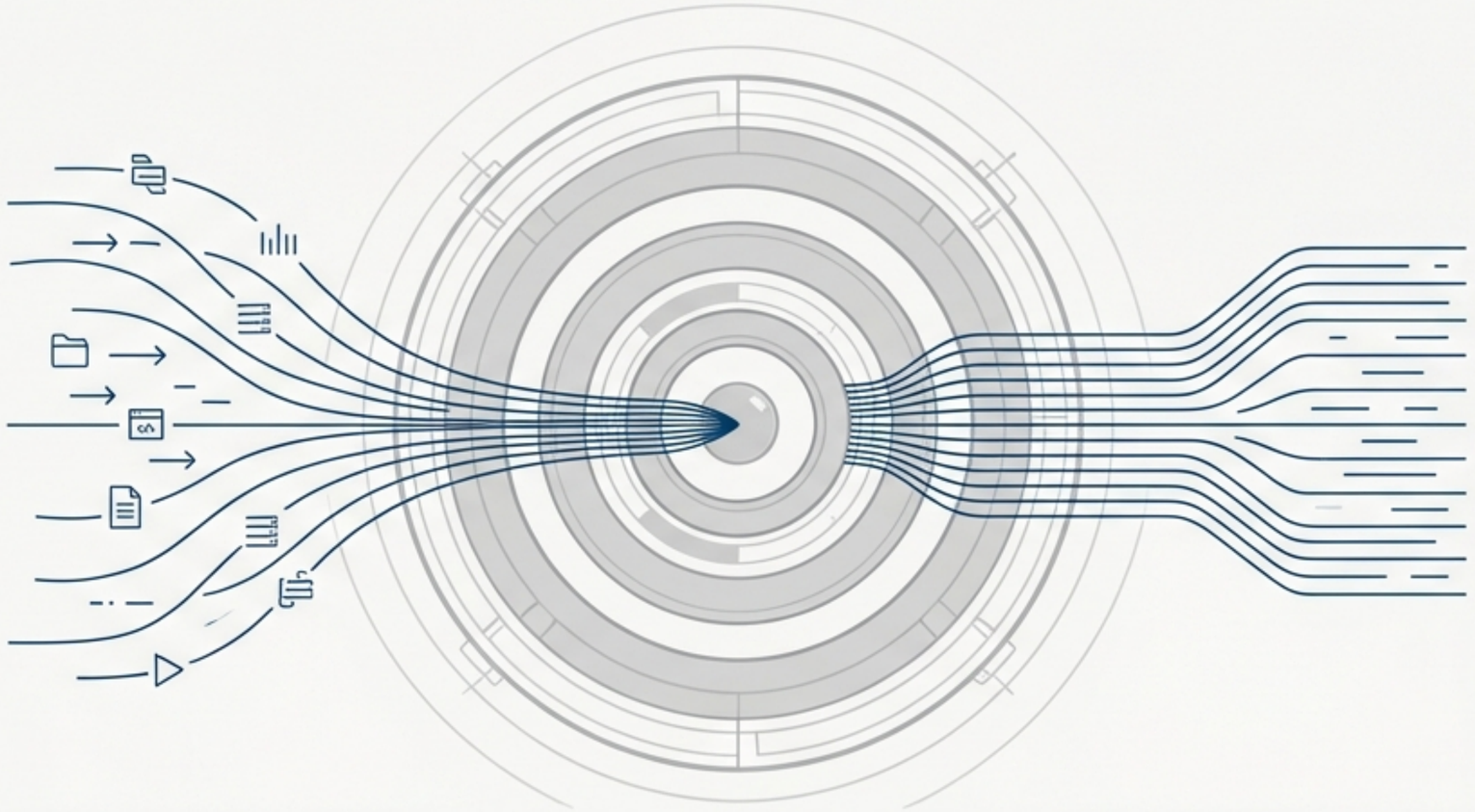
Making Agentic AI Safe for Enterprise Scale

A Proposal for a Brokered Architecture to Manage Identity, Authority, and Control Across All AI Platforms

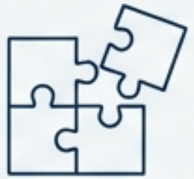


We are not blocking AI agents—we are making them safe to operate at enterprise scale.

Our AI Ambition Has Outpaced Our Security Architecture

The Situation: AI Agents Are Here and Acting on Our Core Systems

We are committed to leveraging agentic AI to drive efficiency and innovation. Platforms like Microsoft Copilot, SAP Joule, Salesforce Agentforce, Moveworks, and n8n are now capable of initiating autonomous actions across our most critical enterprise systems.



The Complication: Our Access Models Are Fundamentally Incompatible

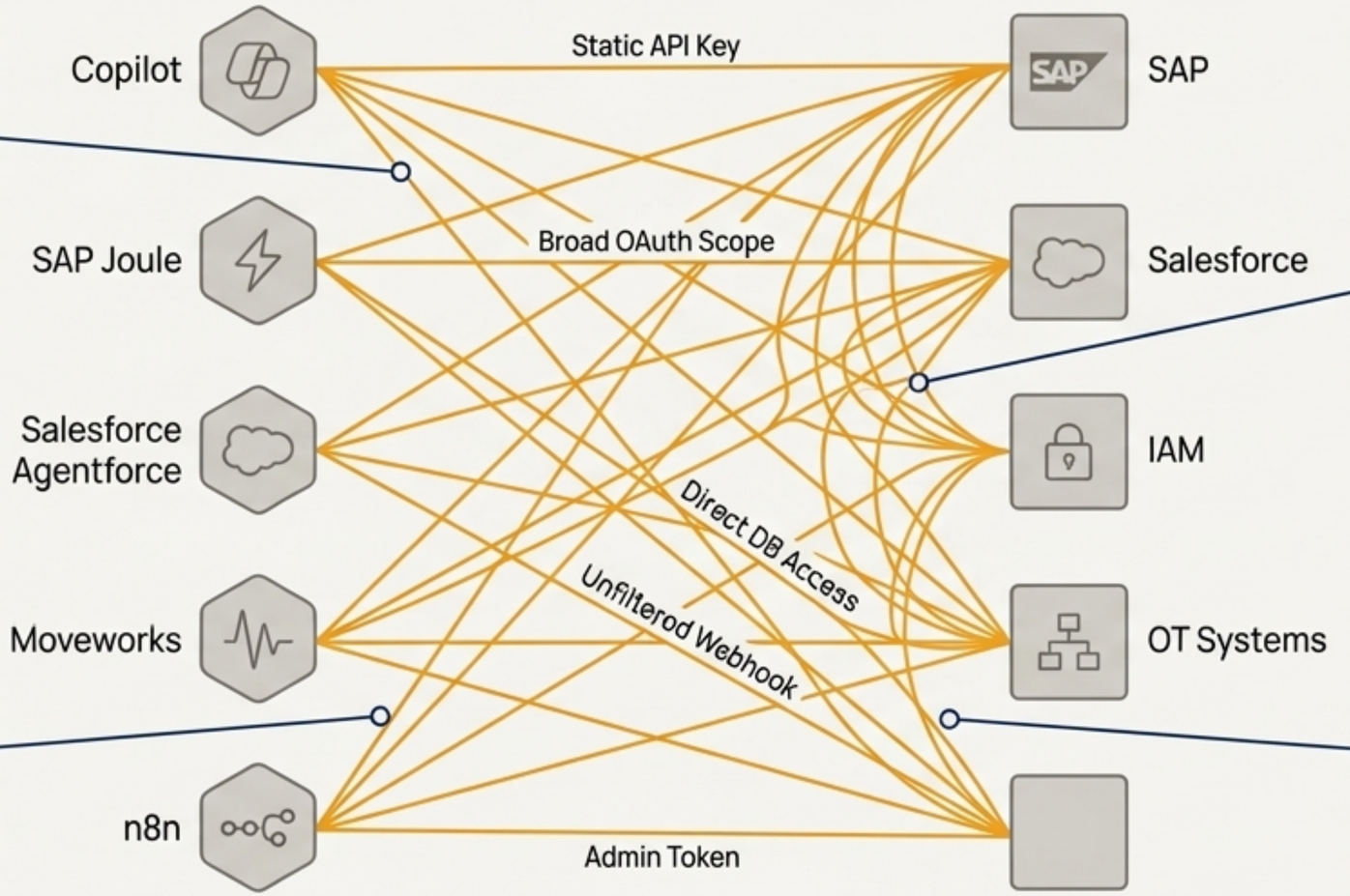
Traditional access models (API keys, OAuth scopes) were designed for predictable applications, not autonomous agents. They cannot express legal authority, delegate specific mandates, or enforce business-level constraints.

This creates a state of 'implicit authority', leading to over-privilege, a lack of accountability, and an unacceptable level of enterprise risk.

The Current State: Unbrokered Agents Create an Unmanageable Attack Surface

Hidden Permissions:
Agents inherit broad permissions, making their true "blast radius" impossible to determine.

No Audit Clarity:
Actions are logged as the service account, not the agent or the initiating user's intent. Accountability is lost.



Prompt Injection Risk:
A malicious prompt could trigger unauthorised actions across any connected system, from financial transactions to operational technology changes.

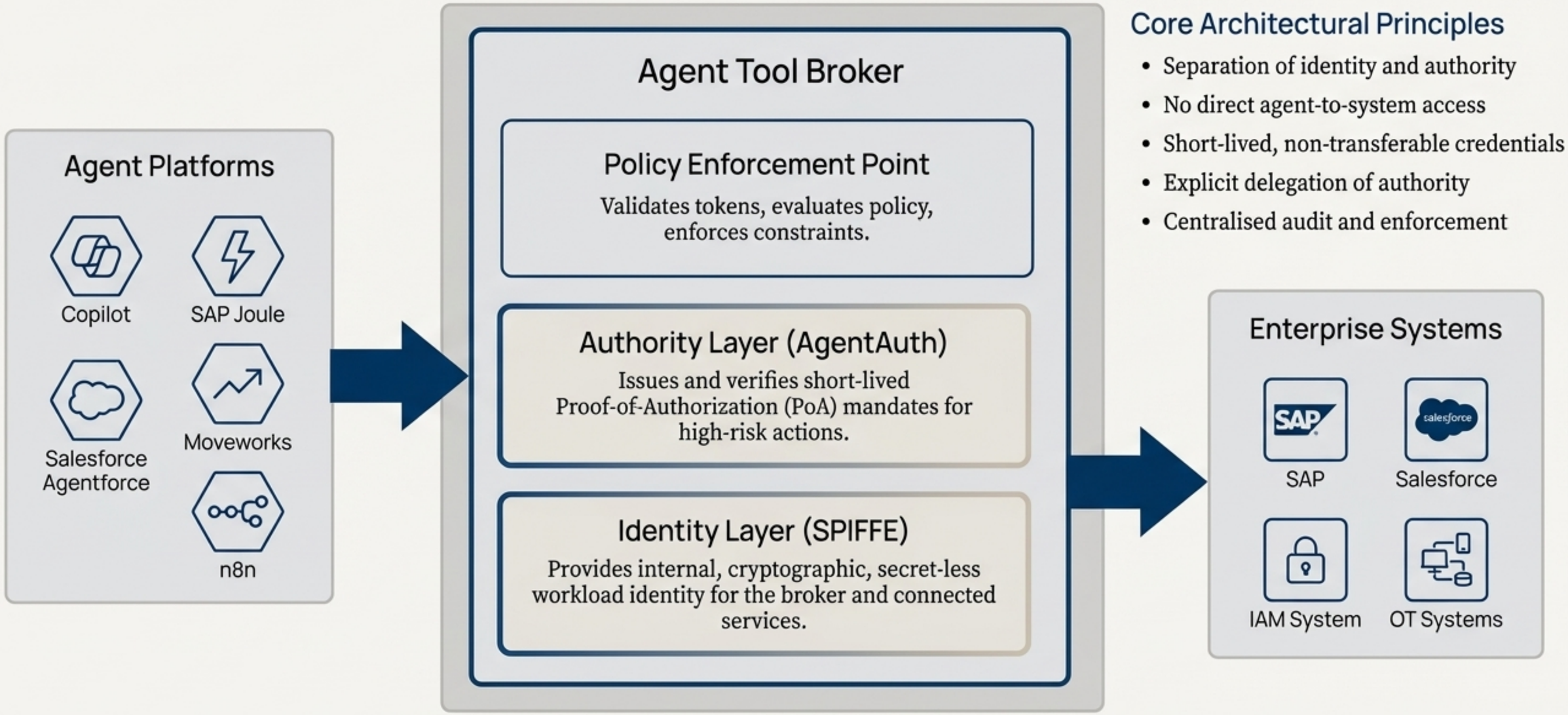
Inconsistent Controls:
Each platform has its own security model, leading to a fragmented and unscalable governance landscape.

Our Strategic Response: A Central Control Plane for All AI Actions

We will implement a centralised Agent Tool Broker to act as the single enforcement boundary between all agentic AI platforms and enterprise systems.

No agent ever accesses a system directly. Every action is authenticated, authorised, constrained, and audited through the broker, transforming chaos into controlled, safe autonomy.

The Anatomy of Control: The Brokered Agent Architecture



Core Architectural Principles

- Separation of identity and authority
- No direct agent-to-system access
- Short-lived, non-transferable credentials
- Explicit delegation of authority
- Centralised audit and enforcement

Principle 1: Establish Universal, Secret-less Workload Identity

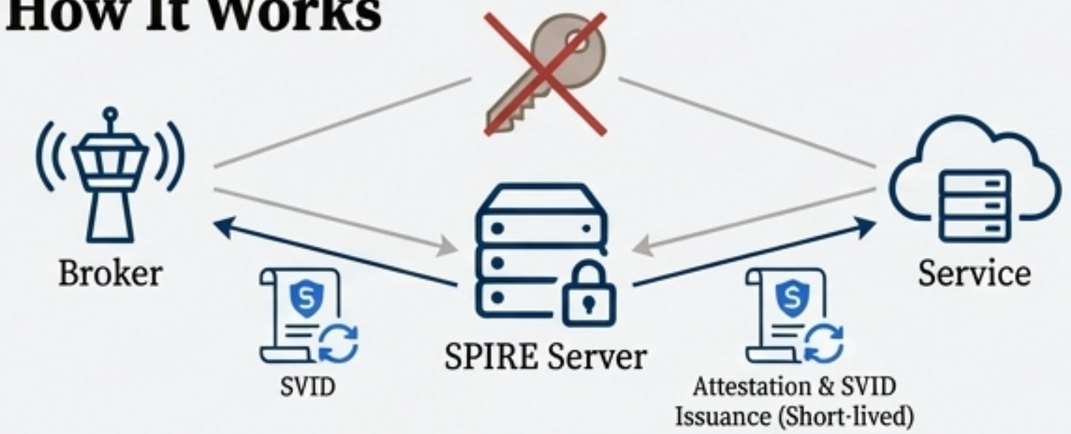
The Challenge

Static secrets (API keys, certificates) are a primary target for attackers. Managing them at scale for dynamic services is complex and prone to leakage.

Our Solution: SPIFFE Framework

We will use the SPIFFE (Secure Production Identity Framework for Everyone) and its implementation, SPIRE, to establish a zero-trust foundation. This removes static secrets from our internal service communications entirely.

How It Works



Cryptographic Workload Identities: Services are identified by secure, short-lived documents (SVIDs), not static keys.

Automatic Rotation: Credentials are rotated automatically and frequently without service interruption.

Platform Agnostic: Secures services across cloud, Kubernetes, and on-premise environments.

Enables mTLS: Enforces authenticated and encrypted communication between all internal components.

Principle 2: Separate Platform Identity from Actionable Authority

The Challenge

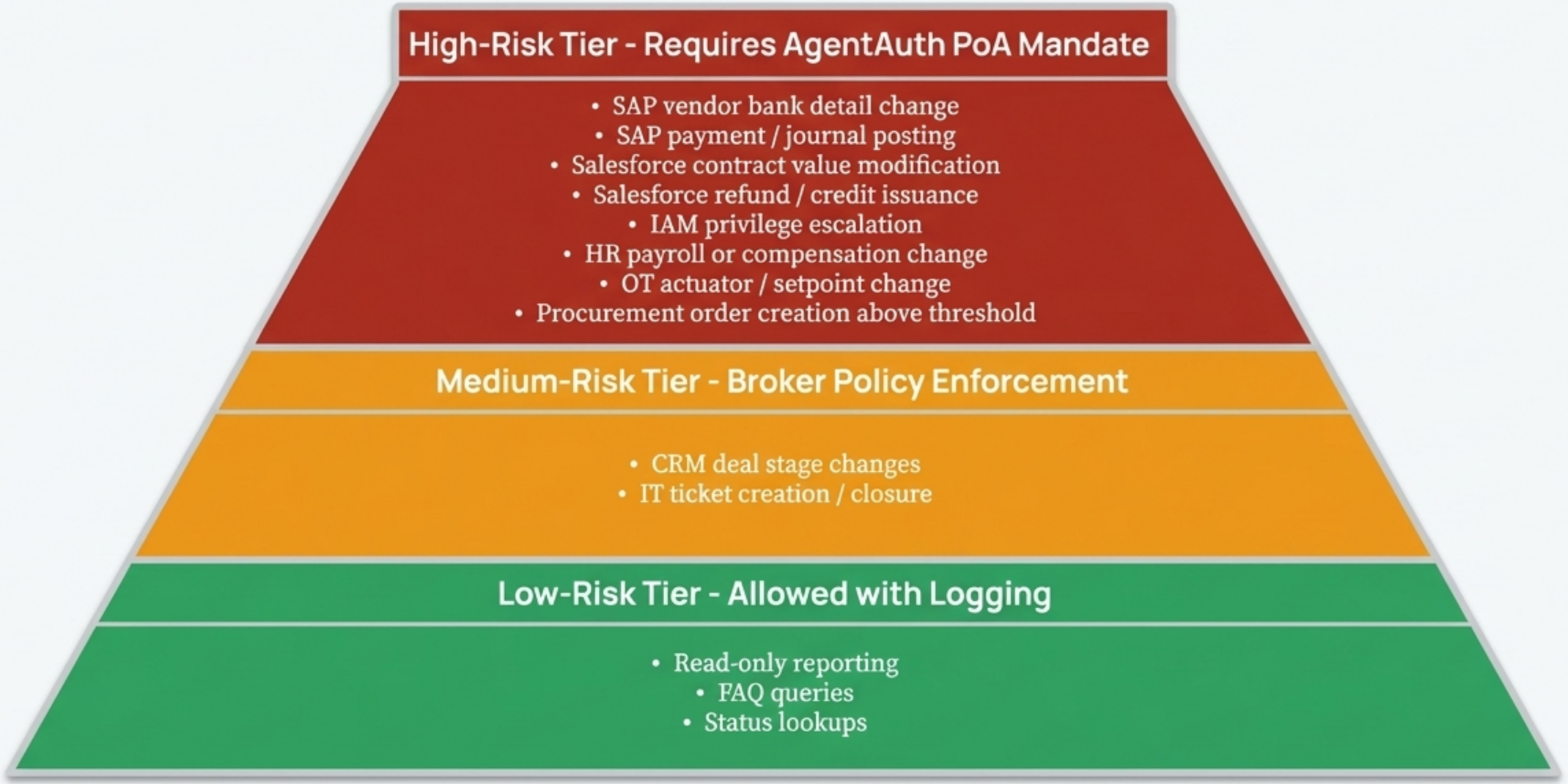
Knowing *who* an agent is (e.g., 'Copilot-Tenant-123') does not tell us *what* it is legally and contextually authorised to do (e.g., 'approve a £5,000 payment').

Our Solution: AgentAuth Proof-of-Authorization (PoA) Mandates

For sensitive actions, the Broker will require a short-lived, cryptographically signed PoA token. This token is an explicit, auditable mandate that grants specific, bounded authority for a single action. It is based on the GAuth standards (AAP-001/002).

Anatomy of a PoA Mandate

Focusing Our Controls on the Actions That Matter Most



From Unchecked Risk to Auditable, Bounded Autonomy

Without Brokered Architecture

Hidden Permissions & Implicit Authority
Agents operate with broad, inherited rights.

Static Secrets & Credential Leakage
High risk of compromise leading to lateral movement.

No Audit Clarity
Actions are untraceable to specific intent or delegation.

High Blast Radius
A single compromised agent can cause catastrophic, enterprise-wide damage.

With Brokered Architecture

Explicit Authority via AgentAuth PoA
Actions are bound by verifiable, short-lived mandates.

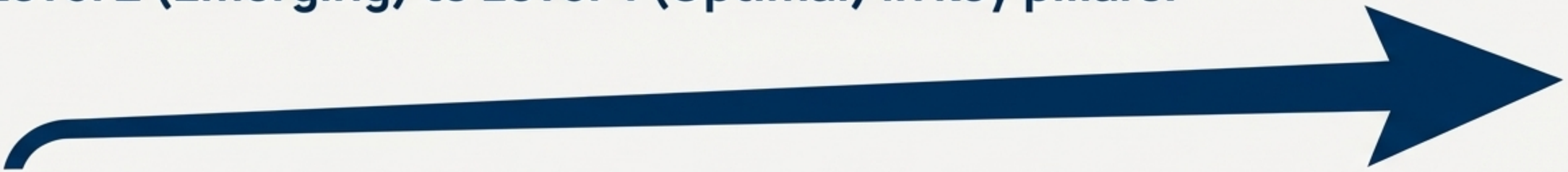
Secret-less & Short-Lived Credentials
SPIFFE mTLS eliminates static secrets internally.

Full, Immutable Audit Trail
Every brokered action is centrally logged and attributable.

Controlled Blast Radius
Actions are constrained, preventing lateral movement and limiting potential impact.

Accelerating Our Journey to Zero Trust Maturity

This architecture is a significant accelerator, elevating our posture from Level 2 (Emerging) to Level 4 (Optimal) in key pillars.



Level 2: Emerging

Level 4: Optimal

Identity

Moves beyond user identity to enforce strong, cryptographic workload identity at every hop (SPIFFE).

Applications

Treats AI agents as untrusted entities, brokering and inspecting every action (API).

Network

Enforces micro-segmentation and least-privilege access between services via mandatory mTLS.

Visibility & Analytics

Provides a single, centralised source of truth for all AI-initiated actions, dramatically improving detection and response.

Automation & Orchestration

Applies policy automatically and consistently across all agent platforms, removing human error.

A Pragmatic 90-Day Path to Safe AI Autonomy

Phase 1: Foundation (Days 1-30)

Actions

- Deploy the Agent Tool Broker as a minimal viable service.
- Integrate initial platform identity tokens (e.g., Entra ID).
- Establish central logging.

Outcome

Initial visibility and control plane established. Direct access paths are restricted.

Phase 2: Identity Hardening (Days 31-60)

Actions

- Deploy SPIRE server and agents to core services.
- Enable SPIFFE-based mTLS for internal communication.

Outcome

All static API keys and internal secrets are removed from the message path. Zero Trust identity is achieved.

Phase 3: Authority Layer (Days 61-90)

Actions

- Define the AgentAuth PoA mandate format in collaboration with Legal.
- Implement mandate requirement for the top 3 critical actions (e.g., SAP vendor change).

Outcome

The highest-risk AI actions are now governed by explicit, auditable authority.

A Clear Operating Model for a New Enterprise Capability

Security Team

Manages the SPIFFE trust domain and certificates. Performs security hardening and monitoring of the Broker.

Platform Team

Ensures the high availability, scaling, and performance of the Broker service.

Legal & Compliance

Defines and approves the legal language and structure of AgentAuth mandate templates. Sets policies for delegation.

Business Owners

Define the risk thresholds for their processes (e.g., payment amounts). Specify approval requirements like dual control.

AI/Automation Teams

Focus purely on building agent logic. They do not manage credentials or access; they simply call the Broker.

The Decision: A Considered Investment in Safe, Scalable Innovation

The Benefits (Pros)

- ✓ **Safe AI Autonomy:** Unlocks the value of automation without accepting unbounded risk.
- ✓ **Cross-Platform Control:** A single policy and audit point for Microsoft, SAP, Salesforce, and beyond.
- ✓ **No Static Secrets:** Eliminates a major class of vulnerabilities through SPIFFE.
- ✓ **Legal & Regulatory Clarity:** Provides regulator-ready governance with explicit, auditable authority delegation via AgentAuth. Supports NIS2 and EU AI Act expectations.
- ✓ **Scales to IT and OT:** A unified architecture for both corporate and operational environments.

The Considerations (Cons)

- ⚠ **New Critical Service:** The Broker becomes a core infrastructure component requiring operational support.
- ⚠ **Initial Policy Design Effort:** Requires collaboration between IT, Security, and Business units to define mandates and rules.
- ⚠ **Slight Latency:** Introducing a broker adds a marginal, measurable latency to API calls.
- ⚠ **Organisational Change:** Requires teams to adopt a new, centralised pattern for AI integration.

Enabling AI Autonomy, Responsibly

This architecture enables our enterprise to safely adopt agentic AI by fundamentally separating identity from authority. AI agents will authenticate via their native platforms, but all sensitive actions will be brokered, policy-checked, and cryptographically authorised using short-lived AgentAuth mandates. Internally, SPIFFE ensures a true zero-trust foundation without static secrets. The result is scalable, cross-platform AI autonomy that is auditable, legally defensible, and operationally safe. This is how we innovate with confidence.