

ANTHROPIC MANAGED AGENTS SPLIT EXECUTION AND ORCHESTRATION OPERATIONAL SIGNAL

DOMAIN: RISK — Risk Surface | CONFIDENCE LEVEL: Preliminary

INSTITUTIONAL DIMENSIONS: Authority | Attribution | Visibility | Escalation | Continuity

Signal ID	Platform	Source	Audience	Status	Date
SIG-RISK-002	Anthropic	Link	PMO Risk Ops IT Architecture	Pilot	2026-05-19

CAPABILITY SHIFT

Anthropic's Managed Agents architecture separates the agent harness, session log and sandbox execution environment. Self-hosted sandboxes, described in May 2026, allow tool execution to run in customer-controlled infrastructure while orchestration, context management and session logging remain separate from that execution environment.

ORGANISATIONAL CONSEQUENCE

Organisations may assume customer-hosted sandboxes keep agent activity fully inside their environment. That assumption may not hold. Tool execution can occur inside the customer boundary while coordination, delegation and event logging remain outside it.

WHAT TO WATCH

- Agent workflows treated as fully internal
- Sensitive activity appearing in session logs
- Multiagent chains creating unclear data movement
- Teams assuming sandbox location equals workflow location

WHY THIS MATTERS NOW

Agent infrastructure is becoming more distributed. The operational question is no longer only where a tool runs. It is where coordination, context and evidence sit across the workflow. Organisations evaluating Managed Agents for sensitive work may need to test those boundaries before deployment.

WHAT TO REVIEW

- Data residency assumptions
- Session log contents and storage location
- Tool inputs or outputs handled by vendor orchestration
- Credential and proxy configurations
- Subagent delegation in architecture and risk review

GOVERNANCE CONTEXT

This is a data boundary and visibility issue. The location of execution may not describe the full agent workflow. Organisations need to know which parts sit inside their infrastructure, which remain with the vendor, and how that split can be evidenced.

Scope Boundary: This artefact translates observed platform capability changes into operational considerations. It does not constitute policy, legal advice, or enforcement guidance.