DOWNLOADABLE ENGINEERING REFERENCE / PDF EDITION

AI Agent Boundary Review Checklist

This checklist is for teams shipping LLM-enabled tools or agents that can read data, call tools, or trigger actions in production systems.

The core idea: the agent is not the trusted component. The boundaries are.

1) Identify Tools and Authority

- List all tools the agent can call.
- For each tool:
  - what data can it read?
  - what actions can it perform?
  - what is the worst-case impact of misuse?
- Ensure the agent does not hold permanent high-privilege credentials.

2) Policy Enforcement at Boundaries

- Enforce policy in the tool layer, not in the prompt.
- Verify tool requests are validated server-side.
- Make allow/deny decisions testable and auditable.

3) Prompt Injection and Tool Misuse (Defensive Framing)

- Treat user-provided content as untrusted input.
- Expect “instruction smuggling” through:
  - retrieved documents
  - attachments
  - web content
  - user-provided text
- Ensure a compromised prompt cannot escalate tool privileges.

4) Data Leakage Controls

- Avoid placing secrets in the prompt context.
- Validate that logs and traces do not capture sensitive inputs/outputs.
- If retrieval is used:
  - scope what the agent can fetch
  - enforce least privilege on document access

5) Auditability

You should be able to reconstruct:

- what input triggered the action (sanitized)
- which tools were called
- what policy allowed the call
- what output was returned
- who was affected

6) Reliability and Failure Modes

- define safe behavior on:
  - model timeouts
  - partial tool failures
  - hallucinated tool arguments
  - retry storms
- ensure “fail closed” behavior for high-risk actions

7) Outputs

Useful deliverables from an AI boundary review:

- a boundary map (agent, tools, data sources, enforcement points)
- policy model with explicit allow/deny rules
- prioritized risk list tied to impact
- validation plan and monitoring signals

Public engineering reference, not a completed assessment, legal advice, audit, certification, or customer deliverable.

Editable Markdown source: demonicbinary.com/downloads/ai-agent-boundary-review-checklist.md