

FOR AGENT BUILDERS / PLATFORM TEAMS / OPERATORS

Agent Operations Field Kit

Keep autonomy
inside the guardrails.



HALLUCINATION HYDRA / VOLUME III

Twelve checks before an agent acts.

01 / THE RUNTIME CHECKLIST

Pick one real workflow. Verify its effective controls, not just the settings label. Rehearse consequential actions in a safe test environment.

01 The delegation is bounded

- ☐ **Name the accountable human.**
Record purpose, owner, allowed actions, forbidden actions, and escalation conditions.
- ☐ **Separate drafts from authority.**
Require a human decision before generated claims become policy or shared memory.
- ☐ **Define evidence of success.**
Check the intended outcome against independent evidence, not the agent's assurance.

WATCH FOR / Hallucination Hydra

03 The budget can stop work

- ☐ **Bound retries and children.**
Set run limits; make subagents share the parent's spend envelope.
- ☐ **Test a real cap.**
Prove calls stop at the limit. Name who can authorize a deliberate extension.
- ☐ **Count accepted results.**
Include failed attempts, fallback, idle capacity, and human review or repair time.

WATCH FOR / Token Furnace

02 The credential can refuse

- ☐ **Inspect effective access.**
List reachable accounts, tools, and hosts. Check inherited permissions and alternate paths.
- ☐ **Try a forbidden action safely.**
Verify the credential or tool boundary rejects it in a controlled environment.
- ☐ **Rehearse revocation.**
Confirm the old credential fails. Account for copies, sessions, and dependent workloads.

WATCH FOR / Secret Zombie Attack

04 Stop and recovery work

- ☐ **Stop the whole workflow.**
Check children, schedules, and restarts; inspect work already dispatched.
- ☐ **Reconcile uncertain actions.**
Before retrying, check the destination or reuse its supported idempency key.
- ☐ **Prove a safe restart.**
Refresh stale context; restore what can be restored. Confirm remaining external effects.

WATCH FOR / Context Rot Swarm

Before autonomy

An unknown owner, permission boundary, budget, or stop path is a reason to keep this workflow supervised.

02 / THE RUNTIME BESTIARY

The agent is running. So are the risks.

Four Volume III monsters that live in tools, memory, credentials, and the bill.



01 / INVENTED AUTHORITY

Hallucination Hydra

"The other agent cited it."

An invented claim becomes a wiki entry, a memory, and the next agent's trusted premise.

YOUR NEXT MOVE

Separate drafts from trusted sources. Verify the original claim and invalidate its downstream copies.

Ask: Where did this claim first become trusted enough to act on?



02 / COST WITHOUT RESULTS

Token Furnace

"Just one more attempt."

Cheap calls become an expensive workflow through retries, fallback, idle capacity, and human cleanup.

YOUR NEXT MOVE

Measure cost per accepted result. Test a spend cap that includes retries and child agents.

Ask: What useful result did this run buy, including repair time?



03 / ACCESS THAT SURVIVES

Secret Zombie Attack

"We already deleted the key."

A forgotten copy, account, or live session still grants access after the apparent cleanup.

YOUR NEXT MOVE

Revoke and verify. Trace dependent sessions and copies; replace broad authority with a narrow credential.

Ask: What proves every old path of access has stopped working?



04 / STALE OPERATING MEMORY

Context Rot Swarm

"It was in the knowledge base."

Expired instructions and old tool assumptions keep returning as current guidance.

YOUR NEXT MOVE

Assign source owners and expiry dates. Remove stale material and test fresh sessions for its return.

Ask: Who owns this source, and when was it last true?

Keep evidence outside the story. A confident completion message is a claim. Verify the destination, the remaining activity, and the controls that actually fired.

03 / THE OPERATORS YOU NEED

Autonomy needs somebody on duty.

These are responsibilities to assign, not four new job titles to hire. Give the person the access and authority the job needs.



KEEP THE FLEET ACCOUNTED FOR

Agentic Animal Control

Knows which agents still justify their access and cost.

Watch for: A friendly avatar turns an obsolete workload into a permanent pet.

THE OPERATING QUESTION

“Can we retire this agent without it restarting tomorrow?”

Make it real.

Inventory owners, schedules, credentials, and last review dates.



MAKE THE BOUNDARY ENFORCEABLE

Dr. No

Turns a broad request for access into a narrow, usable permission.

Watch for: Permanent obstruction pushes work into hidden credentials.

THE OPERATING QUESTION

“Which forbidden action does this credential actually reject?”

Make it real.

Offer a scoped path with expiry and tested revocation.



CHECK THE CLAIM, NOT THE POLISH

The Qualified Reader

Can reject a fluent answer that does not satisfy the task.

Watch for: Review becomes a bottleneck or a check of formatting alone.

THE OPERATING QUESTION

“What must never ship, even if the answer sounds convincing?”

Make it real.

Set rejection criteria and require independent evidence.



ACT WHEN THE THRESHOLD IS CROSSED

Chief Buzzkill Officer

Calls agent-caused harm an incident and starts containment.

Watch for: Enthusiasm keeps the agent running while damage spreads.

THE OPERATING QUESTION

“What has stopped, and what is still able to act?”

Make it real.

Give the responder authority to pause, revoke, and quarantine.

One person may wear several hats. Every consequential workflow still needs a named owner, a qualified reviewer, and someone who can stop it.

Practice the stop. Prove the recovery.

Choose one non-production agent workflow and a harmless failure scenario. Test one control; record what the rehearsal actually proved.

00-03 MIN	03-08 MIN	08-15 MIN
Choose the failure	Trigger and observe	Reconcile and recover
In a test environment, choose a cap, forbidden action, or stop condition. Name the owner and expected result.	Run the harmless case. Inspect rejected calls, child agents, schedules, and external work already in flight.	Check the destination and preserved evidence. Define what must change before a human permits restart.

Record the control you proved

WRITE IT DOWN

Workflow + failure scenario

Control + observed result

Owner

Check by

Before we permit a restart

COPY INTO YOUR AGENT MANIFEST

Make delegation explicit.

Keep this short record with the workflow and its runbook.

Purpose	This workflow exists to... / The human owner is...
Authority	It may... / It may not... / These credentials enforce it...
Budget	Run, retry, and child-agent limits are...
Stop	When this happens, this person stops it by...
Recovery	Evidence required before restart is...

CONTINUE IN VOLUME III

Running agents is easy. Operating them is work.

The Agentic Workforce goes deeper into operator roles, runtime monsters, and the controls that hold when an agent can act.

EXPLORE VOLUME III

slopcodex.com/books/volume-3

