AI Risk Assessment Checklist

RAISEF.AI
AIRAC v0.06 (2025)

Product/System/Use Case:	Assessment ID:
Version IDs (model / data/ prompt/ policy):	Date Created:
Lifecycle Phase: ☐ Ideation/PoC ☐ Design ☐ Development ☐ Testing ☐ Deployment ☐ Monitoring ☐ Other:	
Environment: ☐ Development ☐ Test ☐ Staging ☐ Production ☐ Other:	Evaluator:

Change History

Date	Editor	Change Summary

Instructions

Quick Start: Go item-by-item using the “Check when” criteria, attach evidence and record the approver. Only check a box when the artifact exists, is approved, and meets targets/tolerances (Sec. 5): if higher is better, the result meets or exceeds the target; if lower is better, the result is at or below the tolerance. N/A requires a one-line rationale. Complete GenAI G-items if in scope; otherwise mark N/A. If Conditional Acceptance, related expiry date, ticket number and owner required.

Evidence Quality legend: **L1** – self-attest, **L2** – peer-review, **L3** – independent internal, **L4** – independent external. For **High risk**, Evidence Quality must be **L3+**; and **L2+** for **Medium**.

Sign-off roles: (R) Responsible—prepares & attests; (A) Accountable—approves & accepts risk; co-signs as needed: Legal / Privacy / Security / Safety / Operations / Brand.

See **Detailed Guidance** at the end.

License: Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0). You may share and adapt this checklist for non-commercial use with attribution and indication of changes. Preferred credit: “Richard R. Khan — RAISEF (AIRAC)” + https://raisef.ai/tools/airac + https://creativecommons.org/licenses/by-nc/4.0/. No additional restrictions. See “License & Attribution” at the end for details.

1. Scope & Governance

Notes/Evidence/Rationale

1.01 ☐ Use case defined (intent, boundaries, success criteria)

Check when: use-case document is linked; approver sign-off recorded; success criteria are defined and bounded.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

1.02 ☐ Set risk classification (regulatory/RAISEF) & business criticality

Check when: rubric completed and linked; approver sign-off recorded; classification aligns with documented rules.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

1.03 ☐ Accountable owner & approver(s); key stakeholders named; RACI documented

Check when: RACI table is linked; approver sign-off recorded; roles cover all lifecycle phases.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

1.04 ☐ Risk appetite/tolerances recorded; phase gates (go/no-go) defined

Check when: thresholds and gate criteria are linked; gates reflect the item's risk class; approver sign-off recorded.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

1.05 ☐ Establish oversight model (HITL/HOTL), escalation path, and kill-switch

Check when: oversight criteria and contacts are linked; rollback/kill-switch test results linked; last test recency ≤ tolerance (Sec. 5); approver sign-off recorded.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

1.06 ☐ Identify applicable laws & obligations (privacy, sectoral, IP, consumer, AI regs) and record legal review outcome

Check when: obligations list and counsel decision are linked; any exceptions have approved mitigations or risk-acceptance ticket linked (owner + expiry).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

1.07 □ Assess organizational AI maturity and governance capacity

Check when: organizational AI maturity level (e.g., initial/repeatable/defined/optimized) is documented; capability gaps identified; improvement plan linked.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ____________________________ □ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

1.G. Generative AI Specifics

□ Not Applicable

1.G1 □ Foundation/model family, provider, version & license recorded

Check when: inventory entry is linked; license terms linked; license terms permit intended use.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ____________________________ □ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

1.G2 □ User-data policy for prompts/outputs/memory (collection, retention, purge) defined

Check when: policy link added; retention and purge meet organizational/regulatory requirements; Data Subject Request (DSR)/opt-out flow verification results linked (where required).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ____________________________ □ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

1.G3 □ Disclosure policy (AI-generated labels; limitations notice) finalized

Check when: approved user copy linked; UX screenshot/specification linked; surfaces where required.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ____________________________ □ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

1.G4 □ Media provenance/watermarking approach (e.g., C2PA) chosen

Check when: method and scope are linked; applies to all in-scope output types.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ____________________________ □ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

1.G5 □ Validate training/finetune data rights & consent basis

Check when: sources and bases are listed; gaps resolved or legally risk-accepted (ticket linked: owner + expiry).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ____________________________ □ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

1.Z. Section Approval

Name:	Signature:
Title:	Date:

2. System & Lifecycle Mapping

Notes/Evidence

2.01 ☐ Log architecture diagram (data sources/flows; training vs inference; inputs/outputs)

Check when: current diagram is linked; version matches this release.

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.02 ☐ Models, prompts, tools, integrations, vendors, incl. (Software Bill of Materials) SBOM listed

Check when: inventory/SBOM linked; unknown/unauthorized components count ≤ tolerance (Sec. 5).

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.03 ☐ Human-in/on-the-loop points & decision rights marked

Check when: intervention points are linked; coverage ≥ target (Sec. 5) for the item's risk class.

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.04 ☐ Define environments (dev/test/prod) & deployment/rollback plan

Check when: plan is linked; rollback test results linked; last test recency ≤ tolerance (Sec. 5).

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.G. Generative AI Specifics

☐ Not Applicable

2.G1 ☐ Prompt architecture/governance (system/instructions/policies) documented

Check when: prompt policy and versions are linked; changes tracked with approvals.

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

2.G2 ☐ Sampling/config captured (temperature, top_p, max_tokens, stop)
Check when: current values and change-log linked; all values ≤ tolerance (Sec. 5).

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.G3 ☐ RAG map (vector DB, retriever, chunking, freshness/TTL) if used
Check when: design is linked; freshness ≥ target (Sec. 5).

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.G4 ☐ Review & enforce tool/function-calling permissions (allow/deny, scopes, least privilege)
Check when: allow/deny lists linked; test results linked; least-privilege enforcement rate ≥ target (Sec. 5).

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.G5 ☐ Memory/personalization strategy (consent, retention, user controls)
Check when: strategy linked; opt-out & purge test results linked; success rate for user controls ≥ target (Sec. 5).

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.G6 ☐ Integrate safety pipeline (pre/mid/post moderation) and name vendors/models
Check when: pipeline diagram & thresholds linked; enforcement catch-rate ≥ target (Sec. 5) and test results linked; vendor Service Level Agreements (SLAs) linked.

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

2.G7 ☐ Agent orchestration map & autonomy bounds (single-agent/Multi-Agent Systems (MAS), levels-of-autonomy, termination criteria)
Check when: a current agent-orchestration diagram is linked (topology, coordinator, messaging, and actuation surfaces); levels-of-autonomy per flow are documented; max planning depth/steps/time and concurrency limits are set; termination/rollback criteria and cross-agent kill-switch propagation tests are linked; last test recency ≤ tolerance (Sec. 5).

Responsible: ____________ ____________ ____________ ______
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject

Accountable: ____________ ____________ ____________ ______
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

2.Z. Section Approval	
Name:	Signature:
Title:	Date:

3. Stakeholders & Potential Harms	Notes/Evidence
3.01 □ Identify affected users/groups (incl. vulnerable & accessibility needs) Check when: list is linked; coverage of target markets ≥ target (Sec. 5). Responsible: _____ _____ _____ _____ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: _____ □ Reject Accountable: _____ _____ _____ _____ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
3.02 □ Capture contexts of use & misuse; abuse/dual-use scenarios Check when: scenarios document linked; credible misuse paths ≥ target (Sec. 5) with severities. Responsible: _____ _____ _____ _____ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: _____ □ Reject Accountable: _____ _____ _____ _____ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
3.03 □ Assess harms (safety, fairness/equity, privacy/rights, financial, reputational, environmental) Check when: risk register linked; top harms scored with Likelihood/Impact/(Detectability) (Sec. 5). Responsible: _____ _____ _____ _____ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: _____ □ Reject Accountable: _____ _____ _____ _____ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
3.04 □ Human-oversight failure modes assessed (decision fatigue, vigilance decrement, high-tempo contexts) Check when: a human-factors analysis is linked (expected alerts/hour, review queue lengths, time-to-intervention); oversight load vs. staffing modeled for peak and steady-state; escalation paths for overload are defined; mitigations (rotation, batching, UI affordances) documented. Responsible: _____ _____ _____ _____ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: _____ □ Reject Accountable: _____ _____ _____ _____ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
3.G. Generative AI Specifics □ Not Applicable	
3.G1 □ Assess over-reliance/automation bias & hallucination harms Check when: mitigations (UX copy, confirmations, citations) linked; high-risk flow coverage ≥ target (Sec. 5). Responsible: _____ _____ _____ _____ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: _____ □ Reject Accountable: _____ _____ _____ _____ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

3.G2 ☐ Synthetic media/impersonation/deepfake risk assessed

Check when: policy and response path linked; detection/flag rate for in-scope media ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

3.G3 ☐ Multilingual/locale-specific harms considered

Check when: locales list linked; coverage plan linked; or N/A rationale recorded.

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

3.G4 ☐ Assess harm from code/content generation (e.g., self-harm, illegal, medical/financial advice)

Check when: prohibited domains & escalation paths linked; control test results linked; enforcement coverage ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

3.Z. Section Approval

Name:	Signature:
Title:	Date:

4. Baseline Risk/Threat Catalog

Notes/Evidence

4.01 ☐ Assess accuracy/robustness/drift (incl. out-of-distribution (OOD)/shift)

Check when: baseline & OOD tests linked; metrics ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.02 ☐ Assess bias/fairness/equity across relevant cohorts

Check when: cohorts & metrics linked; gaps ≤ tolerance (Sec. 5) or mitigation plan accepted (link).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.03 □ Assess privacy (leakage, re-ID) & data governance Check when: Data Protection Impact Assessment (DPIA) and leakage tests linked; leakage ≤ tolerance (Sec. 5); Privacy-Enhancing Technologies (PETs) documented (link). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
4.04 □ Assess security (poisoning, evasion, model theft) & supply chain Check when: threat model and vendor review linked; critical open vulnerability count ≤ tolerance (Sec. 5) or formally risk-accepted (ticket linked: owner + expiry). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
4.05 □ Assess misuse/abuse & content safety risks Check when: abuse taxonomy and exposure estimate linked; control coverage for top abuses ≥ target (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
4.06 □ Identify and document IP/compliance/regulatory obligations; confirm licenses/data rights/export controls; record legal sign-off/mitigations Check when: obligations/licensing/export list linked; Legal sign-off linked or risk-acceptance ticket linked (owner + expiry). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
4.07 □ Assess operational resilience & reliability (SPOFs, failover, rate limits) Check when: Single Point of Failure (SPOF) list and failover plan linked; failover test results linked, or justification linked; Recovery Time Objective (RTO)/Recovery Point Objective (RPO) ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4
4.08 □ Assess organizational capacity and readiness for AI operations Check when: capability assessment or maturity report linked; external dependencies identified; training plans and resource commitments confirmed. Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** □ L1 □ L2 □ L3 □ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

4.G. Generative AI Specifics ☐ Not Applicable

4.G1 ☐ Document hallucination/grounding & output factuality risks
Check when: evaluation results linked; target factuality set; eval pass-rate ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.G2 ☐ Assess prompt-injection/data exfiltration & jailbreak risks
Check when: test results linked; attack success ≤ tolerance (Sec. 5).

Responsible: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.G3 ☐ Assess RAG-specific risks (context leakage, retrieval contamination, citation coverage)
Check when: retrieval/citation metrics linked; coverage ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.G4 ☐ Assess toxic/illegal/self-harm content generation risks
Check when: evaluation results linked; category thresholds set; policy violations ≤ tolerance (Sec. 5).

Responsible: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.G5 ☐ Synthetic media (image/audio/video) risks
Check when: likeness/brand policies linked; detection/controls results linked; detection/controls ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.G6 ☐ Assess long-context/truncation, “lost-in-the-middle,” tool-call loops
Check when: stress test results linked; truncation/loop/error rates ≤ tolerance (Sec. 5); limits configured and verified.

Responsible: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________________ ☐ Reject

Accountable: ____________________ ____________________ ____________________ __________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

4.G7 ☐ Agentic/Multi-Agent Systems (MAS) emergent-behavior & recursive-drift risks assessed

Check when: results from closed-loop simulations and MAS stress tests are linked; non-termination/loop rate, invariant-violation rate, unintended tool-chain rate, and policy-deviation metrics are ≤ tolerance (Sec. 5) or risk-accepted (ticket linked: owner + expiry).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

4.Z. Section Approval

Name:	Signature:
Title:	Date:

5. Criteria & Scales

Notes/Evidence

5.01 ☐ Define Likelihood 1–5 with examples

Check when: scale document linked; used in Secs. 4, 7, 9.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

5.02 ☐ Define Impact 1–5 with examples

Check when: scale document linked; used in Secs. 4, 7, 9.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

5.03 ☐ Define Detectability 1–5 (optional) with examples

Check when: either scale document is linked or “not used + rationale” in Notes.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

5.04 ☐ Document “high-risk” threshold & decision rules (per phase/type)

Check when: rules document linked; gates reference these rules.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

5.05 ☐ Assumptions & uncertainty documented Check when: list linked; unknowns tied to follow-ups. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
5.G. Generative AI Specifics ☐ Not Applicable	
5.G1 ☐ Define grounding/confidence scale (e.g., grounded/partial/ungrounded) Check when: scale document linked; applied in 6.G1 & 9. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
5.G2 ☐ Define exposure scale (# users/outputs reach) Check when: scale document linked; applied in prioritization (7). Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
5.G3 ☐ Define reversibility/velocity-of-harm scale Check when: scale document linked; applied in gating. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
5.G4 ☐ Define human-oversight coverage scale (who/when/how) Check when: scale document linked; used to validate 2.03 & 8.03. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
5.G5 ☐ Define autonomy/horizon scale (e.g., AL0–AL4) & gating triggers Check when: scale describing planning horizon, actuation scope, and external-tool reach is linked; each autonomy level maps to oversight requirements, evidence quality, and go/no-go gates; used to validate 2.G7 & 8.07. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

5.G6 ☐ Define oversight workload/alert-fatigue scale & safe reviewer ratios

Check when: scale for alert volume, queue length, reviewer load, and time-to-intervention is linked; thresholds that trigger throttling, rerouting, or additional staffing are defined; used to validate 2.03 & 10.07.

Responsible: ______________ ______________ ______________ ________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject

Accountable: ______________ ______________ ______________ ________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

5.Z. Section Approval

Name:	Signature:
Title:	Date:

6. Evaluation Plan & Evidence

Notes/Evidence

6.01 ☐ Execute data quality/representativeness & lineage checks; publish data card

Check when: checks run & card linked; data quality/representativeness ≥ target (Sec. 5).

Responsible: ______________ ______________ ______________ ________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject

Accountable: ______________ ______________ ______________ ________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

6.02 ☐ Execute performance & robustness metrics (stress/out-of-distribution (OOD)/shift)

Check when: results linked; Key Performance Indicators (KPIs) ≥ target (Sec. 5).

Responsible: ______________ ______________ ______________ ________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject

Accountable: ______________ ______________ ______________ ________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

6.03 ☐ Compute fairness metrics across relevant cohorts

Check when: metrics linked; disparities ≤ tolerance (Sec. 5) or mitigation plan accepted (link).

Responsible: ______________ ______________ ______________ ________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject

Accountable: ______________ ______________ ______________ ________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

6.04 ☐ Produce explainability/interpretability/traceability artifacts

Check when: artifacts linked; fitness-for-purpose ≥ target (Sec. 5).

Responsible: ______________ ______________ ______________ ________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject

Accountable: ______________ ______________ ______________ ________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

6.05 □ Run privacy leakage tests; document Privacy-Enhancing Technologies (PETs) rationale, e.g., Differential Privacy (DP)/Federated Learning (FL)

Check when: tests linked; leakage ≤ tolerance (Sec. 5); PETs documented (link).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: __________________________ □ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

6.06 □ Run adversarial & red-team tests

Check when: report linked; critical open findings count ≤ tolerance (Sec. 5) or formally risk-accepted (ticket linked: owner + expiry).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: __________________________ □ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

6.07 □ Evaluate content safety & guardrails, False Positives/False Negatives (FP/FN) trade-offs

Check when: evaluation set + thresholds linked; FP and FN rates at the chosen operating point ≤ tolerance (Sec. 5).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: __________________________ □ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

6.08 □ Test HITL/HOTL effectiveness under realistic load

Check when: tabletop/usability tests or live fire-drills are linked; intervention latency, reviewer error-catch rate, and override compliance ≥ target (Sec. 5); findings feed training and UX changes.

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: __________________________ □ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

6.G. Generative AI Specifics

□ Not Applicable

6.G1 □ Run hallucination/factuality & grounding evaluation sets

Check when: evaluation sets and results linked; hallucination ≤ tolerance (Sec. 5); grounding ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: __________________________ □ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

6.G2 □ Measure Retrieval-Augmented Generation (RAG) retrieval, recall@k/Mean Reciprocal Rank (MRR), citation coverage/accuracy

Check when: metrics linked; recall@k/MRR and citation coverage/accuracy ≥ target (Sec. 5).

Responsible: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: __________________________ □ Reject

Accountable: ____________________ ____________________ ____________________ ___________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

6.G3 ☐ Execute prompt-injection & jailbreak red-team suites Check when: results linked; attack success ≤ tolerance (Sec. 5). Responsible: ______________ ______________ ______________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________ ☐ Reject Accountable: ______________ ______________ ______________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
6.G4 ☐ Execute toxicity/harassment/Personally Identifiable Information (PII) leakage benchmarks Check when: results linked; rates ≤ tolerance (Sec. 5). Responsible: ______________ ______________ ______________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________ ☐ Reject Accountable: ______________ ______________ ______________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
6.G5 ☐ Run code-gen security tests (secrets/unsafe functions) Check when: findings linked; critical open issues count ≤ tolerance (Sec. 5). Responsible: ______________ ______________ ______________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________ ☐ Reject Accountable: ______________ ______________ ______________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
6.G6 ☐ Validate safety layering (pre/mid/post moderation) Check when: pipeline tests linked; layered catch-rate ≥ target (Sec. 5). Responsible: ______________ ______________ ______________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________ ☐ Reject Accountable: ______________ ______________ ______________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
6.G7 ☐ Test image/video generation safety: Not Safe For Work (NSFW), likeness, brand misuse Check when: tests linked; violation rates ≤ tolerance (Sec. 5). Responsible: ______________ ______________ ______________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________ ☐ Reject Accountable: ______________ ______________ ______________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
6.G8 ☐ Verify watermark/provenance where applicable Check when: steps and outcomes linked; verification pass-rate ≥ target (Sec. 5). Responsible: ______________ ______________ ______________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: __________________ ☐ Reject Accountable: ______________ ______________ ______________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4

6.G9 ☐ Test long-context & multi-turn; tool-call reliability

Check when: tests linked; error/timeout rates ≤ tolerance (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

6.GA ☐ Run closed-loop agent/MAS simulations (long-horizon, self-modifying plans)

Check when: a scenario library (benign + adversarial) and results are linked; termination-on-budget success rate, invariant-check pass-rate, unexpected tool-chain incidence, and loop/non-convergence rates meet targets (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

6.Z. Section Approval

Name:	Signature:
Title:	Date:

7. Analyze & Prioritize

Notes/Evidence

7.01 ☐ Record inherent (pre-mitigation) likelihood/impact/(detectability) per risk

Check when: every risk has values; scales from Sec. 5 used.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

7.02 ☐ Rank top risks; flag single points of failure (SPOFs)

Check when: ordered list linked; SPOFs identified with owners.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

7.03 ☐ Note systemic/cascading risks; compare to appetite/legal constraints

Check when: commentary linked; escalation ticket linked; time-to-escalate ≤ tolerance (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

7.G. Generative AI Specifics ☐ Not Applicable

7.G1 ☐ Elevate high-impact GenAI risks (hallucination, injection, synthetic media)
Check when: risks tagged “GenAI-critical”; gating-review record linked (decision/owner).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

7.Z. Section Approval

Name:	Signature:
Title:	Date:

8. Controls & Mitigations

Notes/Evidence

8.01 ☐ Implement technical controls (data controls, least-privilege, crypto, logging/traceability, rate/usage limits, sandboxes)
Check when: configs linked; negative test results linked; enforcement rate ≥ target (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

8.02 ☐ Implement process/organizational controls (secure SDLC, reviews, change management)
Check when: Standard Operating Procedures (SOPs) linked; adherence sample size ≥ target (Sec. 5) with ticket(s) linked.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

8.03 ☐ Implement human oversight controls (criteria, training, escalation; shadow/veto points)
Check when: materials linked; training completion ≥ target (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

8.04 ☐ Implement UX controls (disclosures, safe defaults, fallback/kill-switch, appeal/recourse)
Check when: UX evidence linked; critical UX safety checks pass-rate ≥ target (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

8.05 ☐ Implement compliance controls: documentation, policy alignment, Data Protection Impact Assessments (DPIAs)/Fairness & Rights Impact Assessments (FRIAs), audit readiness

Check when: artifacts linked; open blocker count ≤ tolerance (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

8.06 ☐ Implement capacity-building and continuous-learning controls

Check when: AI training or awareness program documented; key personnel trained to role-based competencies; learning cadence scheduled.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

8.07 ☐ Enforce autonomy & resource budgets (max steps/horizon/spend; safe-mode downgrade; cascading kill-switches)

Check when: budgets and downgrades are live; breach → automatic downgrade/halt is verified; cross-agent kill-switch propagation tests pass; negative tests show block-rate ≥ target (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

8.08 ☐ Safety invariants & tripwires (pre/post-action checks + replayability)

Check when: a library of domain-specific invariants and tripwires is integrated; violation block-rate and replay auditability ≥ target (Sec. 5); logs support step-level traceability.

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

8.G. Generative AI Specifics

☐ Not Applicable

8.G1 ☐ Enforce Retrieval-Augmented Generation (RAG) grounding with citations; curate index and freshness

Check when: metrics linked; citation accuracy & freshness ≥ target (Sec. 5).

Responsible: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject

Accountable: ______________________ ______________________ ______________________ ____________
Name Title Signature Date

Evidence Quality:
☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

8.G2 ☐ Enforce schema-constrained outputs (JSON/validators); safe prompting patterns Check when: validators active; validation results linked; schema error rate ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ______ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ______ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
8.G3 ☐ Tune content filtering/refusal policies; record precision/recall trade-offs Check when: tuning log linked; False Positives (FP) and False Negatives (FN) rates at the chosen operating point ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ______ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ______ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
8.G4 ☐ Deploy jailbreak/injection mitigations (classifiers/sanitization) Check when: protections on; test results linked; measured attack success ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ______ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ______ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
8.G5 ☐ Define allowed/blocked tools; function permissioning; API quotas Check when: policies live; negative test results linked; block-rate on negative tests ≥ target (Sec. 5). Responsible: ____________ ____________ ____________ ______ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ______ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
8.G6 ☐ Enable media provenance/watermarking for generative outputs Check when: pipeline active; verification results linked; coverage across in-scope output types ≥ target (Sec. 5); verification pass-rate ≥ target (Sec. 5). Responsible: ____________ ____________ ____________ ______ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ______ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
8.G7 ☐ Gate high-stakes outputs to human review Check when: routing rules live; audit samples linked; enforcement compliance rate ≥ target (Sec. 5). Responsible: ____________ ____________ ____________ ______ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ______ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4

8.G8 ☐ Maintain prompt/version control & full audit trail Check when: history retained (link); approval coverage rate ≥ target (Sec. 5) or exceptions documented (ticket linked: owner + expiry). Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
8.G9 ☐ Set cost/latency budgets with throttling Check when: budgets/alerts live (links); key Service Level Objectives (SLOs) remain ≥ target (Sec. 5). Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4

8.Z. Section Approval	
Name:	Signature:
Title:	Date:

9. Decision & Documentation	Notes/Evidence
9.01 ☐ Record residual risk (post-mitigation) scores + rationale Check when: updated scores linked; drivers of residual risk explained. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
9.02 ☐ Record decision (Accept/Mitigate/Defer/Stop) and any conditions Check when: decision and conditions linked; owner assigned. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
9.03 ☐ Sign-offs captured (owner, legal, security, product) Check when: approvers recorded; dates captured. Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: ______________________ ______________________ ______________________ ____________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

9.04 □ File evidence package (system/data map, eval results, risk register, model/data cards, monitoring & incident plan)
Check when: index linked; link resolution error rate ≤ tolerance (Sec. 5).

Responsible: ____________ ____________ ____________ ________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject

Accountable: ____________ ____________ ____________ ________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

9.G. Generative AI Specifics

□ Not Applicable

9.G1 □ Finalize user disclosures (limitations, data use, AI labels)
Check when: copy approved (link); release plan linked; aligns with risk.

Responsible: ____________ ____________ ____________ ________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject

Accountable: ____________ ____________ ____________ ________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

9.G2 □ Align release stage (alpha/beta/GA) with risk; communications reviewed
Check when: stage rationale linked; communications approved (link).

Responsible: ____________ ____________ ____________ ________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject

Accountable: ____________ ____________ ____________ ________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

9.Z. Section Approval

Name:	Signature:
Title:	Date:

10. Operations & Assurance

Notes/Evidence

10.01 □ Set live monitoring metrics & thresholds (performance, drift, safety, abuse, fairness, privacy)
Check when: dashboards + alerts live (links); test alerts fired; alert coverage ≥ target (Sec. 5).

Responsible: ____________ ____________ ____________ ________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject

Accountable: ____________ ____________ ____________ ________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

10.02 □ Verify incident response playbook & on-call contacts
Check when: playbook linked; latest tabletop date linked; tabletop recency ≤ tolerance (Sec. 5).

Responsible: ____________ ____________ ____________ ________
Name Title Signature Date

Decision: □ Accept □ Cond. Accept Conditions/expiry/ticket: ______________ □ Reject

Accountable: ____________ ____________ ____________ ________
Name Title Signature Date

Evidence Quality:
□ L1 □ L2 □ L3 □ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

10.03 ☐ Schedule periodic re-assessment & audits Check when: cadence on calendar (link); re-assessment interval ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.04 ☐ Maintain audit trail & change control; version model/prompt/policy Check when: last 3 changes show approvers; rollback test results linked; mean time to rollback ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.05 ☐ Define decommissioning/rollback; data retention/erasure Check when: plan linked; policy link added; retention periods conform to policy. Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.06 ☐ Periodically review organizational maturity and governance effectiveness Check when: maturity reassessment cadence linked; improvements tracked; external audit or peer review scheduled. Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.07 ☐ Monitor oversight health & auto-throttle exposure Check when: live metrics (alert volume per reviewer, queue length, time-to-intervention, override rates) are tracked; breaching thresholds triggers routing/throttling or additional staffing; monthly review linked. Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.G. Generative AI Specifics ☐ Not Applicable	
10.G1 ☐ Monitor hallucination & policy-violation rates; track False Positives/False Negatives (FP/FN) trends Check when: reports live (links); breach routing defined; detection/violation rates ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ______________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4

© 2025, Richard R. Khan – CC BY-NC 4.0

10.G2 ☐ Track injection/jailbreak attempts; update blocklists/signatures Check when: telemetry live (links); update latency for lists ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.G3 ☐ Track Retrieval-Augmented Generation (RAG) freshness/drift & citation accuracy Check when: metrics linked; freshness & accuracy ≥ target (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.G4 ☐ Manage model/embeddings/prompt updates with canarying Check when: canary results linked; key Service Level Objectives (SLOs) remain ≥ target (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.G5 ☐ Run abuse escalation & user-reporting loops Check when: Service Level Agreements (SLAs) measured (links); escalation response time ≤ tolerance and resolution rate ≥ target (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.G6 ☐ Monitor watermark/provenance efficacy; maintain takedown playbook Check when: periodic spot-checks logged (links); spot-check coverage ≥ target and time-to-takedown ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.G7 ☐ Report compute/latency/cost & sustainability metrics Check when: monthly report linked; metrics ≤ tolerance (Sec. 5). Responsible: ____________ ____________ ____________ ________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ________________ ☐ Reject Accountable: ____________ ____________ ____________ ________ Name Title Signature Date	**Evidence Quality:** ☐ L1 ☐ L2 ☐ L3 ☐ L4

10.G8 ☐ Maintain ongoing red-team cadence Check when: latest report linked; next exercise scheduled; drill cadence interval ≤ tolerance (Sec. 5). Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: _____________________ ______________________ ______________________ ____________ Name Title Signature Date	Evidence Quality: ☐ L1 ☐ L2 ☐ L3 ☐ L4
10.G9 ☐ Monitor agent loops & drift in production Check when: dashboards track steps/run, loop aborts, invariant violations, tool-call fan-out, and plan-vs-outcome deviation; thresholds alert/auto-throttle within tolerance (Sec. 5). Responsible: ______________________ ______________________ ______________________ ____________ Name Title Signature Date Decision: ☐ Accept ☐ Cond. Accept Conditions/expiry/ticket: ____________________________ ☐ Reject Accountable: _____________________ ______________________ ______________________ ____________ Name Title Signature Date	Evidence Quality: ☐ L1 ☐ L2 ☐ L3 ☐ L4

10.Z. Section Approval

Name:	Signature:
Title:	Date:

Detailed Guidance

1. Scope & Governance

Define the use case and boundaries; set risk classification and business criticality; assign accountable owner(s) with a RACI; record risk appetite/tolerances and phase gates; and establish oversight (human-in/on-the-loop), escalation paths, and a working kill-switch. Identify applicable laws and capture legal review. For GenAI, record model family/licenses, user-data policy, disclosures/provenance, and dataset rights. These steps anchor decisions, enforce accountability, and reduce legal, privacy, safety, and reputational risk before later phases proceed.

1.01 Use case defined (intent, boundaries, success criteria)

Define the product/use case in plain terms, purpose, target users, in- and out-of-scope functions, and how success will be measured. Clear scope anchors every later decision (risk class, evidence needs, gates, and tests) and prevents requirement drift. When intent and boundaries are explicit, reviewers can judge whether proposed controls are necessary and sufficient and trace approvals to specific outcomes.

1.02 Set risk classification (regulatory/RAISEF) & business criticality

Classify the initiative's risk level per the prescribed rubric and record its business criticality. Correct classification drives the required evidence quality, oversight model, and gating rigor; misclassification creates legal exposure, weakens controls, and misallocates resources. Aligning with documented rules keeps decisions consistent across teams and phases and sets expectations for escalation and acceptance.

1.03 Accountable owner & approver(s); key stakeholders named; RACI documented

Assign an accountable owner and approver(s) and document a RACI that spans all lifecycle phases. Clear roles eliminate decision gaps, speed escalations, and ensure that legal, privacy, security, safety, operations, and brand functions are engaged at the right moments. A written RACI provides traceability for audits and clarifies who bears responsibility for risk acceptance.

1.04 Risk appetite/tolerances recorded; phase gates (go/no-go) defined

Record the thresholds that express organizational risk appetite and define objective go/no-go criteria for each phase. Explicit tolerances prevent the system from advancing with unresolved high risks and support defensible, repeatable release decisions. Phase gates tied to the item's risk class align teams on what "ready" means and make exceptions visible and accountable.

1.05 Establish oversight model (HITL/HOTL), escalation path, and kill-switch

Define when and how humans supervise decisions (human-in/on-the-loop), who can intervene, and how issues escalate. Implement a working rollback/kill-switch so unsafe behavior can be halted quickly. This structure limits harm propagation from model errors or abuse, ensures high-impact flows receive human judgment, and provides an operational safety net during incidents.

1.06 Identify applicable laws & obligations (privacy, sectoral, IP, consumer, AI regs) and record legal review outcome

List all applicable legal/regulatory obligations and capture counsel's decision, noting any exceptions and how they're risk-accepted. Doing this early reduces privacy, IP, and consumer-protection exposure and prevents costly redesigns later. A documented legal position also clarifies constraints for data use, disclosures, and deployment geography, supporting consistent compliance across releases.

1.07 Assess organizational AI maturity and governance capacity

Evaluate whether the organization has sufficient governance maturity to manage the scope of the AI system. Consider staffing, internal expertise, governance bodies, and prior experience with data-driven systems. Immature organizations should document compensating controls (e.g., external reviews or phased deployment). Misalignment between project risk and maturity increases exposure to compliance, operational, and safety failures.

© 2025, Richard R. Khan – CC BY-NC 4.0

1.G1 Foundation/model family, provider, version & license recorded

Inventory the foundation/model family, provider, version, and license terms for the system. Accurate provenance and licensing ensure the intended use is permitted, support reproducibility and updates, and enable security/vendor reviews. Without this, teams risk breaching license terms, missing critical patches, or losing traceability in audits.

1.G2 User-data policy for prompts/outputs/memory (collection, retention, purge) defined

Define a policy covering what prompt/interaction data and outputs are collected, how long they're retained, where memory is used, and how data are purged and honored for user requests. Clear rules reduce privacy and regulatory risk, limit data-breach blast radius, and align operations with organizational requirements. Documented flows also set expectations for users and downstream teams.

1.G3 Disclosure policy (AI-generated labels; limitations notice) finalized

Finalize standard user-facing disclosures that label AI-generated content and communicate limitations. Transparent messaging mitigates over-reliance, deception, and consumer-protection risk by helping users calibrate trust and take appropriate care. A consistent policy also ensures disclosures appear where required and match approved copy across surfaces.

1.G4 Media provenance/watermarking approach (e.g., C2PA) chosen

Choose and document the method for signaling provenance or watermarking for all in-scope media types. Provenance signals support downstream detection, takedowns, and accountability, reducing impersonation, deepfake, and misinformation risks. A defined approach also harmonizes implementation across products and vendors.

1.G5 Validate training/finetune data rights & consent basis

Validate rights and the consent basis for all training/finetuning datasets, resolving gaps or recording a formal risk acceptance. This prevents IP and privacy violations, reduces litigation and reputational risk, and ensures the model's lineage can withstand audit or challenge. Clear documentation also informs future reuse and decommissioning decisions.

1.Z. Section Approval

Obtain and record section-level approval (name, title, date, signature). Formal sign-off confirms that governance steps have been completed, risks are consciously accepted or escalated, and responsibility is traceable. It also creates an auditable milestone before later phases proceed.

2. System & Lifecycle Mapping

Map the end-to-end system: architecture and data flows, SBOM of models/tools/vendors, human oversight points, and environment separation with deploy/rollback plans. For GenAI, document prompt governance, generation configuration, RAG design, tool/function permissions, memory strategy, and the safety pipeline. This mapping enables privacy/security reviews, reproducibility, and incident response, while least-privilege and versioned configs prevent silent drift and unsafe actuation across the lifecycle.

2.01 Log architecture diagram (data sources/flows; training vs inference; inputs/outputs)

Document and log an end-to-end architecture diagram covering data sources and flows, clear separation between training and inference, and all inputs and outputs. This map gives auditors and engineers a single source of truth for where sensitive data originates and how it moves, enabling privacy, security, and reliability reviews. Accurate, versioned diagrams reduce integration mistakes and speed incident response by showing exactly which components are in play.

2.02 Models, prompts, tools, integrations, vendors, incl. (Software Bill of Materials) SBOM listed

Inventory all models, prompts, tools, integrations, and vendors, and maintain a Software Bill of Materials. A complete inventory is essential for supply-chain and license compliance checks, vulnerability management, and reproducibility. It prevents unapproved or unknown components from entering production, where they can create security, legal, operational, and reputational risk, and provides a basis for change control and vendor accountability.

2.03 Human-in/on-the-loop points & decision rights marked

Mark where humans are in or on the loop and define decision rights for review, escalation, and override. Clear oversight design curbs automation bias and unchecked model actions in high-impact flows, and ensures issues route to accountable roles quickly. Mapping coverage also allows verification that oversight matches the system's risk profile and informs training and staffing plans.

2.04 Define environments (dev/test/prod) & deployment/rollback plan

Define development, test, and production environments, and document a deploy and rollback plan. Clean separation protects data and avoids cross-environment contamination, while a rehearsed rollback path limits downtime and user harm if a release regresses safety or performance. Having this plan codified supports phase gates, incident response, and auditability of changes.

2.G1 Prompt architecture/governance (system/instructions/policies) documented

Document the prompt architecture and governance—including system prompts, instruction layering, and policy constraints—with version history and approvals. Clear governance prevents prompt drift and shadow changes, keeps outputs aligned to policy, and makes investigations reproducible when behavior shifts. This record also enables risk-aware reviews of prompt changes before they reach users.

2.G2 Sampling/config captured (temperature, top_p, max_tokens, stop)

Capture and version the generation configuration—temperature, top_p, max_tokens, and stop sequences—and maintain a change log. Controlled, explainable settings stabilize output quality and variability, support service-level and cost management, and make evaluations comparable over time. Recording changes ensures regressions are traceable and prevents silent parameter shifts that could elevate safety or legal risk.

2.G3 RAG map (vector DB, retriever, chunking, freshness/TTL) if used

Create a Retrieval-Augmented Generation (RAG) map covering the vector store, retrieval method, chunking strategy, and freshness/TTL policies. This blueprint makes context provenance and aging explicit, reducing hallucinations and stale citations and guiding monitoring for index drift. Clear ownership and design notes streamline updates when sources change and enable targeted tests of retrieval quality.

2.G4 Review & enforce tool/function-calling permissions (allow/deny, scopes, least privilege)

Review and enforce tool/function-calling permissions with explicit allow/deny lists, scoped access, and least-privilege defaults. Tight permissioning limits data exfiltration, fraud, and unsafe actuation from prompt-injection or model errors, and provides a defensible control surface for auditors. Documented permissions also accelerate onboarding of new tools without expanding risk unnecessarily.

2.G5 Memory/personalization strategy (consent, retention, user controls)

Define the memory/personalization strategy across consent, retention, and user controls for storing and reusing interaction data. Clear boundaries reduce privacy and regulatory exposure, minimize breach impact through limited retention, and align behavior with user expectations. Documented controls also support DSR/opt-out handling and make cross-device experience predictable.

2.G6 Integrate safety pipeline (pre/mid/post moderation) and name vendors/models

Integrate and document a safety pipeline spanning pre-, mid-, and post-generation moderation, and name the models/vendors involved. A transparent pipeline with calibrated thresholds reduces toxic or policy-violating outputs and sets accountability for third-party services. This structure supports FP/FN trade-offs, vendor SLAs, and incident processes when violations occur.

2.G7 Agent orchestration map & autonomy bounds (single-agent/Multi-Agent System (MAS), levels-of-autonomy, termination criteria)

Document agent orchestration end-to-end: topology, coordinator, messaging, and actuation surfaces; levels of autonomy per flow; and hard limits on planning depth, steps, time, and concurrency. Define termination and rollback criteria, cross-agent kill-switch propagation, and downgrade paths. Exercise negative tests and record recency to prove bounds hold in practice.

2.Z. Section Approval

Obtain and record section-level approval (name, title, date/signature) for System & Lifecycle Mapping. Formal sign-off makes risk acceptance explicit, confirms that mapping and GenAI specifics have been reviewed, and establishes accountability. It also creates an auditable checkpoint before downstream evaluation, gating, and release activities proceed.

3. Stakeholders & Potential Harms

Identify all affected groups, including vulnerable users and accessibility needs, then capture contexts of use and credible misuse/dual-use paths. Assess harms across safety, fairness/equity, privacy/rights, financial, reputational, and environmental dimensions. For GenAI, address over-reliance/hallucinations, synthetic-media/impersonation, multilingual risks, and hazardous content or code generation. This analysis grounds priorities in real-world impact and focuses mitigations and oversight where harm and exposure are highest.

3.01 Identify affected users/groups (incl. vulnerable & accessibility needs)

Identify and document all user and non-user groups affected, explicitly including vulnerable populations and accessibility requirements. Doing so ensures evaluations, UX decisions, and mitigations reflect real-world demographics and needs, reducing fairness, safety, legal, and reputational risk. Clear coverage also anchors later priority-setting and evidence collection by tying harms and tolerances to specific audiences.

3.02 Capture contexts of use & misuse; abuse/dual-use scenarios

Document normal and edge contexts of use alongside credible misuse, abuse, and dual-use pathways with severity considerations. Anticipating how the system can be subverted enables proportionate guardrails, oversight, and routing decisions, limiting safety, security, and legal exposure. This analysis informs downstream testing and gating by focusing attention on high-impact flows and realistic attack surfaces.

3.03 Assess harms (safety, fairness/equity, privacy/rights, financial, reputational, environmental)

Assess and record potential harms across the listed categories, scoring each in a risk register with likelihood/impact (and detectability where used). Consolidated scoring makes trade-offs explicit, supports consistent prioritization, and ties acceptance decisions to documented rationale. This avoids fragmented judgments and ensures material risks are elevated to governance gates and compliance stakeholders.

3.04 Human-oversight failure modes assessed (decision fatigue, vigilance decrement, high-tempo contexts)

Model human-oversight failure modes where alert volume, tempo, or monotony degrade attention. Quantify expected alerts per reviewer, queue lengths, time-to-intervention, and override error-catch rates. Define escalation paths, rotations, batching, and UI aids; set thresholds that pause, throttle, or reroute exposure when staffing or performance falls below safe reviewer ratios.

3.G1 Assess over-reliance/automation bias & hallucination harms

Evaluate the risk that users over-trust outputs and the harms from incorrect or fabricated responses, especially in high-stakes flows. Addressing these failure modes protects user safety and organizational liability by calibrating trust and reducing erroneous actions downstream. Findings guide where confirmations, citations, or other UX safeguards are essential to contain impact.

3.G2 Synthetic media/impersonation/deepfake risk assessed

Assess risks that generated or ingested media could impersonate people, counterfeit brands, or deceive users. Understanding this exposure supports appropriate detection and response paths, reducing fraud, regulatory, and reputational harms. Clear assessment also aligns incident handling and takedown expectations with product scope and threat surface.

3.G3 Multilingual/locale-specific harms considered

Consider how languages, locales, and cultural norms affect output quality and risk profiles; document coverage or justify N/A. This prevents uneven safety or fairness outcomes across regions and user cohorts, avoiding legal, operational, and brand surprises at launch. The analysis guides evaluation set composition and rollout sequencing where risk varies by locale.

3.G4 Assess harm from code/content generation (e.g., self-harm, illegal, medical/financial advice)

Assess the potential for generated code or content to facilitate self-harm, illegal activity, or unsafe medical/financial decisions within the product's scope. Clarifying these domains focuses enforcement and oversight on the highest-risk categories, limiting user harm and compliance exposure. The assessment also informs where escalation paths or prohibited topics are necessary to meet organizational risk tolerance.

3.Z. Section Approval

Obtain and record section-level approval (name, title, date/signature) to confirm the stakeholder analysis and harm assessments are complete and consciously accepted or escalated. Formal sign-off creates an auditable checkpoint, assigns accountability for residual risk, and gates progression to later phases.

4. Baseline Risk/Threat Catalog

Establish baselines for accuracy/robustness/drift, fairness, privacy leakage/data governance, security/supply-chain threats, misuse/content safety, regulatory/IP obligations, and operational resilience. For GenAI, include factuality/grounding risks, prompt-injection/exfiltration, RAG-specific failure modes, toxic/illegal/self-harm content, synthetic media, and long-context/tool-loop issues. A comprehensive threat view guides controls, testing, and gates, preventing normalization of unacceptable risk and informing monitoring thresholds.

4.01 Assess accuracy/robustness/drift (incl. out-of-distribution (OOD)/shift)

Evaluate current model performance for accuracy and robustness, and probe for drift, distribution shift, and OOD fragility. Establishing this baseline prevents silent degradation that can erode product quality, safety, and trust, and it anchors ongoing monitoring thresholds. Clear

© 2025, Richard R. Khan – CC BY-NC 4.0

findings also guide where to harden data, modeling, or oversight paths before exposure scales.

4.02 Assess bias/fairness/equity across relevant cohorts

Measure performance and treatment across the user cohorts that matter for the product, looking for disparate error rates or outcomes. Early detection of inequities limits legal, reputational, and operational risk and informs prioritization of corrective actions. Documented cohort definitions and gaps also enable consistent re-checks as data or usage evolves.

4.03 Assess privacy (leakage, re-ID) & data governance

Assess risks of training or inference leaking sensitive data, enabling re-identification, or violating data-handling rules; document governance for collection, retention, and use. This protects users' rights, reduces breach and regulatory exposure, and constrains blast radius if incidents occur. The assessment clarifies where PETs or stricter access controls are required to meet organizational expectations.

4.04 Assess security (poisoning, evasion, model theft) & supply chain

Map threats across the model and its ecosystem, including data poisoning, evasion, model extraction, and vendor or library weaknesses. Understanding these vectors limits compromise of safety controls, integrity of results, and IP, and it informs where to harden dependencies and verification. A recorded threat view supports accountable risk acceptance and vendor oversight.

4.05 Assess misuse/abuse & content safety risks

Analyze credible misuse and abuse pathways and the product's exposure to harmful or policy-violating content. This frames precision/recall trade-offs for guardrails, reduces harm to users and bystanders, and establishes where escalation or blocking is warranted. Clear risk mapping also guides which scenarios require tighter oversight before scaling.

4.06 Identify and document IP/compliance/regulatory obligations; confirm licenses/data rights/export controls; record legal sign-off/mitigations

Catalog the IP, licensing, data-rights, export-control, and other regulatory obligations that apply, and record the legal position and any accepted mitigations. Doing so prevents unlawful use, costly rework, and reputational harm, and it clarifies permissible deployment scope and data flows. The record also underpins future audits and reuse decisions.

4.07 Assess operational resilience & reliability (SPOFs, failover, rate limits)

Identify single points of failure and evaluate reliability plans, including failover, throttling, and dependencies that could disrupt service. Robustness here limits downtime, cascading incidents, and user harm when models or vendors regress. The assessment sets expectations for recovery and informs capacity, redundancy, and routing decisions.

4.08 Assess organizational capacity and readiness for AI operations

Assess whether internal processes, teams, and resources are adequate to sustain AI lifecycle requirements (testing, monitoring, retraining, security). Weak operational capacity can undermine even well-designed models. Where deficiencies exist, risk acceptance must be explicit, and compensating measures (outsourcing, staggered rollout, external audits) should be documented.

4.G1 Document hallucination/grounding & output factuality risks

Document where the system may fabricate or misstate facts and how grounding is (or is not) ensured. Understanding factuality risk protects users from harmful decisions, reduces legal/brand exposure, and informs where citations or restricted modes are necessary. It also supports clear thresholds for acceptable error in context.

4.G2 Assess prompt-injection/data exfiltration & jailbreak risks

Evaluate susceptibility to adversarial prompts that override policies, extract sensitive context, or cause unsafe tool calls. This limits data loss, fraud, and policy violations and clarifies where defenses or routing need to be strengthened. The assessment also enables targeted monitoring for emerging attack patterns.

4.G3 Assess RAG-specific risks (context leakage, retrieval contamination, citation coverage)

Examine Retrieval-Augmented Generation for risks like leaking private context, retrieving contaminated sources, or weak citation coverage. Making these failure modes explicit reduces hallucinations, stale answers, and privacy incidents, and it guides freshness policies and index hygiene. The output defines what "grounded enough" means for release.

4.G4 Assess toxic/illegal/self-harm content generation risks

Identify the likelihood and impact of generating toxic, illegal, or self-harm content across product flows. This protects users, meets policy and regulatory expectations, and sets the bar for refusal behavior and escalation. Clear articulation of categories and impact supports defensible operating points.

4.G5 Synthetic media (image/audio/video) risks

Assess risks from generating or ingesting synthetic media, including impersonation, deepfakes, and brand misuse. Understanding exposure enables proportionate provenance, detection, and takedown readiness, reducing fraud, safety, and reputational harms. It also clarifies where additional disclosures or constraints are necessary.

4.G6 Assess long-context/truncation, "lost-in-the-middle," tool-call loops

Stress the system for long-context behaviors—token truncation, middle-context loss, and unstable tool-call loops—and document risks. This prevents silent errors, dropped constraints, and runaway actuation that can create safety, cost, or availability incidents. The findings inform safe limits and routing strategies.

4.G7 Agentic/Multi-Agent System (MAS) emergent-behavior & recursive-drift risks assessed

Probe agentic/MAS systems for goal misgeneralization, specification gaming, unsafe fan-out, inter-agent collusion, and recursive plan drift under long horizons. Use closed-loop simulations and MAS stress tests to measure non-termination, invariant violations, unintended tool-chain activation, and policy deviation. Record mitigations or risk acceptances; gate exposure if rates exceed tolerance.

4.Z. Section Approval

Obtain and record formal approval for the section, capturing names, roles, signatures, and dates. This creates an auditable checkpoint that confirms risks in the Baseline Risk/Threat Catalog have been consciously evaluated and accepted or escalated before proceeding. It anchors accountability and supports consistent governance across releases.

5. Criteria & Scales

Define common scales for Likelihood and Impact (and Detectability if used), then codify "high-risk" thresholds and decision rules by phase/type. Record assumptions and uncertainties. For GenAI, add groundedness/confidence, exposure/reach, reversibility/velocity-of-harm, and human-oversight coverage scales. Standardized criteria make scoring comparable, trigger objective gates/escalations, and keep approvals consistent and auditable across teams and releases.

5.01 Define Likelihood 1–5 with examples

Define a 1–5 likelihood scale with clear examples and make it the standard for all scoring. A shared probability yardstick prevents subjective inflation/deflation and enables consistent comparison across

© 2025, Richard R. Khan – CC BY-NC 4.0

baseline risks (Sec. 4), prioritization (Sec. 7), and decision records (Sec. 9). This supports defensible gates and ensures resources track the most probable failure modes, reducing safety, legal, and operational exposure.

5.02 Define Impact 1–5 with examples

Define a 1–5 impact scale with examples that reflect harm severity (e.g., user safety, privacy, financial, legal, reputational). This calibrates what "material" damage means and ties severity to appetite, oversight, and escalation paths. Consistent impact scoring keeps gating decisions proportional to potential harm and avoids under-mitigating high-consequence risks.

5.03 Define Detectability 1–5 (optional) with examples

Decide whether detectability is used; if yes, define a 1–5 scale with examples, and if not, record a brief rationale. Detectability clarifies how reliably and quickly issues will be noticed, shaping triage, monitoring expectations, and residual-risk judgments. A documented stance prevents inconsistent scoring and helps justify control strength when problems are hard to spot.

5.04 Document "high-risk" threshold & decision rules (per phase/type)

Document the numeric thresholds and decision rules that trigger "high-risk" status by phase and use-case type, and ensure gates reference them. These rules translate scales into objective go/no-go criteria, preventing quiet bypass of controls and aligning teams on when escalation, stronger evidence, or additional oversight is mandatory. Clear thresholds create predictable, auditable governance.

5.05 Assumptions & uncertainty documented

List the key assumptions and uncertainties, and tie each unknown to a follow-up action. Making uncertainty explicit reduces overconfidence, surfaces data/model limitations early, and directs additional evaluation where risk is concentrated. This improves planning for safety, privacy, and reliability, and supports transparent residual-risk reasoning later in the process.

5.G1 Define grounding/confidence scale (e.g., grounded/partial/ungrounded)

Define a groundedness/confidence scale for generative outputs (e.g., grounded/partial/ungrounded) and apply it in factuality evaluations and decision logs. This provides a shared language for evidential support, helps calibrate user disclosures and gating for high-stakes flows, and enables tracking of hallucination risk over time.

5.G2 Define exposure scale (# users/outputs reach)

Define an exposure/reach scale based on users affected or outputs produced, and use it to inform prioritization. Exposure determines blast radius: higher-reach features warrant stricter thresholds, faster remediation, and stronger oversight. This scale ensures mitigation effort aligns with potential population-level impact.

5.G3 Define reversibility/velocity-of-harm scale

Define a scale for how quickly harm can occur and how reversible it is, and use it in gating. Rapid or irreversible harms justify conservative releases, tighter controls, and readily available kill-switches. Making this dimension explicit focuses attention on scenarios where delay or rollback is insufficient to protect users and the organization.

5.G4 Define human-oversight coverage scale (who/when/how)

Define a measurable human-oversight coverage scale specifying who reviews, when, and how interventions occur, then use it to validate oversight design and controls. Quantifying coverage reduces automation bias and unsafe actuation, and ties staffing/training to risk. It also creates an auditable link between planned oversight (Sec. 2.03) and implemented controls (Sec. 8.03).

5.G5 Define autonomy/horizon scale (e.g., AL0–AL4) & gating triggers

Define a practical autonomy/horizon scale (e.g., AL0–AL4) describing planning depth, actuation scope, and external-tool reach. Map each level to oversight coverage, evidence quality, canarying, and go/no-go gates. Specify downgrade and halt triggers when budgets or invariants breach. Apply the scale to validate oversight design and release decisions.

5.G6 Define oversight workload/alert-fatigue scale & safe reviewer ratios

Define an oversight workload/alert-fatigue scale with thresholds for alerts per reviewer, queue length, time-to-intervention, and override quality. Establish safe reviewer ratios by risk class and set automatic throttling, rerouting, or staffing triggers when thresholds breach. Use the scale to validate oversight design (Sec. 2.03) and operational monitoring (Sec. 10.07).

5.Z. Section Approval

Record section-level approval (name, title, date, signature). Formal sign-off confirms the scales and rules are complete, applied, and accepted by accountable owners, creating a defensible basis for downstream gating, release, and audit.

6. Evaluation Plan & Evidence

Run data quality/representativeness and lineage checks; measure performance/robustness and fairness; produce explainability/traceability artifacts; and test privacy leakage with PETs rationale, adversarial/red-team suites, and content-safety trade-offs. For GenAI, evaluate hallucination/grounding, RAG retrieval/citation metrics, injection/jailbreak resilience, toxicity/PII, code-gen security, layered safety, media safety/provenance, and long-context/tool reliability. Decision-grade evidence surfaces failure modes early and supports defensible gates.

6.01 Execute data quality/representativeness & lineage checks; publish data card

Execute data quality and representativeness checks, trace lineage, and publish a data card that summarizes sources, sampling, and known limits. This anchors evaluation validity, exposes bias or staleness before results are trusted, and lets reviewers trace issues to specific datasets—reducing accuracy, fairness, privacy, and audit risk. The card also aligns thresholds and interpretability work with the true properties of the data.

6.02 Execute performance & robustness metrics (stress/out-of-distribution (OOD)/shift)

Run performance and robustness evaluations—including stress, OOD, and shift analyses—and record the KPIs. This reveals brittleness under real-world variability, protects reliability and safety, and prevents regressions that could trigger incidents or SLA breaches. Results ground gating decisions and establish baselines for drift monitoring and rollback plans

6.03 Compute fairness metrics across relevant cohorts

Compute fairness metrics across the cohorts that matter for the use case, comparing error rates and outcomes. Quantifying disparities limits equity, legal, and reputational risk, and informs whether rollout controls or mitigations are warranted. Clear cohort definitions and metrics also support repeatable checks as data or usage changes.

6.04 Produce explainability/interpretability/traceability artifacts

Produce explainability, interpretability, and traceability artifacts appropriate to the decision context (e.g., evidence used, decision paths). These enable audit and challenge, support debugging and incident investigation, and let humans-in/on-the-loop make informed interventions. Without them, decisions are opaque, raising safety, legal, and reputational risk.

© 2025, Richard R. Khan – CC BY-NC 4.0

6.05 Run privacy leakage tests; document Privacy-Enhancing Technologies (PETs) rationale, e.g., Differential Privacy (DP)/Federated Learning (FL)

Run privacy-leakage evaluations (e.g., extraction or membership inference) and document why chosen PETs are appropriate. This validates that sensitive information is not exposed and that controls meaningfully reduce breach and regulatory risk. A recorded rationale creates a defensible basis for data handling across environments.

6.06 Run adversarial & red-team tests

Conduct adversarial and red-team exercises that probe abuse paths, policy bypasses, and integrations, and triage findings. This surfaces exploitable weaknesses—such as injection, exfiltration, or unsafe actuation—before exposure scales, reducing security, safety, and operational risk. A structured record supports ownership, timelines, and risk acceptance where necessary.

6.07 Evaluate content safety & guardrails, False Positives/False Negatives (FP/FN) trade-offs

Evaluate content-safety systems and guardrails against a representative set, then select an operating point that balances false positives and false negatives. This calibrates user protection without over-blocking legitimate use, aligning with risk tolerance and legal obligations. Documented trade-offs clarify accountability for edge cases and support consistent enforcement over time.

6.08 Test HITL/HOTL effectiveness under realistic load

Run tabletop exercises and live fire-drills that reflect real alert cadence, ambiguity, and workload. Measure intervention latency, reviewer error-catch and override compliance, and failure patterns by scenario. Use findings to improve reviewer training, summaries, escalation and routing rules, and UX affordances; re-test until targets for high-risk flows are met.

6.G1 Run hallucination/factuality & grounding evaluation sets

Run domain-relevant hallucination/factuality and grounding evaluations. This quantifies misinformation risk and verifies that outputs are appropriately supported by evidence or context, protecting user decisions and organizational credibility. Results inform disclosures, routing, and gating for high-stakes flows.

6.G2 Measure Retrieval-Augmented Generation (RAG) retrieval, recall@k/Mean Reciprocal Rank (MRR), citation coverage/accuracy

Measure retrieval quality for RAG—recall@k, MRR, and citation coverage/accuracy—against curated queries. Strong retrieval underpins grounded answers and traceable citations, reducing hallucination, stale content, and privacy leakage from irrelevant context. These metrics also guide index curation and freshness policies.

6.G3 Execute prompt-injection & jailbreak red-team suites

Execute prompt-injection and jailbreak red-team suites across prompts, tools, and integrations. Demonstrating resilience here mitigates data exfiltration, policy violations, and unsafe tool use triggered by adversarial input. Findings drive hardening and monitoring priorities before broad deployment.

6.G4 Execute toxicity/harassment/Personally Identifiable Information (PII) leakage benchmarks

Run toxicity, harassment, and PII-leakage benchmarks using realistic workloads. This validates that moderation and redaction controls meet expectations, minimizing harm to users and bystanders and reducing legal and brand risk. Clear results inform the chosen operating point and escalation paths.

6.G5 Run code-gen security tests (secrets/unsafe functions)

Test code-generation outputs for security issues such as secret leakage and unsafe APIs. Validating code-gen safety protects downstream systems, limits supply-chain and compliance exposure, and avoids shipping patterns that introduce vulnerabilities for customers. Tracked findings enable targeted safeguards and developer guidance.

6.G6 Validate safety layering (pre/mid/post moderation)

Validate layered safety—pre-, mid-, and post-generation moderation—by exercising the pipeline end-to-end. Defense-in-depth improves catch rates across categories and provides redundancy when a single control fails, reducing harmful outputs and incident load. Evidence supports vendor accountability and threshold selection.

6.G7 Test image/video generation safety: Not Safe For Work (NSFW), likeness, brand misuse

Test image and video generation for NSFW content, likeness misuse, and brand/identity abuse across scenarios. This prevents harmful or unlawful outputs, protects rights holders, and avoids reputational fallout from deepfakes or impersonation. Results guide default blocks, human review triggers, and takedown readiness.

6.G8 Verify watermark/provenance where applicable

Verify that watermarking or provenance signals are correctly applied and detectable where required. Provenance enables downstream detection, attribution, and takedowns, reducing fraud and misinformation risk and supporting platform trust. Verification evidence also aligns expectations with partners and regulators.

6.G9 Test long-context & multi-turn; tool-call reliability

Test long-context and multi-turn behaviors and assess tool-call reliability under realistic sequences. This catches truncation, “lost-in-the-middle” errors, and runaway tool loops that can degrade accuracy, inflate cost/latency, or create safety incidents. Results set safe limits, timeouts, and fallback strategies.

6.GA Run closed-loop agent/Multi-Agent System (MAS) simulations (long-horizon, self-modifying plans)

Run closed-loop agent/MAS simulations using a scenario library with noisy environments, tool failures, conflicting goals, and self-modifying plans. Measure termination-on-budget success, invariant-check pass rate, loop/non-convergence, and unexpected tool-chain incidence. Verify agents respect autonomy bounds and downgrade/kill-switch rules. Record results to inform gating, monitoring thresholds, and remediation playbooks.

6.Z. Section Approval

Record section-level approval with named owner(s), signatures, and dates. Formal sign-off confirms that evaluation plans and evidence are complete, that risks and trade-offs are consciously accepted or escalated, and that accountability is traceable for audits and release gating.

7. Analyze & Prioritize

Assign inherent (pre-mitigation) risk scores using the defined scales, rank the top risks, and flag SPOFs. Note systemic/cascading risks and compare them to appetite and legal constraints to determine escalation. For GenAI, explicitly elevate hallucination, injection/exfiltration, and synthetic-media risks for gating review. Prioritization focuses resources and ensures high-blast-radius issues receive earlier, stricter oversight.

7.01 Record inherent (pre-mitigation) likelihood/impact/(detectability) per risk

Assign baseline, pre-mitigation likelihood and impact (and detectability if used) scores to every identified risk using the §5 scales. This creates a consistent yardstick for comparing disparate issues, prevents subjective inflation/deflation, and enables defensible prioritization and gating. Without inherent scoring, downstream choices about controls and acceptance lack a measurable anchor, increasing safety, legal, and operational exposure.

7.02 Rank top risks; flag single points of failure (SPOFs)

Order the risk register from highest to lowest priority and explicitly flag single points of failure (SPOFs) with ownership. Ranking focuses resources on what most threatens users and the business, while calling out SPOFs surfaces fragility that can trigger outsized outages or harm from a single defect. Clear priorities and SPOF visibility support contingency planning and time-bound mitigation, reducing reliability, reputational, and compliance risk.

7.03 Note systemic/cascading risks; compare to appetite/legal constraints

Document risks with systemic or cascading effects and compare them to established risk appetite and legal constraints to determine whether escalation is required. This surfaces cross-component failures, correlated vendor dependencies, and population-level impacts that can exceed tolerances even when single risks appear acceptable. Aligning analysis to appetite and law enables timely go/no-go decisions and prevents normalization of unacceptable risk.

7.G1 Elevate high-impact GenAI risks (hallucination, injection, synthetic media)

Elevate generative-AI-specific, high-impact risks—hallucination, prompt injection/data exfiltration, and synthetic-media misuse—for dedicated gating review with clear decision ownership. These failure modes can rapidly create user harm, legal exposure, or brand damage at scale, so they warrant earlier and stricter scrutiny than routine defects. Systematic elevation concentrates oversight and mitigations where blast radius is greatest, supporting safe rollout and informed risk acceptance.

7.Z. Section Approval

Record section-level approval (name, title, date, signature) once analysis and prioritization are complete. Formal sign-off makes accountability explicit, confirms that elevated and systemic risks have been consciously accepted or escalated, and creates an auditable checkpoint before controls proceed. This reduces ambiguity in later reviews and ensures decisions reflect organizational appetite and obligations.

8. Controls & Mitigations

Implement technical (data controls, least-privilege, crypto, logging, limits, sandboxes), process/organizational (secure SDLC, reviews, change management), human-oversight, UX (disclosures, safe defaults, fallback/recourse), and compliance controls (documentation, DPIAs/FRIAs, audit readiness). For GenAI, enforce RAG grounding/citations, schema-constrained outputs, tuned filtering/refusals, jailbreak/injection defenses, tool permissions/quotas, media provenance, human gating for high-stakes cases, prompt/version control with audits, and cost/latency budgets. These controls reduce safety, privacy, legal, and operational risk at scale.

8.01 Implement technical controls (data controls, least-privilege, crypto, logging/traceability, rate/usage limits, sandboxes)

Implement foundational technical controls—restrict and encrypt data access, enforce least-privilege, log actions for traceability, throttle usage, and isolate risky execution. These measures curb confidentiality and integrity failures, limit the blast radius of compromise or misuse, and provide the forensic trail needed for incident response. Strong baselines also stabilize higher-level governance and oversight, reducing operational, legal, and reputational risk.

8.02 Implement process/organizational controls (secure SDLC, reviews, change management)

Institutionalize secure SDLC practices, cross-functional reviews, and formal change management. Repeatable processes prevent drift and regressions, ensure accountability for risk-bearing decisions, and align releases with policy and legal constraints. Documented reviews and approvals also create an audit trail that lowers compliance and reputational exposure while enabling safer, faster iteration.

8.03 Implement human oversight controls (criteria, training, escalation; shadow/veto points)

Define when humans must review or intervene, train reviewers, and establish escalation routes with clear shadow/veto points. Effective oversight mitigates automation bias and catches high-impact errors before harm occurs, assigning responsibility where judgment is required. Clear criteria and pathways ensure timely intervention during incidents and maintain legal defensibility.

8.04 Implement UX controls (disclosures, safe defaults, fallback/kill-switch, appeal/recourse)

Embed disclosures and limitations, default to safe behaviors, provide reliable fallbacks/kill-switches, and offer user appeal/recourse. These UX controls calibrate trust, reduce over-reliance, and give users safe exits when outputs are wrong or unsafe, cutting consumer-protection and reputational risk. Clear recourse paths also support incident handling and continuous improvement.

8.05 Implement compliance controls: documentation, policy alignment, Data Protection Impact Assessments (DPIAs)/Fairness & Rights Impact Assessments (FRIAs), audit readiness

Maintain complete documentation, align with governing policies, and complete DPIAs/FRIAs where applicable to ensure audit readiness. Doing so prevents unlawful data use and inequitable outcomes, creates traceable rationale for risk acceptance, and reduces late-stage rework. Strong compliance hygiene protects users and the organization from regulatory, financial, and brand harm.

8.06 Implement capacity-building and continuous-learning controls

Ensure staff understand AI risk, limitations, and governance processes. Capacity-building reduces human error, strengthens oversight, and supports continuous improvement. For SMBs and new entrants, formal training is a critical control compensating for lack of institutional experience.

8.07 Enforce autonomy & resource budgets (max steps/horizon/spend; safe-mode downgrade; cascading kill-switches)

Treat autonomy, step, horizon, and spend budgets as hard safety limits. On approaching limits, or on invariant breach, automatically downgrade to safe-mode, require human confirmation, or terminate. Test cross-agent kill-switch propagation and negative scenarios; record block/downgrade rates and alerting. Budgets should be versioned, monitored, and tied to risk class and environment.

8.08 Safety invariants & tripwires (pre/post-action checks + replayability)

Encode domain-specific safety invariants and tripwires as pre-/post-action checks on tool calls and plan steps. Block and alert on violations, then capture replayable traces for audit and learning. Maintain coverage maps and target block rates; ensure logs support step-level traceability to reproduce incidents and refine invariants over time.

8.G1 Enforce Retrieval-Augmented Generation (RAG) grounding with citations; curate index and freshness

Require answers to be grounded in retrieved sources with visible citations, and actively curate the index for coverage and freshness. Grounding and citation enable verification, reduce hallucinations, and increase user trust, while index hygiene prevents stale or contaminated context. This control underpins factuality for high-stakes decisions and supports auditability.

8.G2 Enforce schema-constrained outputs (JSON/validators); safe prompting patterns

Constrain outputs to strict schemas (e.g., JSON with validators) and use safe prompting patterns. Structured outputs reduce parsing errors and injection of untrusted text into downstream systems, improving reliability and safety. Consistent prompting keeps behavior aligned with policy and simplifies detection and rollback of regressions.

8.G3 Tune content filtering/refusal policies; record precision/recall trade-offs

Tune filtering and refusal behavior to an operating point appropriate to the product's risk and explicitly record precision/recall trade-offs. Clear choices limit harmful outputs without over-blocking legitimate use and provide a defensible rationale for enforcement. Documented trade-offs support governance reviews and consistent treatment of edge cases.

8.G4 Deploy jailbreak/injection mitigations (classifiers/sanitization)

Deploy mitigations that detect and neutralize jailbreaks and prompt-injection attempts, including classifiers and input/output sanitization. These controls reduce data exfiltration, policy bypass, and unsafe tool invocation, limiting security, privacy, and operational harm. Defense-in-depth here protects users and integrated systems from adversarial manipulation.

8.G5 Define allowed/blocked tools; function permissioning; API quotas

Publish explicit allow/deny lists for tools, enforce fine-grained function permissions, and set API quotas. Least-privilege and quotas cap blast radius, cost, and unintended actuation when prompts are manipulated or models err. Clear, enforceable rules create a verifiable control surface and enable safer scaling.

8.G6 Enable media provenance/watermarking for generative outputs

Enable provenance signals or watermarking for generated media across in-scope types. Provenance supports detection, attribution, and takedowns of synthetic or impersonating content, mitigating fraud, rights violations, and brand risk. It also aligns with emerging platform and regulatory expectations for labeling AI-generated media.

8.G7 Gate high-stakes outputs to human review

Route high-stakes outputs to qualified human review before release. Human gating prevents irreversible or severe harm, aligns decisions with risk appetite, and preserves legal defensibility when expert judgment is required. It also provides feedback that strengthens models and policies over time.

8.G8 Maintain prompt/version control & full audit trail

Maintain version control and approvals for prompts and preserve a full audit trail. Traceability prevents shadow edits, enables rapid rollback, and makes behavior reproducible for investigation and audits. This discipline reduces reputational and compliance risk by tying changes to accountable owners.

8.G9 Set cost/latency budgets with throttling

Define cost and latency budgets and enforce them with throttling and alerts. Predictable performance envelopes protect user experience and service reliability while preventing runaway spend or abuse patterns that create availability and safety incidents. Budgets also guide capacity planning and vendor management.

8.Z. Section Approval

Record section-level approval with named approver(s), signatures, and dates. Formal sign-off confirms controls are implemented, risks are consciously accepted or escalated, and responsibility is traceable. This auditable checkpoint gates exposure and aligns accountability before release or scale-up.

9. Decision & Documentation

Record residual risk scores with rationale; log the decision (Accept/Mitigate/Defer/Stop) and conditions with owners; capture required sign-offs; and file a complete evidence package. For GenAI, finalize user disclosures and align release stage (alpha/beta/GA) with residual risk. Clear decisions and traceable documentation prevent launch creep, support audits, and tie exposure to accountable acceptance.

9.01 Record residual risk (post-mitigation) scores + rationale

Record post-mitigation likelihood/impact (and detectability if used) for each prioritized risk and briefly explain the drivers that remain. Clear residual scoring ties implemented controls to the actual risk posture, making go/no-go and risk-acceptance decisions defensible and auditable. It also sets expectations for monitoring and future re-assessment by focusing attention on what remains material after mitigation.

9.02 Record decision (Accept/Mitigate/Defer/Stop) and any conditions

Document the release decision explicitly with accept, further mitigate, defer, or stop, and capture any conditions with an accountable owner. This creates a binding record that connects evidence to action, prevents silent launch creep, and ensures conditional approvals translate into tracked work. A clear decision log reduces operational, legal, and reputational risk by making accountability and follow-through testable in later audits.

9.03 Sign-offs captured (owner, legal, security, product)

Capture dated approvals from the accountable owner and required functions (e.g., legal, security, product). Formal sign-off confirms informed acceptance of the documented risks and evidences cross-functional review. This traceability is essential for audit readiness and for demonstrating that releases align with organizational appetite and regulatory obligations.

9.04 File evidence package (system/data map, eval results, risk register, model/data cards, monitoring & incident plan)

File a complete evidence package behind a stable index, including architecture and data maps, evaluations, risk register, model/data cards, and monitoring/incident plans. Centralizing proof reduces rework, accelerates audits and investigations, and prevents loss of institutional memory across releases. Link integrity and completeness guard against gaps that could undermine compliance, safety, or reproducibility claims.

9.G1 Finalize user disclosures (limitations, data use, AI labels)

Finalize approved user-facing copy and release plans for disclosures about AI use, data handling, and system limitations. Clear, consistent disclosures calibrate user trust, reduce over-reliance and deception risk, and align with consumer-protection duties. Ensuring the language and placement match the product's risk profile lowers legal and reputational exposure at launch.

9.G2 Align release stage (alpha/beta/GA) with risk; communications reviewed

Select the release stage (alpha/beta/GA) to match residual risk and have launch communications reviewed and approved. Staged exposure limits blast radius, ensures appropriate expectations, and provides room to validate controls before broad rollout. Documented rationale and vetted messaging make gating decisions transparent and defensible to auditors and stakeholders.

9.Z. Section Approval

Record section-level approval with name, title, date, and signature. This creates an auditable checkpoint that confirms decisions, disclosures, and evidence are complete and that residual risks are consciously accepted or escalated before proceeding. The approval anchors accountability for the release posture and closes the governance loop for this phase.

10. Operations & Assurance

Instrument live monitoring and alerts (performance, drift, safety/abuse, fairness, privacy); verify incident response and on-call readiness; schedule re-assessments/audits; maintain change control and audit trails; and define decommissioning/rollback and data retention/erasure. For GenAI, monitor hallucination/policy-violation and injection attempts, track RAG freshness/citation accuracy, canary model/prompt changes, run abuse-escalation loops, check provenance efficacy, report compute/latency/cost/sustainability, and sustain red-team cadence. Continuous assurance guards against drift and emergent harms.

10.01 Set live monitoring metrics & thresholds (performance, drift, safety, abuse, fairness, privacy)

Define and instrument live monitoring across performance, drift, safety, abuse, fairness, and privacy with explicit thresholds and alerting tied to owners. Continuous observability turns evaluation snapshots into ongoing assurance, enabling rapid detection, triage, and rollback when models, data, or behavior shift in production. Without this, degradation and policy violations remain invisible, amplifying legal, reputational, and operational risk as usage scales.

10.02 Verify incident response playbook & on-call contacts

Verify an incident-response playbook and keep on-call contacts current; exercise the plan so roles, communications, and escalation paths are clear. Practiced response minimizes time to detect and contain failures such as model regressions, policy breaches, or data leaks. Preparedness reduces downstream harm to users and operations and demonstrates accountable, rapid remediation.

10.03 Schedule periodic re-assessment & audits

Place periodic re-assessments and audits on the calendar for models, data, controls, and vendors. Scheduled reviews catch drift, emerging threats, and regulatory changes that invalidate earlier assumptions, and they verify that mitigations remain effective. Regular audits also produce defensible evidence of ongoing due diligence for regulators and customers while guiding re-prioritization of risk work.

10.04 Maintain audit trail & change control; version model/prompt/policy

Maintain a complete audit trail and structured change control for models, prompts, and policies, including versioning, approvals, and rollback procedures. Traceability and disciplined release management deter shadow changes and enable rapid root-cause analysis when behavior shifts. Strong change control limits blast radius, supports reproducibility and compliance, and shortens time to restore safe service after a regression.

10.05 Define decommissioning/rollback; data retention/erasure

Define decommissioning and rollback procedures alongside retention and erasure rules for data and artifacts. Planned retirement prevents orphaned systems from lingering with unresolved liabilities, and clear retention/erasure aligns operations with privacy and contractual obligations. Thoughtful rollback paths protect users during reversions and ensure historical evidence is preserved appropriately for audit while sensitive data are removed on schedule.

10.06 Periodically review organizational maturity and governance effectiveness

AI governance maturity should evolve as the organization gains experience. Regular reassessment ensures that governance, controls, and expertise remain aligned with system complexity and scale. Stagnation in maturity can erode safety, fairness, and compliance assurance over time.

10.07 Monitor oversight health & auto-throttle exposure

Treat oversight health as an SLO. Continuously track alerts per reviewer, queue length, time-to-intervention, override quality, and escalation latency. When thresholds breach, auto-throttle or shed low-value traffic, reroute to standby reviewers, or trigger staffing. Review trends monthly and adjust thresholds, training, and routing to keep high-risk flows within tolerance.

10.G1 Monitor hallucination & policy-violation rates; track False Positives/False Negatives (FP/FN) trends

Monitor hallucination rates and policy-violation incidents in production and track evolving false-positive/false-negative trade-offs. These signals validate that chosen operating points remain safe and that disclosure, routing, or moderation stays calibrated to real usage, not just test sets. Continuous measurement enables targeted hardening and prevents silent drifts that could mislead users or breach trust.

10.G2 Track injection/jailbreak attempts; update blocklists/signatures

Instrument telemetry for prompt-injection and jailbreak attempts and maintain blocklists/signatures with timely updates. Visibility into attack patterns and response latency lowers the risk of data exfiltration, unsafe tool invocation, and policy bypass. Routine updates turn post-mortem lessons into proactive defenses and provide a measurable deterrence posture.

10.G3 Track Retrieval-Augmented Generation (RAG) freshness/drift & citation accuracy

Track RAG index freshness, retrieval drift, and citation accuracy in live traffic. Monitoring ensures that sources remain current and relevant, that retrieval quality does not decay, and that citations continue to support outputs. Without these checks, grounded answers can turn stale or misleading, increasing factual, legal, and reputational risk.

10.G4 Manage model/embeddings/prompt updates with canarying

Manage updates to models, embeddings, and prompts with canary releases and health gates before broad rollout. Gradual exposure confines the blast radius of regressions in accuracy, safety, or latency, enabling rollback based on evidence rather than intuition. Canarying preserves service levels while allowing controlled experimentation and faster, safer iteration.

10.G5 Run abuse escalation & user-reporting loops

Run user-reporting and abuse-escalation loops with measured handoffs from intake to resolution. Direct feedback channels broaden detection beyond automated filters, surface emergent harms, and provide context for tuning guardrails. Efficient escalation and closure protect users, reduce legal exposure, and demonstrate accountable operations to auditors and partners.

10.G6 Monitor watermark/provenance efficacy; maintain takedown playbook

Monitor real-world efficacy of watermarking/provenance mechanisms and keep a takedown playbook ready with roles and partners. Regular spot-checks and coordinated removals curb impersonation, deepfakes, and brand misuse that escape initial controls. Operational readiness shortens time-to-takedown and supports trust with affected users, rights holders, and regulators.

10.G7 Report compute/latency/cost & sustainability metrics

Publish recurring reports on compute, latency, and cost, including sustainability metrics tied to utilization. Visibility into resource efficiency and performance underpins capacity planning, budget control, and service-level reliability. Tracking environmental impact also aligns operations with organizational goals and stakeholder expectations.

10.G8 Maintain ongoing red-team cadence

Maintain an ongoing red-team cadence that exercises the production system, not just pre-release builds. Regular adversarial probing uncovers new jailbreaks, data-exfil paths, and unsafe emergent behaviors created by updates or scale. A living program keeps defenses current and provides fresh evidence to inform gating and monitoring adjustments.

10.G9 Monitor agent loops & drift in production

Instrument production telemetry for agent loops and drift: steps per run, loop aborts, invariant violations, tool-call fan-out, plan-versus-outcome deviation, and convergence time. Alert when limits breach; auto-throttle, downgrade autonomy, or trigger kill-switches. Correlate incidents to updates/canaries and record mitigations, keeping long-horizon behavior within tolerance as exposure scales.

10.Z. Section Approval

Record section-level approval with names, roles, signatures, and dates to confirm Operations & Assurance controls are in place. Formal sign-off makes risk acceptance explicit, anchors accountability for ongoing monitoring and response, and creates an auditable checkpoint before further exposure or scale-up. This closes the governance loop for steady-state operations.

License & Attribution

This work—AI Risk Assessment Checklist (AIRAC)—is licensed under Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0). Canonical license: https://creativecommons.org/licenses/by-nc/4.0/ (see Legal Code for full terms).

What you may do

- Share – copy and redistribute in any medium or format.
- Adapt – remix, transform, and build upon the material.

These freedoms are irrevocable as long as you follow the terms.

What you must do

- Attribution (TASL): Include Title, Author, Source (link), and License (link); indicate if changes were made. Example:

“AI Risk Assessment Checklist (AIRAC v9.9, YYYY) by Richard R. Khan — RAISEF, source link, licensed CC BY-NC 4.0; changes: [describe].”

- NonCommercial: Use is limited to purposes not primarily intended for or directed toward commercial advantage or monetary compensation.
- No endorsement: Do not imply that the licensor endorses you or your use.
- No additional restrictions: You may not apply legal terms or technological measures that restrict others from doing anything the license permits. See CC BY-NC 4.0, Section 2: No downstream restrictions.

What is not covered

- Trademarks & logos: “RAISEF” and related marks/logos are not licensed under CC; do not reuse them without permission.
- Third-party materials: Any third-party content included here may have separate rights; mark such items and their licenses and treat them as excluded from CC BY-NC.

Attribution guidance

- Use TASL and link to the license; where attribution details are long, CC 4.0 allows you to point to a page that consolidates them.
- For web/PDFs, mark the work and embed license metadata (e.g., license URL in PDF/XMP) so it’s human- and machine-readable.

Commercial use

For uses outside “NonCommercial” (e.g., resale, paid training bundles, inclusion in commercial products), please request a separate commercial license: licensing@raisef.ai.

Disclaimer

The license and related information are provided “as-is”; Creative Commons is not a law firm and provides no warranties.