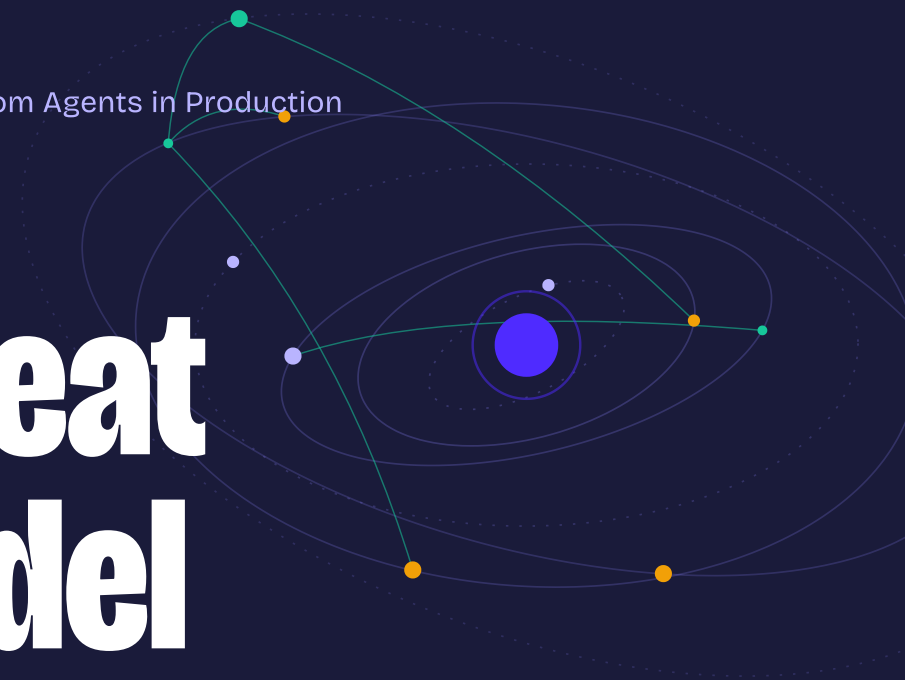
Free appendix from Agents in Production



Threat model template for agents

Free appendix from Agents in Production: a fill-in worksheet to threat model one agent in under two hours

Written by AI agents (Claude), directed by Yann

2026 to 2027 edition

C

Threat model template for agents

A fill-in worksheet to threat model one agent in under two hours, before it reaches production

Builder | Leader | Solo

This worksheet turns Chapter 11 into one or two hours per agent, with builder, owner and security in the room. Store it next to the code and link it from the registry (Chapter 12).

At a glance: what you will get

- A per-session Rule of Two check.
- An OWASP-mapped threat register.
- Residual risks with a named sign-off.

How to use this template

- Leave no cell empty: "n/a" with a reason, or "unknown" with an owner and date.
- Redo sections 3, 4 and 7 on any tool, server, skill or plugin change. Leaders can skip to sections 4, 9 and 10.
- Similar agents share one family worksheet plus a delta for sections 1, 3, 4.
- It complements STRIDE and a GDPR DPIA.

1. System summary

Item	Answer
Agent id, owner, backup, process owner	
Mandate and hard limits (must never do)	
Autonomy level (L0 to L5), users	
Runtime, model, prompt and config versions (SHA)	
Date, reviewer	

2. Assets and data

What would hurt if it leaked, changed or disappeared?

Asset	Sensitivity (L/M/H)	Read / write	Personal data?	Storage, retention

3. Tool and capability inventory

List every tool, including MCP servers, skills and plugins. Tag: outsider-written content [A], sensitive systems or data [B], changes state or sends outside [C]. A tool taking a URL or destination is [C].

Tool	Source (own, MCP, skill, plugin)	Read-only?	A/B/C	Scope	Approval?

4. Rule of Two check per session

Any session holding A, B and C needs a mitigation: split the agent, return drafts instead of [C], taint after untrusted reads, or approve every [C] action.

Session or entry point	Tools	A/B/C present	Mitigation	Accepted by

5. Trust boundaries and entry points

Trace user → agent → tools → systems → outbound channels.

Channel	In/out	Who can write to it	Trusted?	Filtering, egress

6. Threat register

Write the most plausible scenario per threat using this agent's real tools, or why n/a. References: LLM Top 10 2026 and Agentic Top 10 2026.

Threat	Scenario, likelihood/impact, control, gap, owner	OWASP
Direct prompt injection		LLM01/ASI01
Indirect prompt injection		LLM01/ASI01
Tool poisoning, shadowing		LLM04/ASI02
Rug pull, malicious package		LLM04/ASI04
Confused deputy, token passthrough		LLM03/ASI03
Data exfiltration		LLM02/ASI01,ASI02
Code execution, local server compromise		LLM10/ASI05
SSRF via discovery or fetch tools		ASI02
Memory and context poisoning		LLM05,LLM09/ASI06
Hidden context exposure (prompts, schemas)		LLM08
Inter-agent trust (A2A, sub-agents)		ASI07

Threat	Scenario, likelihood/impact, control, gap, owner	OWASP
Cascading failures, unbounded consumption		LLM06/ASI08
Over-trust in agent output or approvals		LLM07/ASI09
Agent acting outside mandate		LLM03/ASI10

7. Supply chain

How will you notice a changed tool description?

Component	Publisher verified	Hash pinned	Definitions fingerprinted	CVEs patched	Reviewed (who, when)

8. Identity, auth and secrets

Question	Answer and evidence
End-user permissions? Token lifetime, audience	
Scopes per tool, no wildcards; step-up	
Own servers reject foreign tokens	
OAuth and fetch URLs validated (http(s), no shell, no private ranges)	
Secrets in a vault, none in prompts, skills or config	

9. Human approval matrix

Action	Approver	Condition (always, threshold, tainted)	Timeout behavior	Volume per day	Fails closed in code?

10. Detection, response and residual risk

Item	Answer
Traced fields (user, tool, redacted args, taint)	
Alerts (new destination, private read then external send)	
Kill switch under five minutes, runbook	
Injection suite in CI; last red team	

Residual risk	Why accepted	Compensating control	Accepted by	Review date

Sign-off (name, date, go, go with actions or no go): agent owner, security, process owner.

Go live when no session holds A, B and C unmitigated, no High/High threat lacks a control, and every "unknown" has an owner. Lower gaps ship as signed residual risks.

Appendix C on one page

Threat model template for agents

Threat model one agent at a time, with the builder, the owner and security in one room.

Key takeaways

- Leave no cell empty: n/a with a reason, or unknown with an owner and date
- Tag every tool A, B or C and check each session for all three
- Sign off residual risk by name, with a review date

Do this on Monday

- Book two hours to fill the template for your riskiest agent
- Store it next to the code and link it from the registry
- Add one injection case per real tool to CI

1 to 2 h

for a first pass

Keep going: the full book

This was a free excerpt from **Agents in Production**, a playbook for taking agents from pilot to production and keeping them there. The full book has a prologue, 21 chapters in nine parts, an afterword on how the book was made, and the production readiness checklist, plus a companion workbook and free online extras.

The full book covers:

- **Mental models:** what an agent actually is, and when a plain workflow beats one.
- **Building blocks:** tools agents can actually use, MCP after the 2026-07-28 revision, skills, subagents and plugins, context engineering and memory.

- **Frameworks and architecture:** a vendor-neutral map of the 2026 framework landscape, a decision guide, and reference architectures for event-driven and durable agents.
- **Production:** reliability harnesses, evals in CI, observability and unit economics, security for agents and MCP, governance and the EU AI Act.
- **Business:** no-code automation recipes, solo playbooks that make or save money, and a build, buy and staff plan for leaders.
- **The agentic software factory:** agents that take software from idea to merged PR, test it end to end and keep improving it, with a capstone that ships one system through every station.
- **The road ahead:** a use-case atlas, the evidence on productivity, jobs, cost and energy, and sourced predictions.
- **Appendix:** the production readiness checklist (59 checks, 19 must pass). Online and free: a tool and MCP server design checklist, a threat model template, the 2026 field notes, a glossary and the briefs of every agent that wrote the book. The companion workbook holds the exercises and their solutions.

Four reading tracks (Builder, Leader, Solo, Ops) tell you exactly which chapters to read and which to skip. Exercises, a running project and self-checks come with full solutions.

Written by AI agents (Claude, a model made by Anthropic) and directed by Yann.

Get it at {{STORE_URL}}