

Monitoring, Rollout, and Postmortem Playbook

Operational playbook for taking an agent from canary to full production, watching it in flight, and learning from every incident. Closes the reliability loop opened by the canvas and the failure-mode register.

1. Metrics that matter

Every metric is per-agent and per-version. Roll-ups across agents are useful for the library but never for go/no-go on a specific agent.

- Success rate by failure_id class: For each register entry, fraction of runs in the window where that mode did not fire. Report separately for critical, high, medium. Window: rolling 24h, 7d, 30d. A drop of >2 sigma vs. trailing 14d baseline opens a drift incident.
- MTTR (Mean Time To Recovery): From first Escalated or FailedTerminal event of a given failure_id to the next clean Succeeded run of the same agent after a fix. Reported per severity. Targets: critical <= 60 min, high <= 4h, medium <= 24h.
- Override rate: Fraction of HITL gates where decision_origin=human-override and action_type in {permit-with-modification, refuse-and-instruct}. Target ceiling default 8% rolling 7d.
- Escalation rate: Fraction of runs that reach Escalated for any reason. Track by tier. Tier 3 escalations are a board-level signal; target near zero outside of declared incidents.
- Fallback fire rate: Fraction of steps that traversed the Fallback state. A rising fallback rate alongside a steady success rate means a connector is degrading silently.
- Boundary-cross attempts: Count of unsafe-action-attempted and pii-leak-risk events. Target zero. Any non-zero week is a postmortem trigger.
- Adversarial suite delta: Pass rate on s1-adversarial-test-plan.md run nightly. Below 100% blocks the next deploy.
- Cost and latency: p50 / p95 wall clock per run, per-step token cost, fallback-vs-model cost share. Used for compute-economics tuning, not for reliability go/no-go.

All metrics are computed from the event stream defined by the override event schema (contract C2). No bespoke logging paths.

2. Escalation playbook

When an alarm fires, the responder follows this script. Do not improvise during an incident.

1. Acknowledge within SLA. Tier 1 within 4h, Tier 2 within 60 min, Tier 3 within 15 min.
2. Classify the failure mode. Use the register failure_id. If none fits, mark UNKNOWN and add a candidate entry to the register backlog at end of incident.
3. Decide blast radius. Quarantine the agent (block new runs, drain in-flight) when severity is critical, when pii-leak-risk fired, or when override rate breached ceiling in the last hour.

4. Restore service. Roll back to the last green version unless a one-line config change (canvas threshold, retry cap) cleanly resolves. Roll-forward only if S2 evals pass on the patched version.
5. Notify stakeholders. Operator pool (Tier 1), agent owner (Tier 2), accountable executive and any regulator-facing role (Tier 3). Use the canvas-declared channels.
6. Hand off to postmortem. Within 24h for high, within 72h for critical (postmortem itself, not the fix).

3. Rollout checklist

A new or substantially changed agent moves through four stages. Each stage has gates that must be green to advance.

3.1 Pre-deploy

- Canvas v1.0 signed by the agent owner.
- All seven S1 artifacts present and cross-referenced.
- Adversarial suite passes 100%.
- S2 eval baseline recorded; eval delta on the change is non-regressive.
- Fallback paths exercised in staging at least once each.
- HITL reviewer pool and channels confirmed reachable; SLA targets agreed.

3.2 Canary (24-72h)

- 1-5% of eligible traffic, or a single named tenant.
- 100% HITL sampling. No autonomous-execute on medium band.
- Boundary-cross attempts: zero.
- Override rate within ceiling.
- No new failure_id types not in the register.

Exit criterion: 48 consecutive hours green on all four canary metrics.

3.3 Percentage rollout

- Step up at 10%, 25%, 50%, 100% with at least 48h dwell at each step.
- HITL sampling drops to policy default (10% on medium reversible) at 25%.
- Drift detector armed at 25% onward.
- One named owner on-call at each step.

Roll back, do not push through, on any of: override rate >1.5x ceiling, MTTR breach, adversarial-suite regression, Tier 3 escalation.

3.4 Full

- Weekly reliability review (see section 5).
- Monthly canvas re-read with the owner.
- Quarterly adversarial-suite expansion.

4. Postmortem template (Five Whys + corrective hardening)

Postmortems are blameless, time-boxed, and end in a backlog of hardening tasks linked back to the override events that surfaced the issue.

```
incident_id: <uuid>
agent_id: <kebab-case>
agent_version: <semver>
detected_at: <ISO-8601>
resolved_at: <ISO-8601>
severity: <critical | high | medium | low>
failure_id_refs: [<failure_ids from register>]
blast_radius:
  runs_affected: <count>
  external_actions_taken: <count and description>
  reversibility: <reversible | partially-reversible | irreversible>
timeline:
  - t: <ISO-8601>
    event: <what happened>
five_whys:
  surface: <what users or operators observed>
  why_1: <one level deeper>
  why_2: <one level deeper>
  why_3: <one level deeper>
  why_4: <one level deeper>
  root_cause: <statement; should be addressable by a hardening task>
contributing_factors:
  - <organizational, tooling, or data factor>
what_went_well:
  - <protections that fired correctly>
what_did_not:
  - <protections that should have fired and did not>
corrective_hardening_tasks:
  - id: HT-<short>
    type: <deterministic-fallback | threshold-tightening | prompt-or-tool-change | canvas-
edit | new-failure-mode>
    description: <one sentence>
    owner: <role>
    due: <date>
    linked_override_events: [<event_ids>]
    linked_failure_ids: [<failure_ids>]
    s2_regression_seed: <yes | no>
    s3_provenance_followups: <yes | no>
followups:
  - canvas_version_bump: <yes | no, with reason>
  - register_changes: <added | severity changed | none>
  - threshold_changes: <yes | no>
  - rollout_status_change: <pause | rollback | continue>
sign_off:
  agent_owner: <role>
  reliability_owner: <role>
  date: <ISO-8601>
```

Postmortem outputs feed back into the canvas, the register, and the adversarial suite. A postmortem with no hardening tasks is incomplete; the answer cannot be "be more careful."

5. Weekly reliability review

Thirty minutes, one room or one thread. Standing agenda: top three failure_id classes by volume, override rate by agent, open hardening tasks aging past due, any boundary-cross attempts. Output is a one-page status that anyone in the library can read in two minutes. Patterns surfaced here become the next quarter's library-wide priorities.