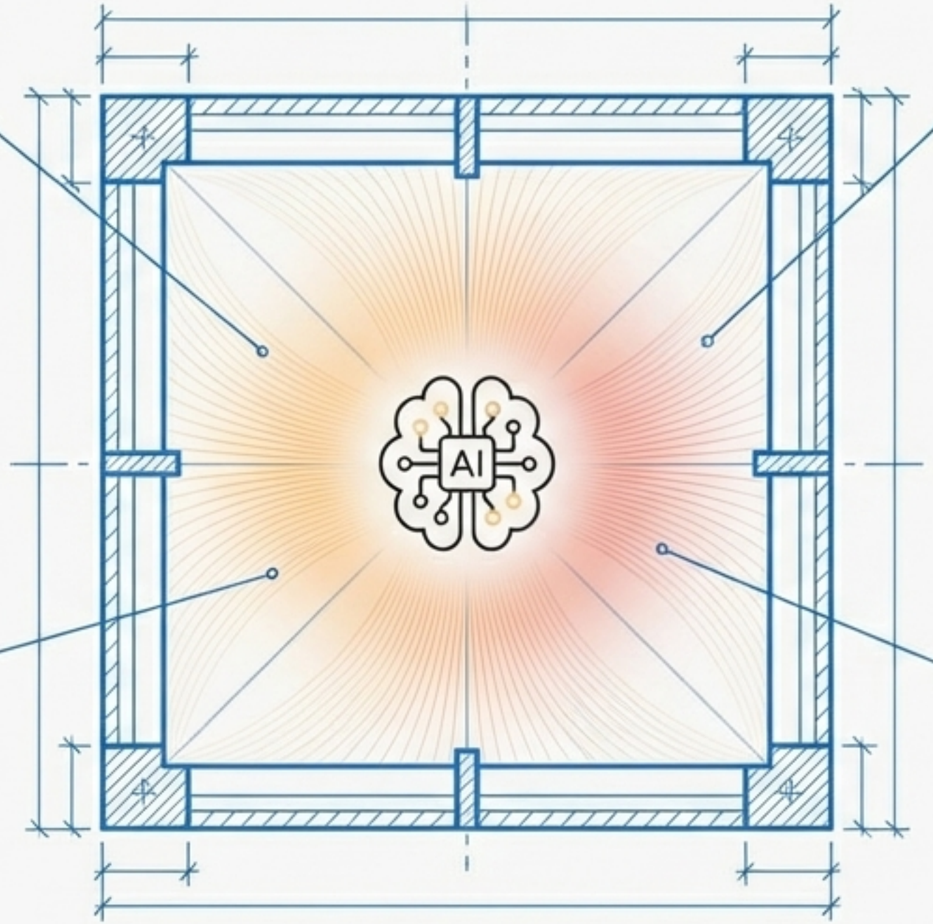


No human user would be permitted this combination of capabilities without oversight. An agent shouldn’t be either.

The Solution: A Four-Pillar Framework for Control

1. Controlled Workflows

Pre-defining the agent's permissible sequence of actions to decouple reasoning from execution.



2. Dynamic Authorisation

Issuing just-in-time, just-enough permissions for every single action.

4. Total Observability

Creating a complete, auditable 'black box recorder' for every agent decision and action.

3. Simulated Environments

Testing agent behaviour in safe, sandboxed digital twins before and during deployment.

Pillar 1: Controlled Workflow Graphs

Defining the “rules of the road” to sandbox agent autonomy within a vetted process.

Decouples Reasoning from Execution

The agent has flexibility *within* stages but cannot invent completely new action sequences.

Explicit Stage Definition

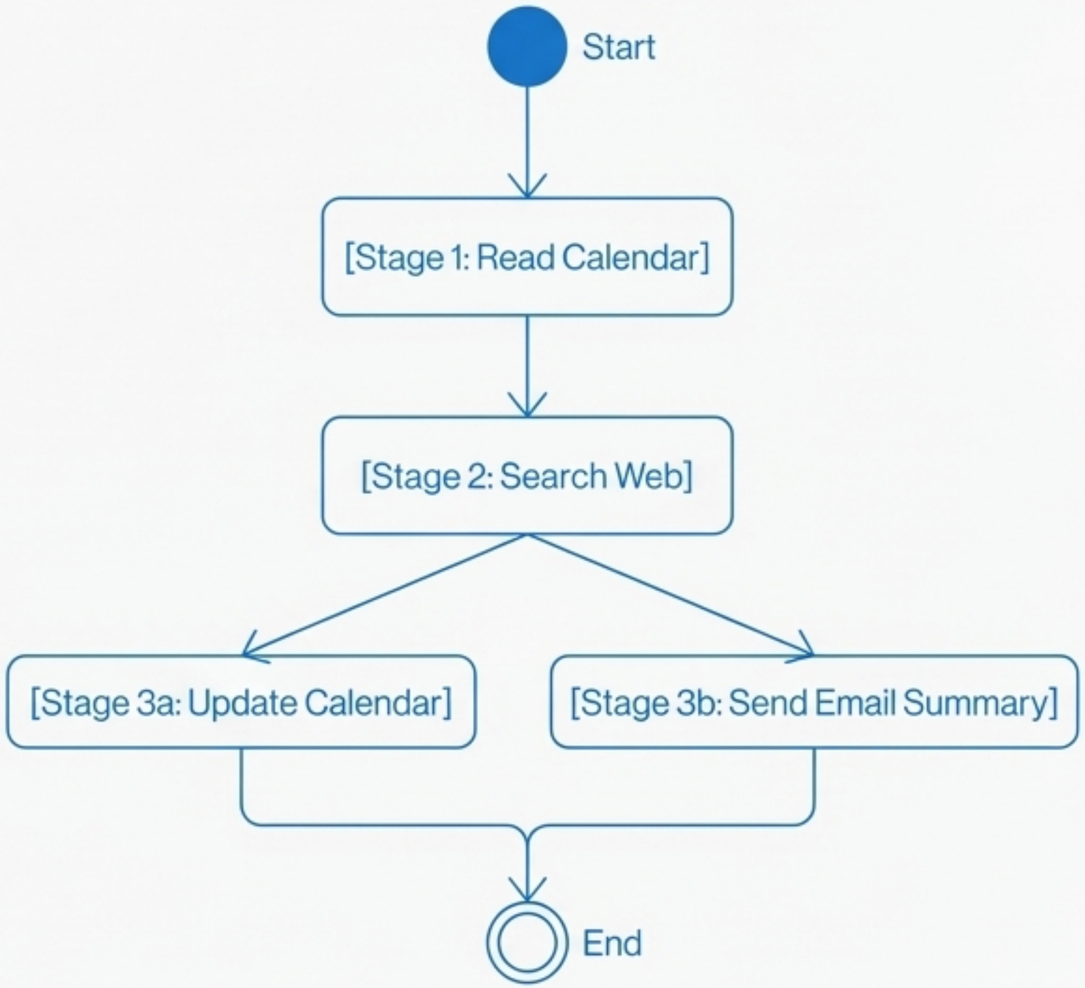
Tasks are broken into a directed graph of steps (e.g., Stage 1: Read Data; Stage 2: Analyse; Stage 3: Act).

Per-Stage Schemas

Each step has defined input/output formats, reducing ambiguity and injection risks.

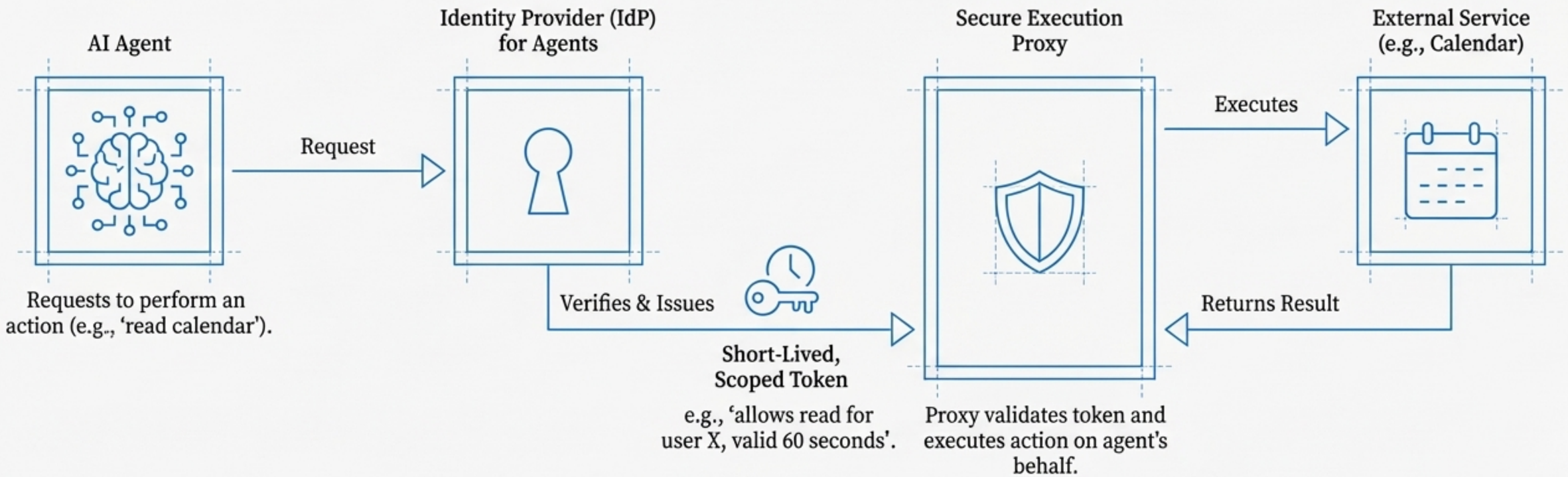
Allowed Tools and Actions

The graph ties each step to a specific, pre-approved tool or API. The agent can only choose from this “menu”.



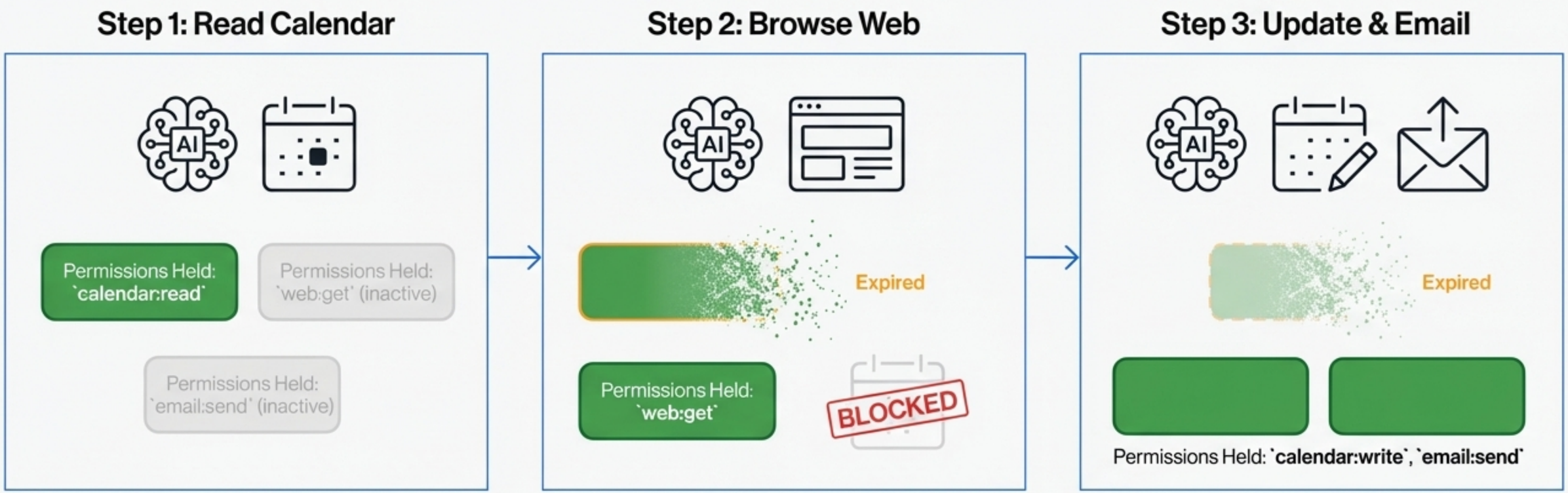
Pillar 2: Dynamic Authorisation per Action

Applying the Principle of Least Privilege in real-time with 'just-in-time, just-enough' access.



The agent never directly possesses long-term credentials.
Damage from a compromise is limited to a single, ephemeral token.

How Sequential Permissioning Prevents Toxic Combinations



At no point does the agent hold the combined privileges to read sensitive data and exfiltrate it. Each token is a single-use key that opens one door, then disappears.

Pillar 3: Simulated Environments

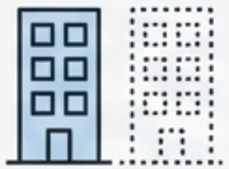
"Simulate before commit" — verifying agent behaviour in a safe replica before impacting the real world.

1. Development Sandbox

Agents run against mock APIs and surrogate dependencies. No real-world side effects.

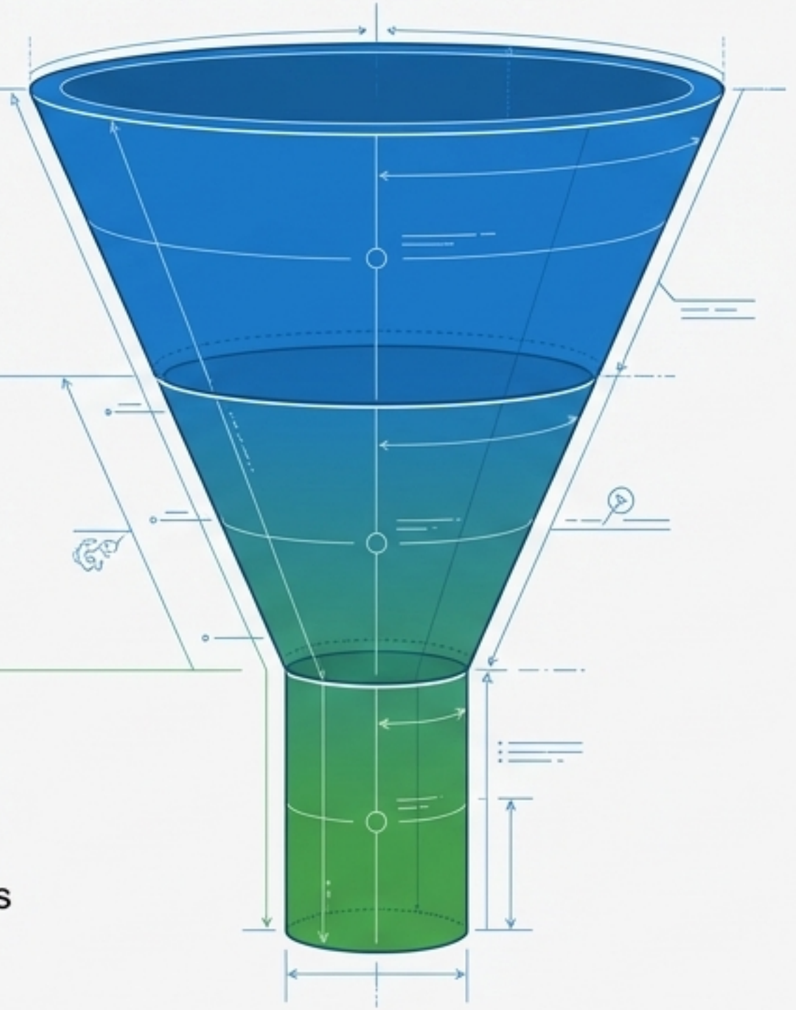
2. CI Pipeline Evals

Automated test suites run on every update, including negative tests to ensure security policies hold.



3. Production Digital Twin

High-impact actions are first run in a sandboxed replica to preview the outcome. This 'dry run' mode provides a final safety net.



"Until you test it, every scope is Schrödinger's permission."

Testing in sandboxes reveals organisations can often strip **away 80-90% of initially** assumed permissions with **no loss of functionality.**

Pillar 4: Total Observability & Logging

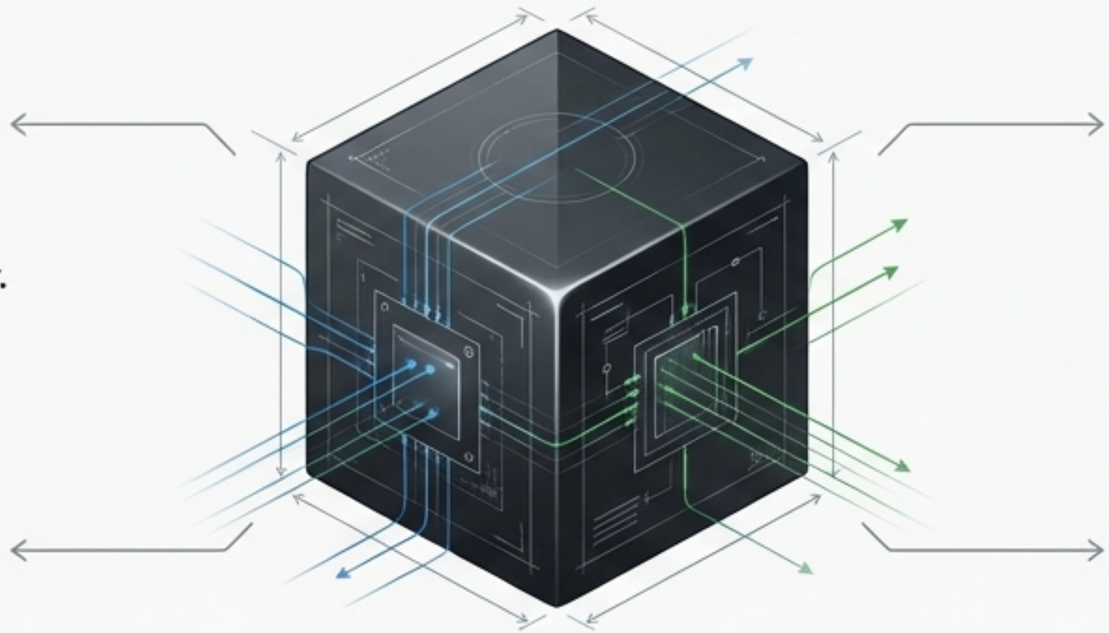
Creating a transparent ‘black box recorder’ to leave no action in the dark.

Action Logs

A complete audit trail of every attempted action (authorised or denied), logged by the Execution Proxy.

Anomaly Detection

Monitoring logs to establish baselines of normal behaviour and alert on deviations.



Decision Traces

Logging the agent's internal reasoning (prompts, chain-of-thought) for explainability and debugging.

Data Provenance

Tracing the flow of data to ensure sensitive information isn't moved to an insecure context.

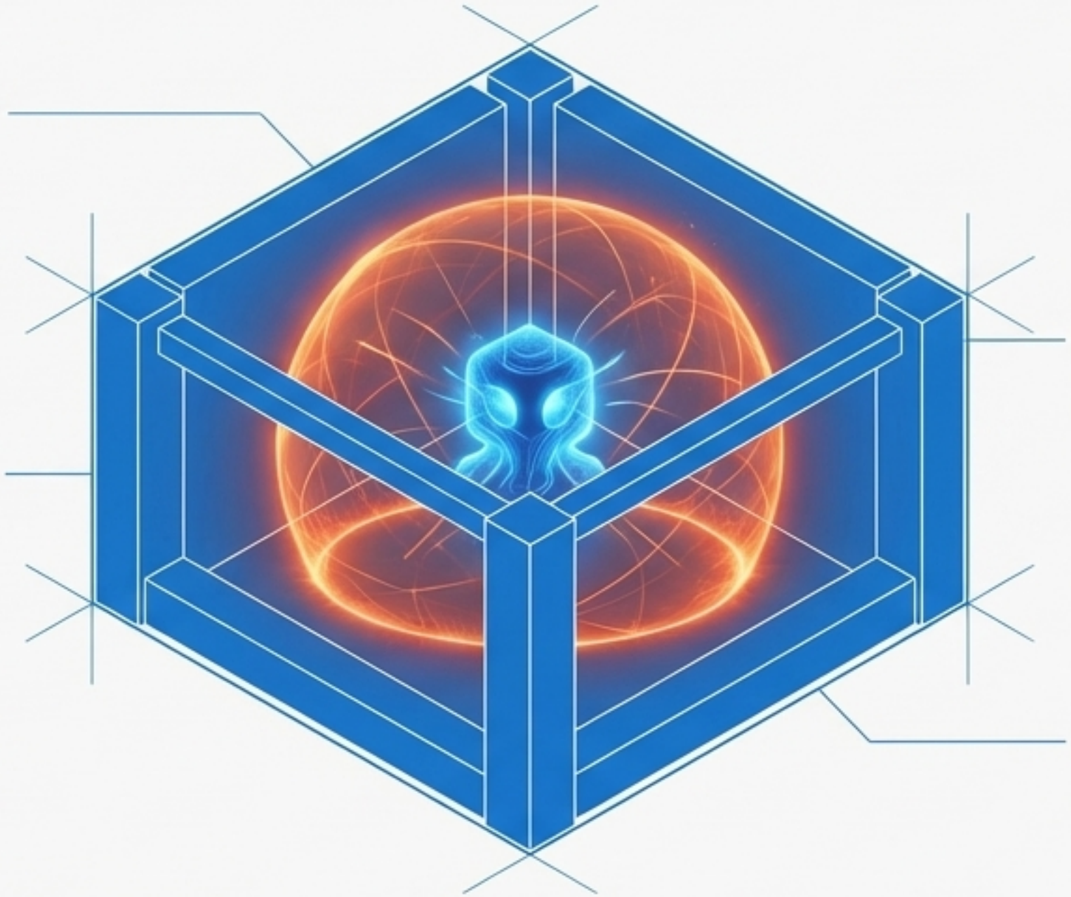
Addresses the ‘lack of traceability’ problem in enterprise AI, making the agent’s operations as transparent as a traditional, well-audited system.

The Framework in Synergy: A Multi-Layered Defence

The pillars work in concert to contain the blast radius:

Controlled Workflows define *what* can happen.

Dynamic Authorisation controls *if* it can happen.



Simulated Environments verify *how* it will happen.

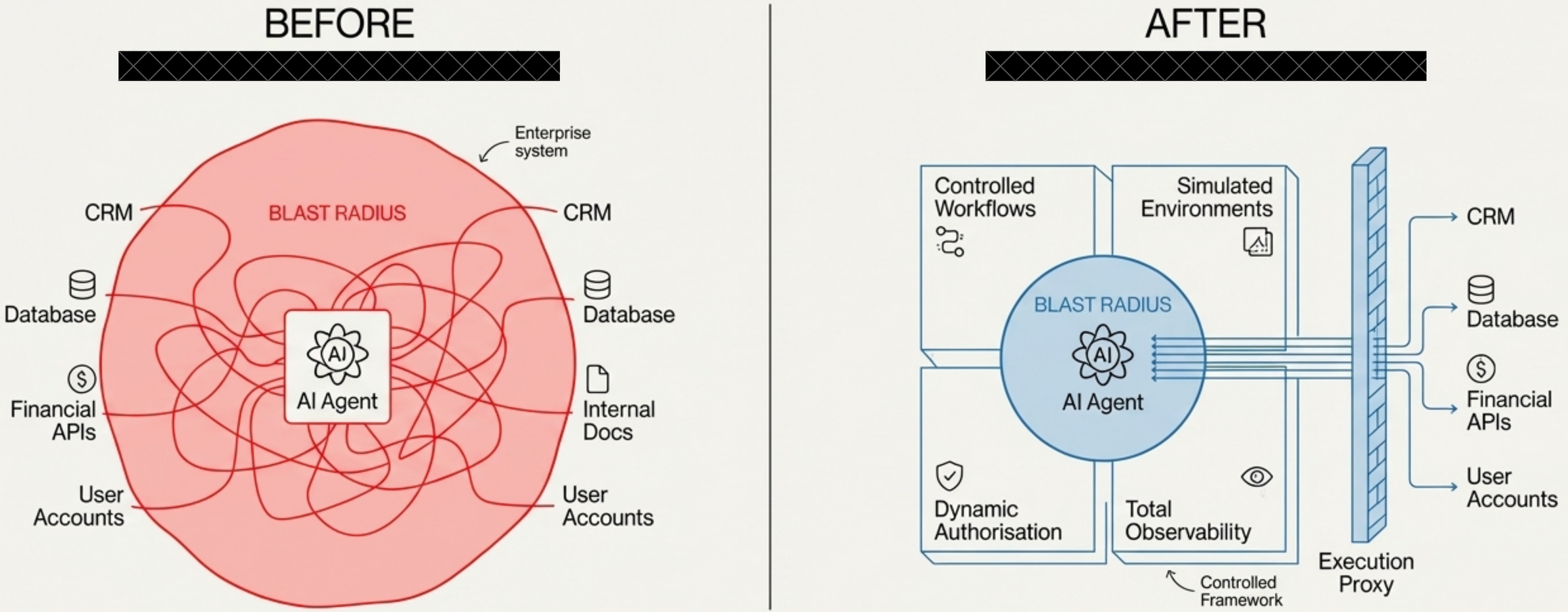
Total Observability records *that* it happened.

This integrated approach transforms the agent from a wild card into a governed entity, creative and autonomous only within the safe bounds we set.

Practical Considerations: Balancing Control and Complexity

Challenge	Mitigation
Performance Overhead Per-action token requests can add latency.	**Optimised Architecture** Caching low-risk authorisations; highly optimised identity providers.
Integration Complexity Requires integration with existing IAM and APIs.	**Standards-Based Approach** Leveraging modern standards like OAuth 2.1; treating agents as first-class identities.
Workflow Design Effort Defining graphs requires a new design discipline combining development and security expertise.	**Reusable Frameworks** Using Policy-as-Code frameworks; developing libraries of common, reusable workflows.

The Result: From Uncontrolled Risk to Governed Autonomy



This model brings the best practices of modern security—least privilege, rigorous testing, and complete auditing—into the realm of AI agents. Early evidence shows strict scoping can reduce permission sprawl by up to 90% without hindering functionality.

A New Paradigm for AI Governance

This is more than a technical fix; it is a necessary evolution in how we manage non-human entities, aligning with core Zero Trust principles.

> **“Agents aren’t people, and they can’t be governed as if they were. Assign ownership. Scope least privilege. Prefer federation over static secrets. Monitor actions. Retire fast.”**

Our model embodies these principles by design, treating each agent action as a discrete event that must be authenticated, authorised, and logged.

Conclusion: Enable Creativity by Ensuring Safety

To safely harness the power of agentic AI, we must impose structure and limits externally rather than expecting agents to be inherently safe.

1. Unconstrained agentic AI poses an unacceptable and unprecedented **risk** to the enterprise.
2. A holistic model combining Controlled Workflows, Dynamic Authorisation, Simulation, and Observability provides robust, multi-layered **control**.
3. This structured approach allows AI to be **creative** in solving problems while being fundamentally incapable of breaching **safety constraints**.

Speed the business, without widening the blast radius.