PHANTOM WORKFLOW

Human review and AI agent guardrails

Define permissions, review evidence, and recovery procedures for operational AI agents.

Last reviewed: 2026-10-05

This is a planning guide for your own workflow. It is not a client result or a finished build. The guide text matches the web page.

Read this whitepaper online

Define the action boundary

List the information an agent may prepare, the actions it may propose, and the actions it may execute. Decisions about money, access, customers, or regulated work need explicit ownership and approved boundaries.

A rule or prompt is not proof that an AI system cannot make mistakes. Describe observable controls and failure tests rather than claiming an error-free system.

Make review actionable

A review request should identify the task, source information, missing context, proposed action, and reason approval is required. The reviewer should be able to approve, reject, or ask for more information.

An alert is not a completed review. Define what the system may do while waiting and how a stalled review is assigned or escalated. Avoid sending a customer confirmation before the decision is recorded.

Plan for changed context

Before executing an approved action, verify that the relevant business record has not changed in a way that invalidates the decision. An old approval may refer to a different appointment, amount, or request state.

Keep a recovery procedure for failed actions and uncertain results. For webhook-based workflows, use the provider's signature requirements and document how repeated or delayed notifications are handled.

Review the risk framework

NIST provides a framework for governing, mapping, measuring, and managing AI risk. Stripe documents webhook signature verification. These are references for a review process, not certifications or proof of a deployed control.

Test rejected actions, unanswered review requests, missing evidence, changed records, and invalid notifications. Have the process owner verify the final state and the trail of decisions.

Decisions to write down

Checkpoint	Decision
Permissions	What may the system read, propose, and change?
Reviewer	Who owns each approval and what evidence do they receive?
Waiting state	What happens while review is pending or overdue?
Recovery	Who resolves a failed or uncertain action?

Sources

NIST: AI Risk Management Framework. Checked 2026-10-05.

Stripe: Resolve webhook signature verification errors. Checked 2026-10-05.

Related guides

How to define human review for operational AI agents

How to review webhook recovery and failed-event queues