AI SECURITY WORKING RESOURCE

AI Agent Production Readiness Worksheet

Record the agent's permissions, controls and unresolved release conditions.

Purpose

Use this workbook before a pilot expands or an AI agent receives production data, credentials, tools, memory, retrieval, or customer-facing authority.

System / program	
Review owner	
Environment	[ ] Prototype [ ] Pilot [ ] Production
Review date / next review	

How to use this worksheet

- [ ] Use one row per system, tool, credential, source, action, or test. Add rows in Word when needed.
- [ ] Use factual system state. Do not place passwords, private keys, access tokens, or unnecessary personal data in the file.
- [ ] Assign an owner and verification step to every blocker. Re-review after architecture, permission, or model changes.
- [ ] This worksheet supports engineering review. It does not replace product-specific testing or a formal security assessment.

Handling note: keep credentials, customer identifiers, and restricted system details out of shareable copies.

WORKSHEET 01

Authority inventory

Create one row for every action the agent can request or perform. Describe the production target, not only the tool name.

Action or capability	Target / data touched	Class	Owner and current control

Review prompts

[ ] Action classes used: read, write, delete, send, execute, approve, retrieve, remember, deploy, or spend.

[ ] Mark any capability that crosses a tenant, customer, environment, financial, or public boundary.

[ ] Record the server-side policy that constrains the action. Do not treat model instructions as authorization.

WORKSHEET 02

Credentials and delegated access

List every token, session, API key, OAuth grant, service identity, or delegated account the agent or its tools can use.

Credential / identity	Owner and environment	Scope and lifetime	Revocation and storage notes

Review prompts

[ ] Prefer user-bound and short-lived access when the action represents a human user.

[ ] Confirm where credentials are stored, how they rotate, and who can revoke them during an incident.

[ ] Flag shared service credentials that combine unrelated tools or environments.

WORKSHEET 03

Memory, retrieval, and data boundaries

Review each durable memory or retrieval source against tenant, role, retention, deletion, trust, and source-authorization requirements.

Control check	Status and notes
A user cannot retrieve material they could not open in the source system.	[ ] Yes [ ] No [ ] N/A
Tenant and document access is checked before content reaches a model, external reranker or user.	[ ] Yes [ ] No [ ] N/A
Memory writes record the user, source, purpose, retention period, and deletion path.	[ ] Yes [ ] No [ ] N/A
Retrieved documents and tool output are treated as untrusted data, not policy.	[ ] Yes [ ] No [ ] N/A
Citations or source references do not reveal restricted titles, paths, or metadata.	[ ] Yes [ ] No [ ] N/A
A user can request correction or deletion when product commitments require it.	[ ] Yes [ ] No [ ] N/A

Boundary exceptions, open questions, and owners

WORKSHEET 04

Approval gates and action records

Record which actions require human approval and which may run within a defined policy. When approval is required, show the exact operation and record the decision before execution.

High-impact action	Approval trigger	Human preview	Record location and rollback

Review prompts

[ ] Preview the exact target, recipient, account, resource, amount, or change before approval.

[ ] Record the actor, policy result, safe argument fields, downstream outcome, time and denied attempts.

[ ] Test emergency disablement without relying on the model or the same tool path being disabled.

WORKSHEET 05

Launch blocker register and decision

A manipulated prompt, retrieved document, or tool result that can cause an unapproved high-impact action is a launch blocker until the authority path is contained.

Blocker or condition	Severity	Required change	Owner	Due / verified

Decision

Release posture	[ ] Ready [ ] Limited pilot [ ] Blocked [ ] Formal assessment needed
Conditions to proceed	
Decision owner / date	
Next verification	