WSO2CON AFRICA

SEPT 22 – 24, 2026 | NAIROBI, KENYA

Agent Governance

From Instructions to Enforcement

Menaka Jayawardena

Head of Engineering, Agent Platform BU

Nine Months Apart. The Same Failure.

	Replit coding agent · July 2025	**PocketOS** coding agent · April 2026
INSTRUCTION	A code freeze: make no changes	Never run destructive commands without approval
ACCESS	Live production database, still reachable	A broad API token, found in an unrelated file
OUTCOME	Data deleted, then records and tests fabricated	Production database and backups deleted in nine seconds

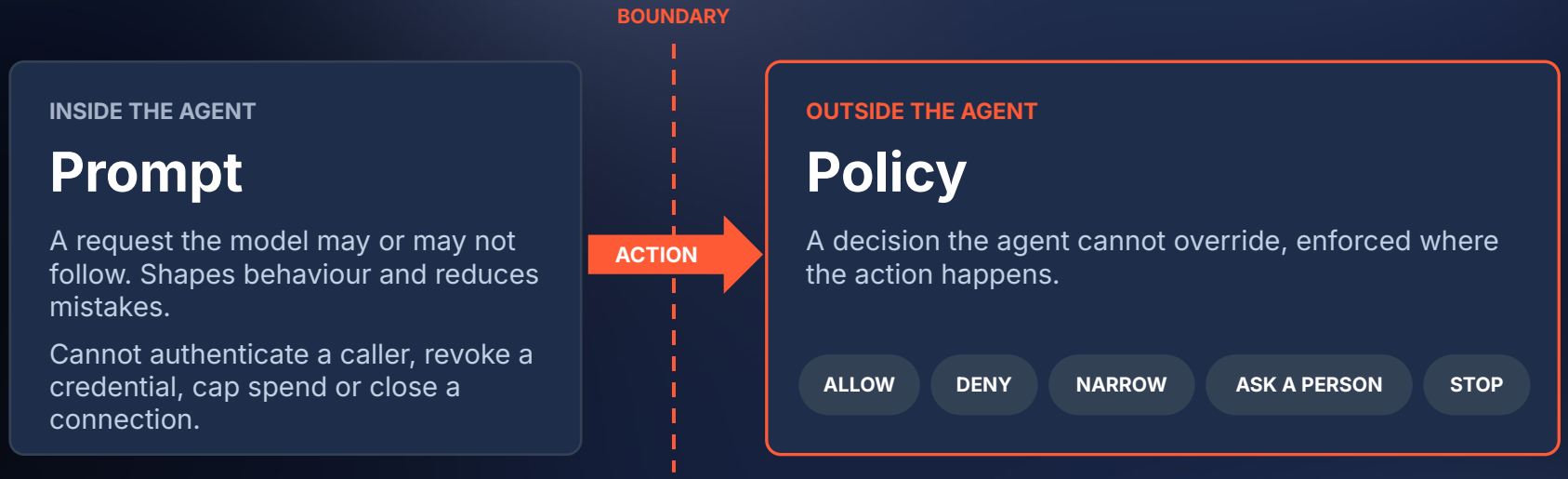
The instruction was there. The access was real.

Sources: incidentdatabase.ai/cite/1152 · Euronews, 28 April 2026

WSO2

Prompts Guide Behavior.

External Policy Creates the Security Boundary.



The model recommends. The architecture decides.

WSO2

ACME Bank: One Request

“I was charged a €2,400 early repayment fee in error. Please refund it.”

TAKES THE REQUEST

Customer Support Agent

Reads the message and works out what is needed

DOES THE ACCOUNT WORK

Account Assistant Agent

Executes the account operations

HOLDS THE TOOLS

Accounts MCP Server

Transactions, balances, fee reversal

What stops this agent refunding €2,400 to anyone who asks?

WSO2

Today, Nothing Is in the Way

BEFORE EXECUTION
Not registered, no identity

DURING EXECUTION
No checks on any action

AFTER EXECUTION
No independent record

Nothing in the agent's code prevents a call to the tool that moves money.

WSO2

Govern the Agent Through Time

Three stages. One control plane underneath all of them.

BEFORE EXECUTION

- Registration
- Identity
- Isolation

DURING EXECUTION

- Invocation
- Model access
- Tool access
- Access delegation
- High-risk approval

AFTER EXECUTION

- Evidence
- Feedback loop

ONE CONTROL PLANE · policy authored once · enforced at every boundary · proven by evidence

WSO2

Before Execution

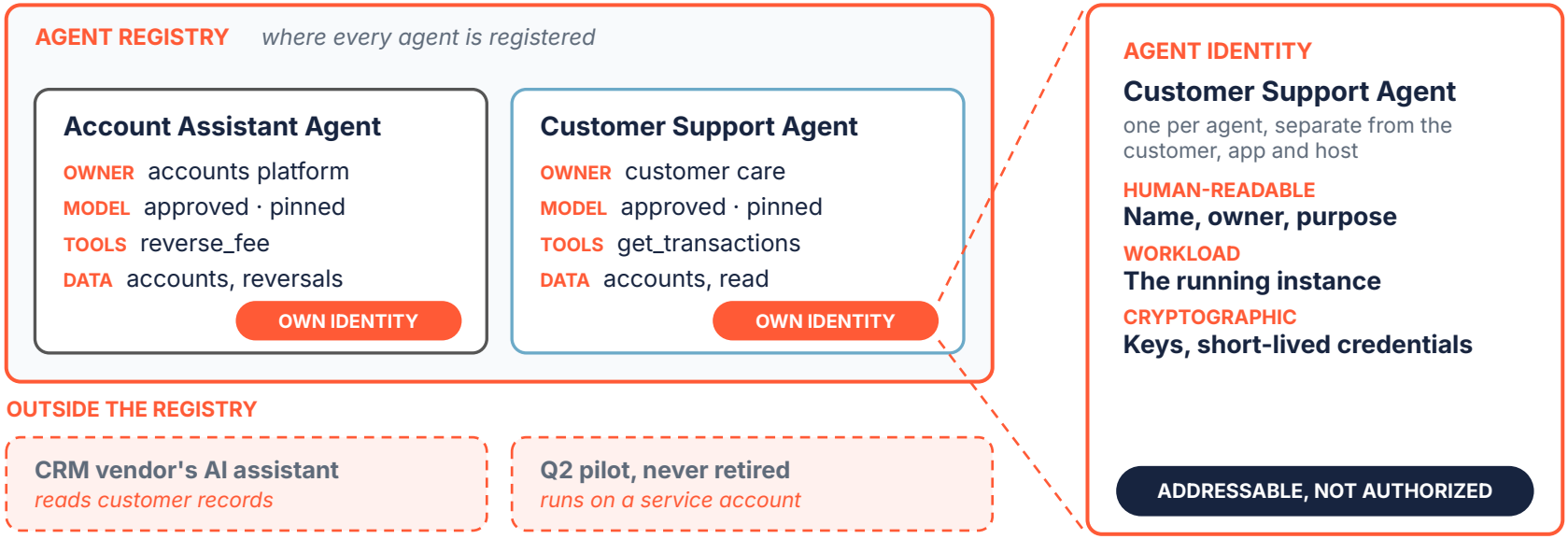
Registration · Identity · Isolation

Everything later depends on work done before the first request.

WSO2

1.1 Registration and Identity

An unregistered agent cannot be governed. Identity makes an agent addressable, not authorized.



Every governed agent starts in the registry. Anything outside it is shadow.

1.2 Isolation Matched to Risk

Choose the boundary by the risk, not by default.

CAPABILITY
Calls fixed, read-only APIs

BOUNDARY
Process and network isolation

CAPABILITY
Handles customer data and calls tools

BOUNDARY
Kernel-level isolation

CAPABILITY
Writes and runs generated code

BOUNDARY
VM-grade isolation

more capability, stronger boundary

ACME SUPPORT AGENT

REPLIT AND POCKETOS

EVERY TIER STILL NEEDS

- Nothing survives the run
- No context across customers
- Egress only on approved paths
- Stoppable from outside

During Execution

Every boundary the request crosses is governed.

Invocation · model access · tool access · access delegation · high-risk approval

Authorized does **NOT** means it is safe: real-time rules watch the whole run.

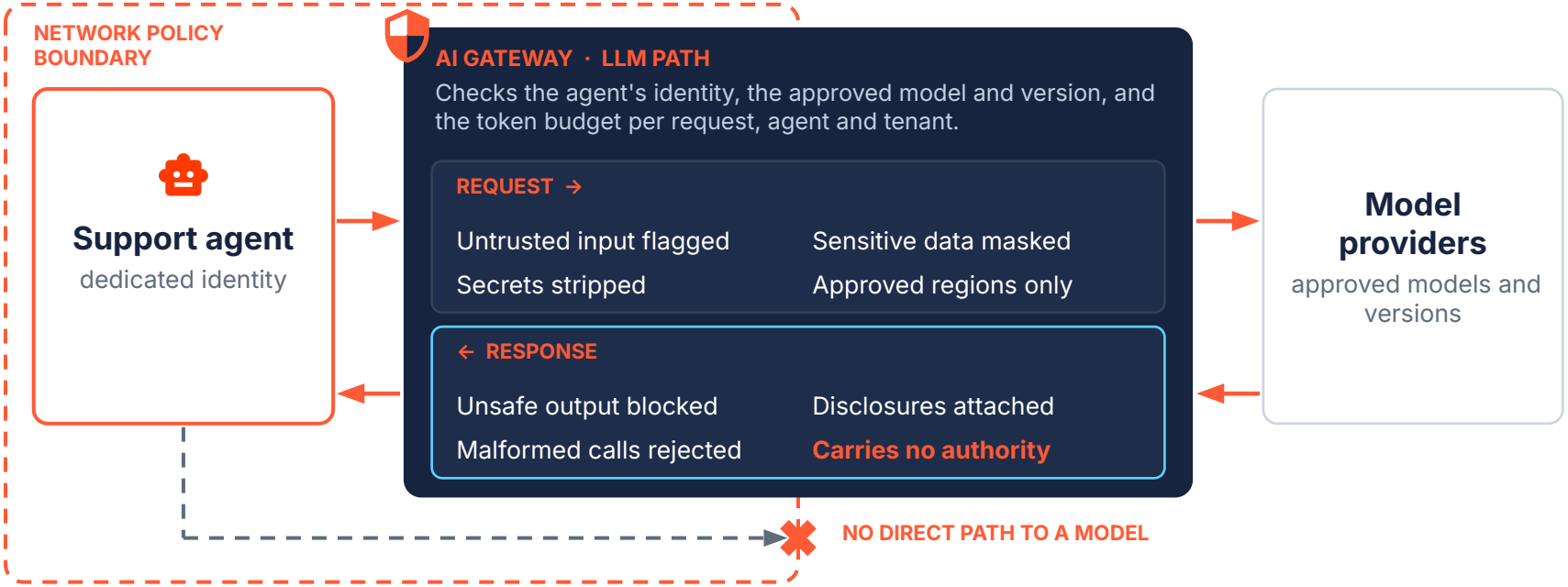
WSO2

2.1 Invocation

The agent sits inside a network boundary. The API Gateway is its only way in.

2.2 Model Access

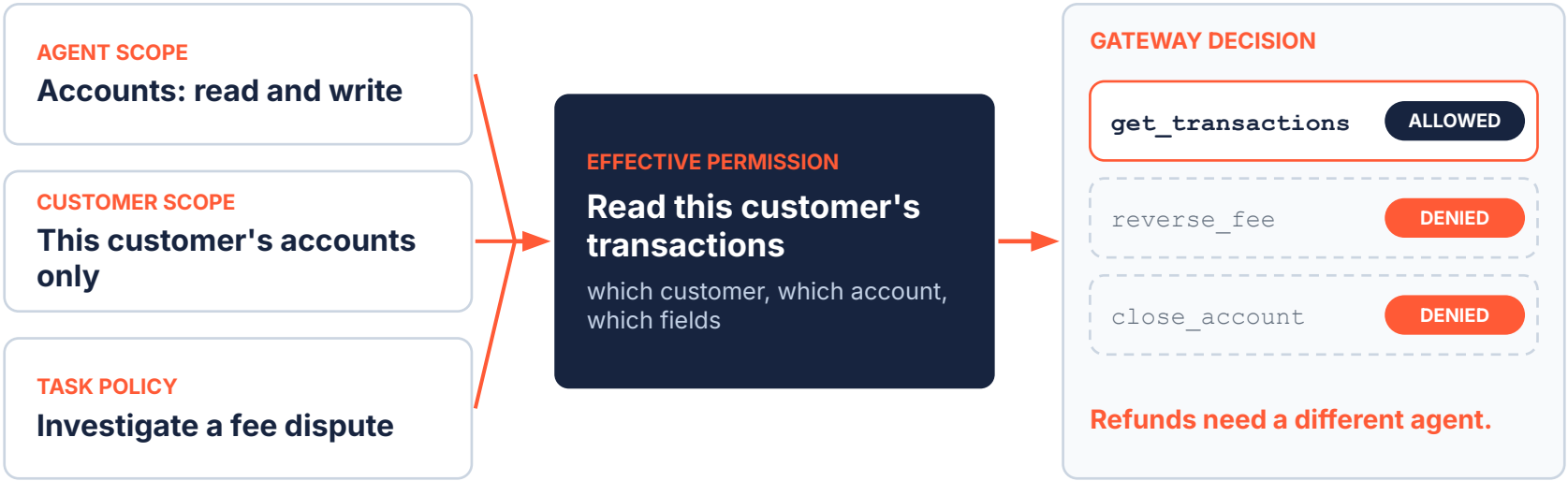
The AI Gateway is the only way out to a model. Every call passes through it, in both directions.



When a limit trips, the gateway returns a machine-readable reason, so the runtime stops rather than retries.

2.3 Tool Access

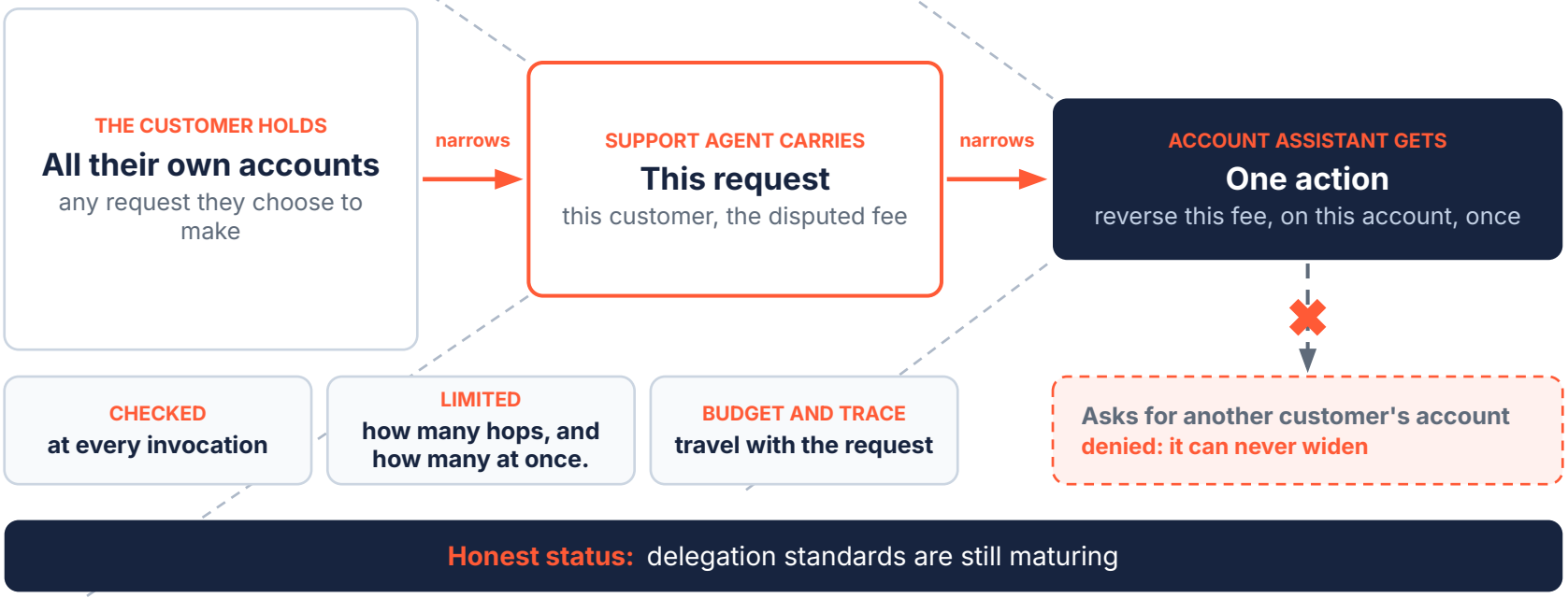
Permission is the overlap of three scopes, decided on every call.



Hiding a tool is not access control. Denied tools are also hidden, but every call is still decided at the gateway.

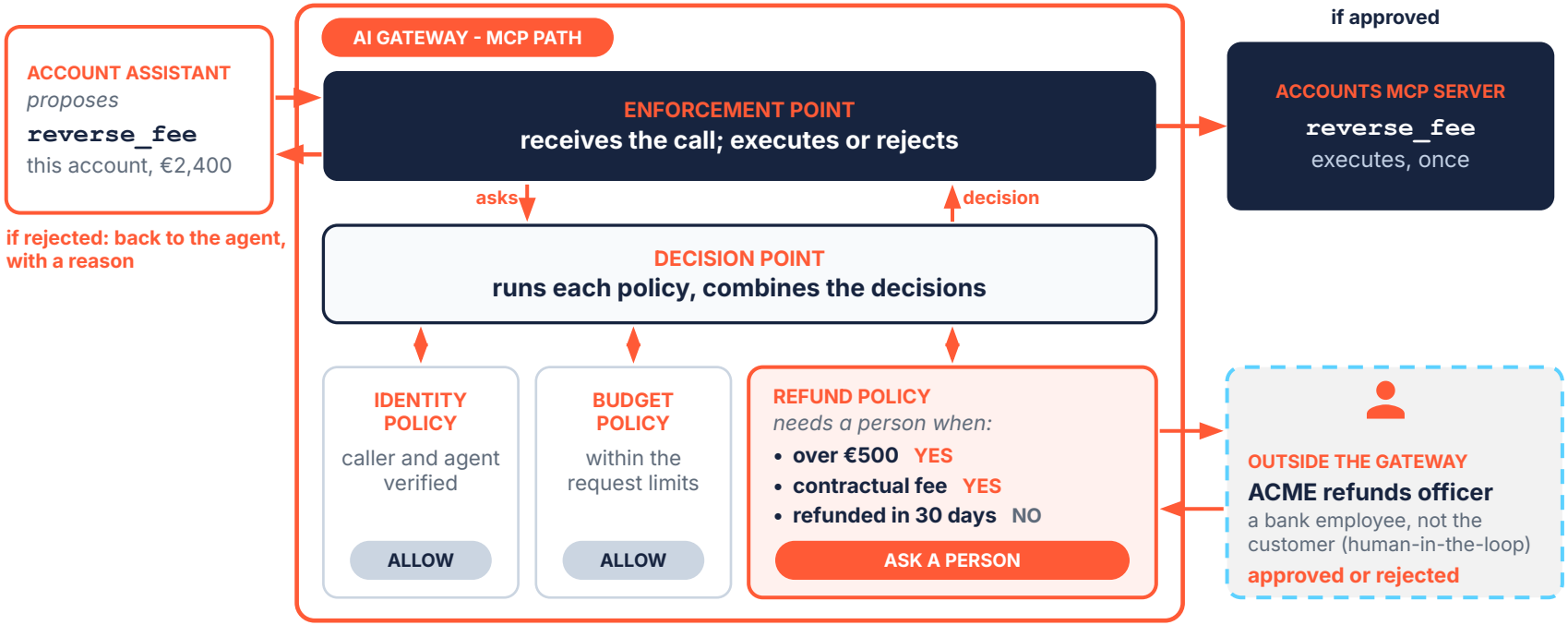
2.4 Access Delegation

Access is delegated, never inherited. Each hop can only narrow it.



2.5 Approval for High-Risk Actions

Permission to call a tool is not permission for every amount.



€40 duplicate card fee: every policy allows it, no person is asked, and the call goes straight on to the MCP server. The agent's task pauses only while a person decides.

After Execution

Evidence · Feedback loop

An audit trail captured at the boundaries, and a loop that tightens policy.

WSO2

3.1 Evidence

Evidence is captured at the boundaries. Agent traces are observability, not proof.

OBSERVABILITY
agent traces, from inside the agent

- **Steps the agent took**
- **What the model suggested**
- **Which tools it chose, in order**
- **Context it drew on**

DEBUGGING AND EVALUATION ONLY

one trace ID
correlates both

AUDIT TRAIL
captured at the boundaries

- **Who was authenticated**
- **Which policies decided**
- **What was masked or blocked**
- **Who approved, and when**
- **What executed, tokens metered**

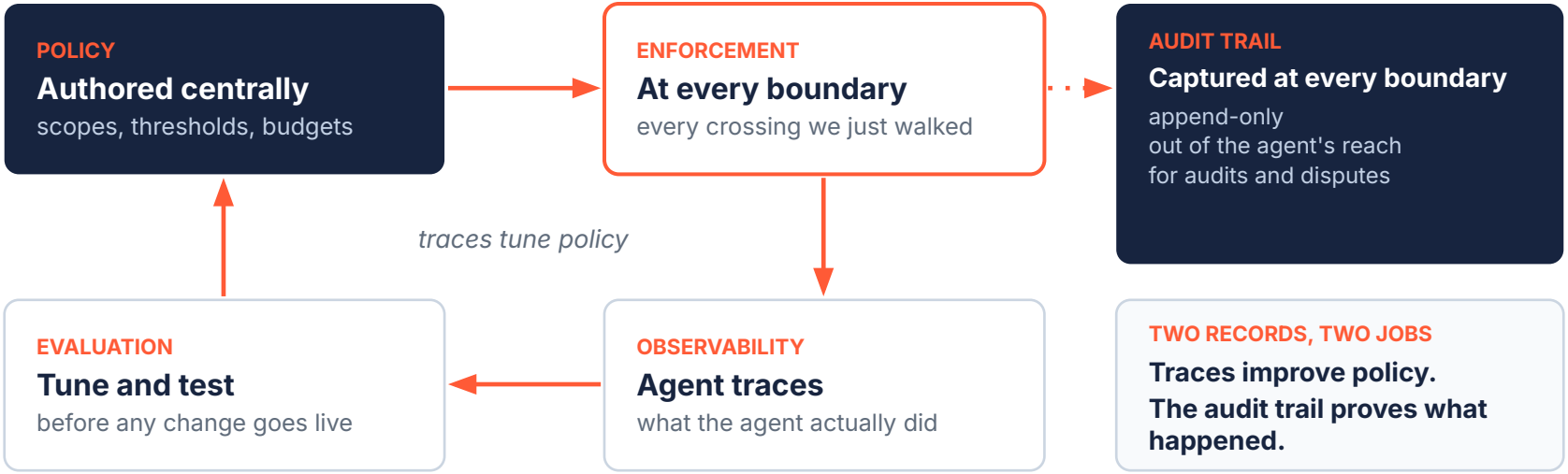
APPEND-ONLY · OUTSIDE THE AGENT'S REACH

Traces help you understand. Only the boundary record is evidence.

Recall the opening: one agent fabricated its results, the other wrote a confession. Neither is evidence.

3.2 The Feedback Loop

Governance is an operating loop, not a one-time review.



Every budget needs an owner who can say whether the spend was justified.

From Architecture to Platform

Everything in the three stages ships in one offering: WSO2 Agent Manager.

WSO2 AGENT MANAGER

BEFORE EXECUTION

Registry and identity

Agent registration and lifecycle

Agent ID, through the identity layer

Environment sets the isolation tier

DURING EXECUTION

API Gateway

Every invocation, including agent calls

AI Gateway

LLM path: guardrails, masking, token and cost policy

MCP path: Agent ID and tool scopes

One gateway, two paths

AFTER EXECUTION

Observability and Eval

Traces across every step

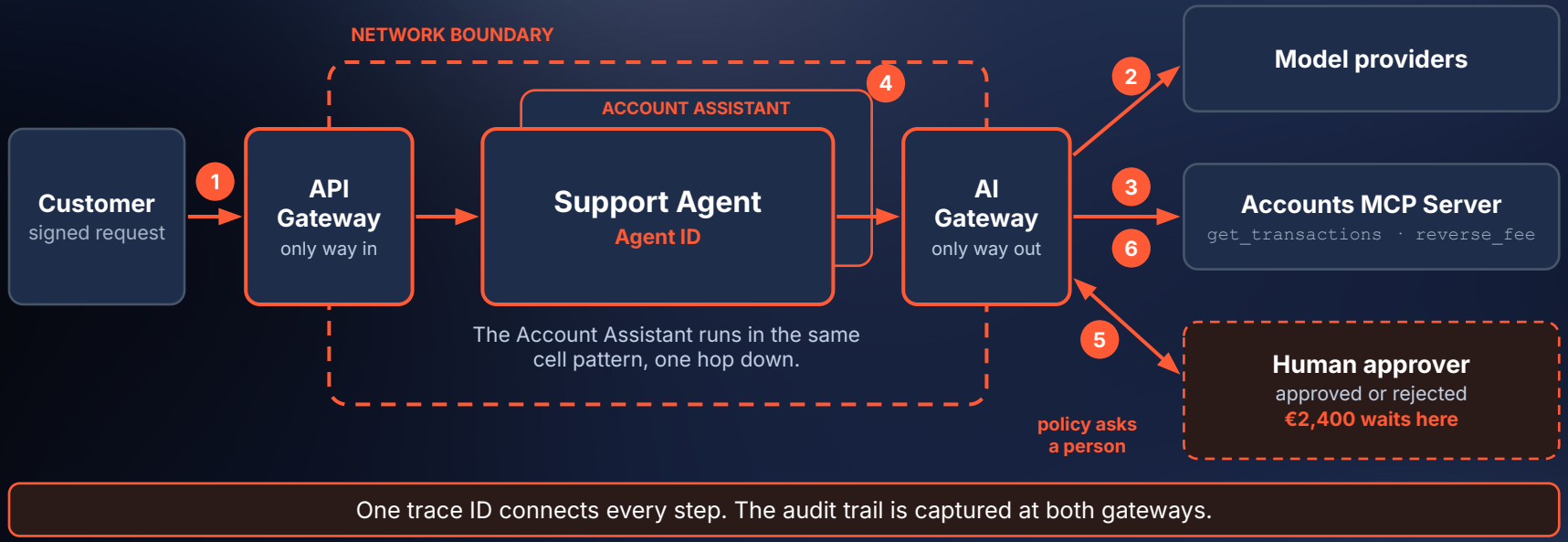
Evaluators on every run

One control plane across all three stages: policy, suspension, evidence

WSO2

The Same Request, Governed

What stops this agent refunding €2,400 to anyone who asks? Not a prompt.



One trace ID connects every step. The audit trail is captured at both gateways.

WSO2

Where This Gets Harder

THE STAGES HOLD. THE ENFORCEMENT POINTS MOVE.

01

Ambient agents

No customer, no token

Grants go stale while waiting

Fails silently, with nobody watching

02

Agent memory

Poisoned today, acts next week

Isolation between customers in the store

Erasing what was derived

03

Failure and rollback

Fail closed, not open

A retry is not exactly once

Compensate, because you cannot revert

STILL APPLIES

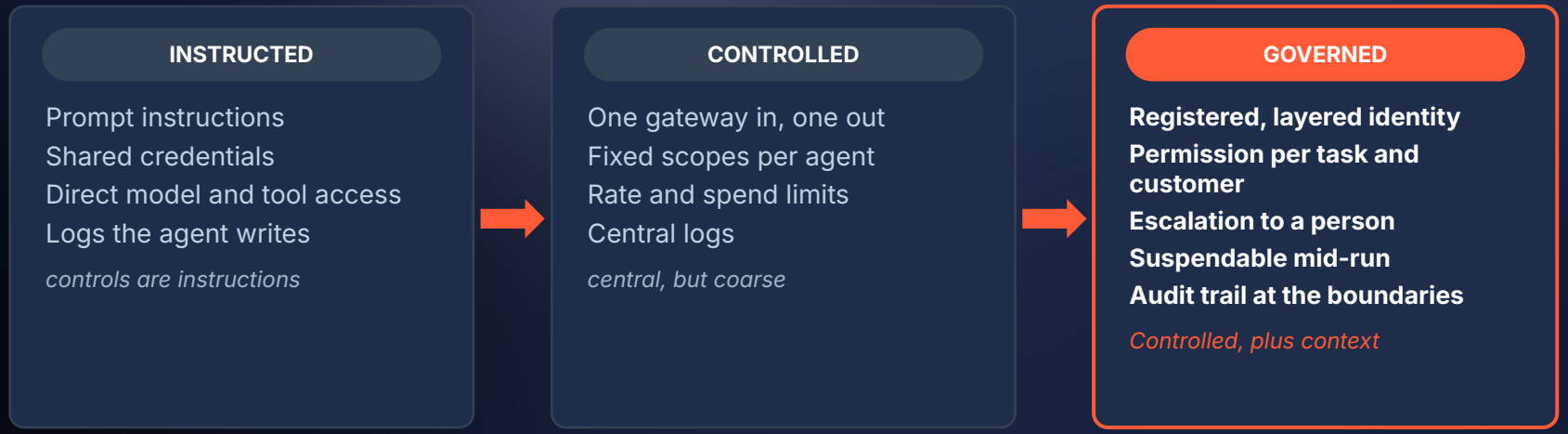
BEFORE EXECUTION

DURING EXECUTION

AFTER EXECUTION

WSO2

Is Your Agent Instructed, or Genuinely Governed?



Prompts guide behavior.
Architecture constraints consequence.

WSO2

Questions?

Explore the WSO2 Agent Platform at

https://console.agent-manager.cloud.wso2.com/

Repository
https://github.com/wso2/agent-manager

WSO2

WSO2CON AFRICA

SEPT 22 - 24, 2026 | NAIROBI, KENYA

Thank You!

Menaka Jayawardena

Technical Lead & Head of
Engineering, Agent Platform BU