CHEAT SHEET

Securing Agentic AI

A practical checklist for reducing risk when designing, deploying, and operating AI agents.

LLM-based systems need defense in depth, especially when instructions and untrusted data share the same channel.

Christian Schneider
Security Architect · Hacker · Trainer
christian-schneider.net

ZONE 1 Input surfaces

Everything that reaches the context window — not just the prompt.

- **Treat every retrieved artifact as untrusted data.** Documents, email, web pages, API responses, and tool descriptions may contain instructions. Instructions embedded in retrieved content must not be treated as authoritative merely because the content enters the context window.
- **Tag provenance and trust on every chunk.** Record source identity, integrity, tenancy and freshness so validation and retrieval can reflect actual trust.
- **Classify all external content, not just the user turn.** Treat a detector as a routing signal, never as a trust or authorization decision. (NeMo Guardrails · Guardrails AI · Wolf-Defender)
- **Inspect and canonicalize every modality before use.** Extract and inspect text from images and audio, reject ambiguous encodings, remove unnecessary metadata and scan exactly what the model receives.
- **Minimize data before context or tool use.** Redact, tokenize or exclude secrets and personal data that the task does not require.

ZONE 2 Planning and reasoning

The control center. Hijack it and the whole plan changes, not one answer.

- **Bind every run to an immutable task envelope.** Store the goal, scope, constraints and permitted effects outside model control in auditable configuration.
- **Goal-lock plans and actions to the authorized task.** Diff proposed steps and tool calls against the task envelope, not merely the user's text.
- **Use an independent checker with separate context.** Give it only the task envelope and structured plan; treat any second model as advisory, never as authorization.
- **Never authorize on the model's reasoning.** A deterministic, auditable policy engine decides. For high-impact actions, show the exact target, arguments and diff before approval. (OPA)
- **Treat plans, arguments and model output as untrusted input.** Validate and encode for the target context; never hand raw output to a shell, query engine, browser or peer agent.

ZONE 3 Tool execution · MCP · skills

Every tool is a capability and a liability. Descriptions are executable context.

- **Treat tools, skills, identity files and configuration as code.** Review, version and sign them; pin immutable hashes and re-verify on installation, connection and change. (Snyk Agent Scan · Cisco MCP Scanner · NVIDIA SkillSpector)
- **Default-deny permissions and egress per tool and task.** Allow only the operations, data and destinations required for the authorized goal.
- **Never forward the user's token downstream.** Exchange it for an audience-bound, purpose-bound, short-lived token scoped to the current task.
- **Isolate every tool and server.** Give each one a distinct identity, credentials, filesystem, memory scope and network policy.
- **Treat every tool response as untrusted data.** Enforce typed schemas, preserve provenance and keep returned strings out of privileged instruction context.
- **Bound and sandbox every run.** Limit tool calls, time, tokens, cost, retries and delegation fan-out; trip a circuit breaker when limits are exceeded or behavior drifts.

ZONE 4 Memory and state

Where an injection stops being an incident and becomes persistence.

- **Authorize and validate every memory write before commit.** A request to "remember" is untrusted input, even when it arrives through model output or a tool.
- **Rank retrieval by trust, not relevance alone.** Weight provenance, tenancy, integrity, freshness and review status before a memory can influence decisions.
- **Segment memory by tenant, user, session, task and domain.** Shared state can turn one poisoned interaction into everybody's context.
- **Never auto-promote generated output into trusted memory.** Store model- and tool-generated entries as unverified by default, and promote them only after source-backed validation. Reserve human approval for entries that could influence sensitive or high-impact actions.
- **Version and snapshot memory so you can recover.** Keep timestamps and known-good states; support quarantine, rollback and expiry of unverified entries.

ZONE 5 Inter-agent and cross-system

Agents bridge systems that were never meant to reach each other.

- **Treat every agent as a trust boundary.** Scope it like a privileged account, with explicit identity, permissions, audiences and data access.
- **Break the toxic combination.** Do not let one agent combine low-trust input, sensitive data access and high-impact exfil/write without a deterministic gate.
- **Re-validate every message at every hop.** Check sender, audience, authorization, integrity, freshness, schema and semantic intent.
- **Give each agent instance its own short-lived identity.** Bind credentials to the task and purpose; never share keys across agents.
- **Log the full action chain.** Record actor, goal ID, tool, target and arguments, policy decision, result and side effects in tamper-evident logs; alert on deviations.

CROSS-ZONE Model it, then prove it

Attacks are chains. Audit the path, not the components.

- **Model one high-risk scenario end to end.** Turn it into an attack tree spanning every applicable zone; annotate assumptions and the control at each node.
- **Use two controls with different failure modes per high-risk node.** Two classifiers or two prompts are not independent controls.
- **Test each control at its own decision point.** Then propagate the result through the full chain to verify containment.
- **Re-verify whenever the system changes.** Models, prompts, tools, policies, schemas, dependencies and identity scopes can all invalidate prior evidence.
- **Keep an emergency stop.** Be able to revoke identities, disable tools, halt queued work and isolate affected memory or credentials immediately.

IF YOU DO NOTHING ELSE THIS MONTH

1. **Inventory every agent, tool and skill.** Record its owner, identity, permissions, data, destinations and high-impact actions.
2. **Default-deny permissions and egress.** Allow only the operations, data and destinations required for the task.
3. **Gate every high-impact action.** Enforce policy, show the exact target, arguments and diff, then require explicit HITL approval.
4. **Keep external content out of authority.** Preserve provenance; never let it rewrite goals, policy or trusted memory automatically.
5. **Log and contain every action.** Record actor, goal, tool, arguments, decision and result — and maintain an immediate kill switch.

Full series, with the reasoning behind each control: **christian-schneider.net/securing-agentic-ai**
Every engagement starts with a free scoping call: **mail@christian-schneider.net**