VCRI
THE VALUE CHAIN
RISK INSTITUTE

VCRI LEARN · AI SECURITY

Is your agent still the agent you approved?

Drift and reconciliation: checking a running AI agent against what was approved.

Joshua Marpet and Cairn Viktor

VCRI
THE VALUE CHAIN RISK INSTITUTE

WHY YOU CARE

Nobody has to touch it for it to change

MCP is the Model Context Protocol, the common way agents connect to tools.

The provider

Repoints a model version alias to a new snapshot. Same name, different model.

The tool author

Changes a tool's definition after you reviewed it, or an MCP server's version pin moves.

The agent itself

Creates an account or publishes a token that keeps working after the agent is gone.

If you only check at approval time, you are checking a picture of an agent that no longer exists.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

TODAY'S TOOLS

Two ways to secure an agent, and the gap between them

Constrain

Limit what the agent can do.

Fails closed and breaks workflows. And an agent optimized to finish the task finds the indirect path: another agent's tool, a person it persuades, content injected upstream.

Judge

Watch what the agent is doing and decide if it is bad.

Fails open, and needs a ground truth about intent that nothing in the stack has.

Measure against a manifest

Compare the agent to a written record of what it is built from and what it may reach.

That ground truth is about composition and reach, not intent, and both can be inventoried.

Source: The Agent Manifest and Agent Reconciliation, VCRI method note v0.2 (valuechainrisk.org/blog).

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

BORROWED FROM INFRASTRUCTURE

Desired state, observed state, reconcile

Infrastructure solved this shape of problem twenty years ago, and did not call it constraint.

Desired state
What it should be. In infrastructure: the configuration database, the golden image.

→

Observed state
What is actually running, inventoried from the live system.

→

Drift
The measured difference between the two.

→

Reconcile
A loop that acts on the difference: replace, kill, or alert.

valuechainrisk.org/learn · CC BY 4.0

THE MODEL

Four terms

Agent manifest

The desired state of one agent, per workflow: model and version, harness, tools and MCP servers with definition hashes, memory, credentials, scopes, and the permissions this job needs and no other.

Agent drift

The measured difference between the manifest and the running agent. A set of quantities, not a verdict, reported per field.

Agent reconciliation

The loop: inventory, diff, act. The action is a small module the owner keeps: open a ticket, revoke a scope, replace the agent, or stop it.

Least variance

The principle: how far may the running agent sit from its manifest before something happens?

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

PER-FIELD DRIFT

What goes in a manifest, and what drift looks like

FIELD	DRIFT SIGNAL
Model identity and version	Provider repoints a versioned alias to a new snapshot
System prompt and policy pack	The hash of the system prompt or policy pack changes
Tool definition version or hash	Definition hash changes after approval
Tool access scope	Scope wider than declared. The core drift field.
Permission exemplar	Any held permission outside the minimum set this workflow needs

Selected rows from Appendix A of the VCRI method note, which maps manifest fields to the OWASP AI Bill of Materials (AIBOM) Foundations Guide v1.0.

valuechainrisk.org/learn · CC BY 4.0

THE PRINCIPLE

Least privilege, least agency, least variance

PRINCIPLE	THE QUESTION IT ASKS
Least privilege	How little may the agent hold?
Least agency	How little may the agent decide?
Least variance	How far may the running agent sit from its manifest before something happens?

Source: The Agent Manifest and Agent Reconciliation, VCRI method note v0.2.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

A REAL EXAMPLE

Plugin4Shell is drift

Declared
A marketplace pins a plugin to a 40-character commit hash.

→

Arrived
The repository owner names a branch with that same hash. Git prefers the branch, so the agent checks out the owner's code.

→

Reported
Git prints one warning. The agent reports the pinned version is installed.

→

Missing step
Nobody compared what was declared to what arrived.

Source: VCRI field note "Plugin4Shell Is Agent Drift" (valuechainrisk.org/blog), on Air Security's Plugin4Shell disclosure, 18 September 2026.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

THE PERSISTENCE FINDING

The agent ends. Its accounts don’t.

What happened

In a UK AI Security Institute (AISI) evaluation, one agent created a GitHub account and published its own access token in a public gist, on purpose, as a channel that would survive.

The agent stopped

It hit its token limit and its run ended. Tunnels it opened died with the machines. The account, token, repositories and DNS records did not.

Others inherited it

Agents in other, isolated runs found the token and used it, one from a different vendor. A found credential was treated as a trusted one.

What ended it

Nothing in the AI systems. GitHub suspended the account after AISI contacted it, days later.

Source: VCRI blog “The Agent Ends. Its Accounts Don’t.” on UK AISI incident report INC-2026-07-28-01 (4 August 2026). AISI calls its analysis preliminary.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

PUT IT IN THE MANIFEST

Your blast radius is what your agent can register

- Can your agent create accounts anywhere? Email, source forges, DNS, package registries, form services.
- If it can, do those accounts belong to you, and can you list them?
- Can your agent publish a credential?
- When a run ends, what did it leave behind? Teardown that reclaims compute but not registrations removes only the ephemeral half.

The four deployer questions from VCRI's "The Agent Ends. Its Accounts Don't."

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

WHAT TO DO MONDAY

Write one manifest, run one loop

- Pick one agent workflow. Write its manifest: model and version, tools with hashes, memory, credentials, scopes, permissions.
- Inventory the running agent (the free AIBOM Constellation Inventory does this) and diff it against the manifest.
- Decide the action before you need it: ticket, revoke a scope, replace the agent, or stop it. Name the owner.
- List every account and token your agents have created, and revoke what nobody owns.

valuechainrisk.org/learn · CC BY 4.0

VCRI

THE VALUE CHAIN
RISK INSTITUTE

Free to use, teach and share.

This deck is part of the VCRI Learn series. Use it in your class, your team meeting or your chapter, with credit to The Value Chain Risk Institute.

valuechainrisk.org/learn

valuechainrisk.org/learn · CC BY 4.0