FREE PRACTITIONER PACK

Agentic AI Deployment Pack

20 checks across 8 control areas. Agents are new attack surface: tool access, data access, and autonomy all need explicit controls.

Pack	Agentic AI Deployment Pack
Document	Security checklist (20 items)
Use	Risk review · go-live gate
Date	2026-09

AnovaCloud — Engineering the AI-Native Enterprise. Frisco, Texas. This pack is a working instrument, not a brochure: complete it with your team, score honestly, and bring it to your discovery call.

AnovaCloud — anovacloud.com

PROMPT & INPUT SAFETY

☐ **Untrusted content is treated as data, never instructions**
Web pages, documents, tickets, and tool outputs are untrusted input. Instruction hierarchy is enforced: system > developer > user > tool output.

☐ **Prompt-injection test cases exist**
A red-team set of injection attempts (direct, indirect via documents, tool-output poisoning) runs in CI.

☐ **Output is validated before action**
Agent outputs that drive tool calls or user-facing text pass schema and policy validation first.

TOOL-USE SANDBOXING

☐ **Tools run with least privilege**
Each tool exposes only the operations the use case needs; destructive operations are separate, gated tools.

☐ **Side-effecting tools require confirmation at the right tier**
Write/delete/execute tools map to the HITL tier per the permission matrix — never open by default.

☐ **Tool inputs are parameterized and escaped**
No string-concatenated commands, queries, or scripts built from model output.

IDENTITY & ACCESS

☐ **The agent has its own identity**
Service principal / workload identity — never a shared human credential, never ambient admin rights.

☐ **Human approvals are authenticated and non-repudiable**
Approvers authenticate; approvals are logged with identity, timestamp, and scope.

SECRETS & CREDENTIALS

☐ **No secrets in prompts, code, or logs**
Secrets live in a vault/KMS; the agent receives short-lived, scoped tokens at runtime.

☐ **Model outputs are scanned for secret leakage**
Responses and tool arguments are checked for credential patterns before leaving the boundary.

DATA EXFILTRATION GUARDS

☐ **Egress is allow-listed**
The agent can only reach approved endpoints; everything else is denied and logged.

☐ **PII handling rules are enforced in the pipeline**
Masking, redaction, or blocking applied before data reaches the model or leaves the boundary.

MODEL & SUPPLY CHAIN

☐ **Model and dependency provenance is recorded**
Which model version, which libraries, which prompts — pinned and recorded per deployment.

☐ **Third-party tools are vetted**
External APIs and MCP-style servers assessed for data handling, auth, and abuse potential before connection.

LOGGING & AUDIT

☐ **Immutable audit log of agent actions**
Every tool call, approval, and override recorded append-only with enough context to reconstruct the decision.

☐ **Logs exclude secrets but include accountability**
Who/what/why per action; retention period defined and enforced.

INCIDENT RESPONSE

☐ **Kill switch is implemented and tested**
A single control halts all agents; tested quarterly like a fire drill.

☐ **Agent-specific incident runbook exists**
Detection, containment, and communication steps for agent misbehavior — distinct from generic app incidents.

☐ **Post-incident review feeds the eval set**
Every incident becomes regression test cases in the evaluation framework.

SIGN-OFF

Security reviewer: _____ Date: _____ Architect: _____ Date: _____

Security sign-off is a launch gate, not a checkbox exercise — unresolved items defer go-live.