VCRI
THE VALUE CHAIN RISK INSTITUTE

VCRI LEARN · AI SECURITY

What's in an AI agent?

The agent constellation: why an agent is many parts, and why you inventory them as one unit.

Joshua Marpet and Cairn Viktor

VCRI
THE VALUE CHAIN RISK INSTITUTE

WHY YOU CARE

You approved an agent. Do you know what code it is running?

Most people would answer with the name of the model. That is one star. The agent you run today also depends on a harness, prompts, tools, plugins, memory, credentials and other agents, and any of them can change without anyone telling you.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

VCRI'S FRAMING

> *"Security tooling that targets the model alone is like cooking Thanksgiving dinner and only having cranberry sauce. Incomplete!"*

From The Agent Manifest and Agent Reconciliation, VCRI method note, Joshua Marpet and Cairn Viktor, September 2026.

valuechainrisk.org/learn · CC BY 4.0

THE CONSTELLATION

Eight stars, one agent

Model and version

Which model, which exact version or snapshot.

Harness

The agent framework that runs the loop and calls the model.

System prompt

The standing instructions, plus templates and policy packs.

Tools and MCP servers

What it can call. MCP, the Model Context Protocol, connects agents to tools.

Plugins

Add-ons and skills it loads, usually from a marketplace.

Memory stores

What it remembers between runs: vector stores, files, databases.

Credentials and scopes

The accounts and keys it holds, and what each can reach.

Other agents

The agents it calls, trusts, or hands work to.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

CHANGE ANY STAR

Change one part and the agent changes

WHAT CHANGED	WHAT IT LOOKS LIKE
Model	The provider repoints a version alias to a new snapshot. Same name, different model.
Tool or MCP server	A tool’s definition changes after you approved it, or a version pin quietly moves.
Plugin	The agent installs something other than the pinned version, and reports the pinned one.
Credentials	A scope is wider than declared, or the agent holds a key it should not.
Memory or agents	A memory store it was never granted, or a new agent it now talks to.

Drift examples from VCRI’s Agent Manifest method note and the Plugin4Shell field note (valuechainrisk.org/blog).

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

WHY ONE UNIT

Inventory the whole constellation, together

- Approval should bind to the whole set, not to a model name. The agent is the combination.
- Parts change on different schedules and by different owners: the provider, the vendor, your own team.
- An inventory that misses a component looks complete. That is why discovery and an accountable owner matter.
- Once you have one list, you can compare it to next month’s list. The difference is drift.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

EVIDENCE

Know how you know

Every line of the inventory says which kind of evidence stands behind it.

Observed

The tool hashed the bytes on disk, or read the entry from a configuration file.

Example: a plugin directory, a system prompt file, an MCP server entry in your client config.

Stated

Someone typed it in and nothing on disk was checked.

Example: a hosted model. Its weights are not on your disk, so its version is always stated.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

FREE TOOL

The AIBOM Constellation Inventory

What it does

Lists everything an agent depends on, with versions and content hashes.

Writes the list as a CycloneDX 1.6 AIBOM (AI Bill of Materials), an open format other tools read.

Built to be approved

One Python file, standard library only. No network access. Runs no other programs.

Never writes secret values. Credentials are recorded as a name and a scope.

What it does not do

It does not score, rate or judge anything. It records what is there.

Open source, Apache-2.0.

github.com/value-chain-risk-institute/constellation-inventory

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

TRY IT

Four commands

Write the list, build the AIBOM, check for gaps, and compare against last time.

```
python3 constellation.py init > manifest.json      # starter list; edit it
python3 constellation.py build manifest.json --discover --out aibom.cdx.json
python3 constellation.py check aibom.cdx.json      # missing versions, hashes
python3 constellation.py diff  last.cdx.json aibom.cdx.json   # what changed
```


valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

WATCH

You Don't Know What Code You're Running

A Security Weekly short with Joshua Marpet on the agent constellation: the model, MCP servers, plugins, prompts, memory and tools an AI agent depends on. Find it at valuechainrisk.org/media.

valuechainrisk.org/learn · CC BY 4.0

VCRI
THE VALUE CHAIN RISK INSTITUTE

WHAT TO DO MONDAY

Inventory one agent

- Pick one agent your team runs. Write down all eight stars, even the ones you have to guess.
- Mark each line observed or stated. Chase the stated ones that matter most.
- Run the Constellation Inventory, save the output, and run diff against it next month.
- Ask each AI vendor you buy from: can you give us the constellation of the agent you run for us?

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