BILYTICA AI OPERATING SYSTEM

# Agent Operations Framework

Autonomy tiers, tool boundaries and human checkpoints — so agents can act without becoming a liability.

AGENT · Act

**WHAT IT IS FOR**

An agent that can act is a system with permissions, not a chat feature. AGENT defines how much autonomy a workload gets, what it may touch, who reviews it and how it is stopped.

**WHO RUNS IT**

Owned jointly by the process owner and security; operated by the platform team.

**TYPICAL TIMELINE**

Tiering and boundaries defined in two weeks; then applied per agent at delivery time.

## STAGES AND EXIT GATES

**1. Tier** — Set autonomy from 0 (suggest) to 4 (act and settle) based on reversibility and impact.

**Exit gate:** Irreversible actions never sit above the reviewed tier.

**Artefacts:** Autonomy tier record · Reversibility analysis

**2. Bound** — Give each tool a scope, a rate limit, a spend limit and an identity of its own.

**Exit gate:** The agent's identity has least privilege and is separate from a human account.

**Artefacts:** Tool permission map · Budget and rate limits

**3. Checkpoint** — Insert human approval where value, risk or irreversibility crosses a threshold.

**Exit gate:** Every checkpoint has a target response time and a default if it times out.

**Artefacts:** Checkpoint policy · Approval queue design

**4. Evaluate** — Score task completion, tool-call correctness, cost per task and harmful-action attempts.

**Exit gate:** Agent evaluations run per release, including adversarial cases.

**Artefacts:** Agent eval suite · Adversarial test log

**5. Observe** — Trace every run: plan, tool calls, data touched, cost, outcome.

**Exit gate:** Any action can be replayed from the trace during an investigation.

**Artefacts:** Run traces · Cost per task view

**6. Contain** — Keep a kill switch, a rollback path and a defined incident response for agent behaviour.

**Exit gate:** The kill switch is tested, not assumed.

**Artefacts:** Kill-switch test · Incident playbook

## STANDARDS CROSSWALK

| STANDARD | WHAT IT SATISFIES |
|---|---|
| **OWASP LLM Top 10** | Injection, excessive agency and insecure tool use covered in Bound and Evaluate. |

| STANDARD | WHAT IT SATISFIES |
|---|---|
| **MITRE ATLAS** | Adversarial technique catalogue used for the agent red-team set. |
| **NIST AI RMF** | Manage function — incident response, containment and rollback. |
| **NCA ECC** | Identity, least privilege and logging requirements applied to non-human actors. |

## WHAT YOU GET

- **Autonomy tier policy** — Which action types may run at which tier, and who approves changes.
- **Tool permission map** — Scopes, identities, rate and spend limits per tool.
- **Agent evaluation suite** — Task, tool-correctness and adversarial tests wired into release.

## WHAT WE REFUSE TO DO

- **One shared service account** — Without a distinct identity per agent, no action can be attributed after the fact.
- **Unbounded tools** — A tool with broad scope turns a prompt-injection into a data incident.
- **No cost ceiling** — Looping agents are a budget risk before they are a safety risk.

## FREQUENTLY ASKED

**What are the autonomy tiers?**

0 suggest, 1 draft for approval, 2 act on reversible steps, 3 act with post-review, 4 act and settle within limits.

**Does this replace our security review?**

No. It gives the security review a defined object to assess: identities, scopes, limits, traces and a tested kill switch.

**What makes an AI agent safe to put in production?**

A tier that sets its blast radius, bounded tools, its own identity and credentials, human checkpoints on irreversible actions, and containment that can stop it.

**Why does each agent need its own identity?**

A shared service account makes actions untraceable and over-permissioned. Per-agent identity gives least privilege and an audit trail per action.

**Book the Data Foundation Assessment**

Full framework detail: https://bilytica.com/frameworks/agent-operations-framework/

bilytica.com
info@bilytica.com