

AI Safety Review Checklist

The checklist I run in an AI safety review: forty checks across scope, guardrails, evaluation, reversal, governance and operations.

[Download PDF](#)[Print](#)

1. Scope and permissions

- Decision inventory exists and is current
- Each decision has an autonomy tier
- Tools per agent scoped to minimum
- No agent can call another agent's tools
- Data access limited to what the decision needs
- PII stripped or masked before the model

2. Guardrails

- Hard bounds (caps, floors, ceilings) in code
- Step limits per decision cycle
- Rate limits per agent, per segment
- Policy checks after the model, before the action
- Executing system re-validates independently
- Guardrail-hit rate is monitored and alerted

3. Evaluation

- Fixed evaluation set of real cases, 50+
- Scorer defined per decision
- Every change runs the suite before release
- Suite re-run nightly on fresh cases
- Cascade tiers measured for false negatives
- Cost and latency per decision tracked

4. Reversal and audit

- Snapshot before every action
- Rollback window defined and tested
- Batch rollback by agent and time range
- Audit record: inputs, proposal, guardrails, action, approver
- Audit store is append-only and agent cannot write to it
- Records retained per policy

5. Governance

- Named owner per agent
- Escalation rules documented
- Human-in-the-loop points defined
- Compliance and legal have reviewed the audit format
- Customer-facing disclosures in place
- Vendor and model terms reviewed

6. Operations

- Kill switch, tested, usable by non-engineers
- Incident runbook exists
- On-call knows the agent exists

- Dashboards: decisions, hits, cost, accuracy
- Weekly sample review of decisions
- Monthly review of tool-use logs