Enforceable AI Risk Assessment Checklist

1. Design & Scope (MAP / GOVERN)

Objective: Make intent, ownership, and risk boundaries explicit before anything is built.

Control Checks

☐ System purpose, decision scope, and autonomy level explicitly defined

☐ Impact tier assigned (low, medium, high) with documented rationale

☐ Affected users, data subjects, and downstream systems identified

☐ Human override requirements defined for high-impact decisions

☐ Approval authority defined by impact tier (product owner, risk committee, exec sponsor)

☐ Regulatory obligations mapped based on use case and geography

Evidence Artifacts

- System definition document
- Risk classification record
- Model card or system datasheet (draft)
- Approval authority matrix

2. Build & Validate (MEASURE + Pre-Production Gates)

Objective: Prevent unsafe systems from reaching production through eval-driven development.

Control Checks

☐ **Eval-driven development in place**

  - Quality, safety, security, and policy evals defined
  - Evaluation datasets documented
  - Thresholds approved and versioned

☐ **CI/CD promotion gates enforced**

  - Models or agents cannot deploy if eval thresholds fail

- Regression eval suite runs on every model or agent change

☐ **Security testing completed**

- Prompt injection testing (direct and indirect)
- Data leakage and inversion testing
- Misuse and jailbreak scenarios exercised

☐ **Fairness and reliability validated**

- Fairness metrics selected with thresholds
- Cohort-level performance evaluated
- Calibration error measured

Evidence Artifacts

- Eval results with thresholds
- CI/CD gate configuration
- Security testing report
- Fairness and reliability metrics

3. Deploy & Enforce (Runtime Governance / TRiSM Controls)

Objective: Enforce risk controls during live operation, not after failure.

Control Checks

☐ AI runtime defense or gateway deployed

☐ Prompt and output inspection enabled

☐ Blocking, redaction, and rate limits configured

☐ Tool calls monitored and constrained

☐ Policy enforcement applied consistently across environments

☐ **Information governance enforced**

- Data classification applied to inputs and outputs
- Access controls and retention policies enforced
- Sensitive data handling verified

Evidence Artifacts

- Runtime policy configuration
- Gateway or guardian architecture
- Data access control definitions
- Runtime enforcement logs

4. Operate & Improve (Continuous Evals + Incident Response)

Objective: Detect, contain, and recover from failures in production.

Control Checks

☐ Continuous evals running in production

☐ Monitored signals explicitly defined:

  - Drift magnitude
  - Hallucination rate
  - PII leakage events
  - Prompt injection success rate
  - Tool-call anomalies
  - Latency and cost budgets

☐ Incident response playbook exists and is actionable

  - RACI defined
  - Response SLAs documented
  - Tabletop exercise completed

☐ Retraining, rollback, and shutdown triggers automated

Evidence Artifacts

- Monitoring dashboards or metric definitions
- Incident runbook and drill notes
- Alert thresholds and response logs

5. Agentic/Tool-Calling Systems (Applied When Relevant)

Objective: Control the dominant failure mode in modern AI systems: agent access and tool misuse.

Control Checks

☐ Tool registry maintained (tool, purpose, risk tier)

☐ Least privilege enforced per tool

☐ User identity propagated through agent and tool chains

☐ Approval gates for high-impact tool calls

☐ MCP or agent-to-agent integrations governed and authenticated

☐ Indirect prompt injection controls for RAG and untrusted sources

☐ Agent memory controls defined (scope, retention, poisoning detection, isolation)

Evidence Artifacts

- Tool registry and permission matrix
- Identity propagation design
- MCP governance configuration
- Agent memory policy documentation

6. TRiSM Capability Coverage Check (Executive View)

Objective: Expose foundational gaps quickly.

Capability Presence

☐ AI system inventory exists (models, apps, agents, embedded AI)

☐ AI runtime defense or gateway deployed

☐ AI security testing program exists

☐ AI supply chain or AI-BOM tracked

☐ Information governance controls implemented

☐ AI usage control for embedded and external AI enforced

Missing any of the above indicates systemic risk, regardless of policy maturity.