SNAPWIRE

Vanguard User Guide

Generated: March 08, 2026 at 08:00 UTC

1. Hold Window

The Hold Window is Snapwire's human-in-the-loop control mechanism. When an AI agent makes a tool call that triggers a constitutional rule, the action is held for human review instead of being immediately blocked or approved.

How It Works

- When a tool call matches a hold-eligible rule, Snapwire returns HTTP 202 (Accepted) to the calling agent, signaling the action is pending review.
- The configurable hold timeout ranges from 0 to 60 seconds. If no human responds within the timeout, the action is auto-released according to your configured policy (approve or deny).
- Operators can approve or deny held actions from the Snapwire dashboard, Slack, or via the API.
- All hold decisions are logged in the immutable audit trail with the operator identity, timestamp, and decision rationale.

Configuration

Set the hold window duration in your tenant settings (0-60 seconds). A value of 0 disables the hold window and reverts to immediate block/approve behavior. The recommended setting for production deployments is 30 seconds.

2. Slack Alerts

Snapwire integrates with Slack via Socket Mode to deliver real-time alerts when actions are held for review. This enables operators to triage agent actions without leaving their existing workflow.

Setup

- Create a Slack app at api.slack.com with Socket Mode enabled.
- Add the SLACK_BOT_TOKEN and SLACK_APP_TOKEN environment variables to your Snapwire instance.
- Configure the notification channel in your Snapwire dashboard under Settings > Webhooks & Notifications.
- Snapwire will send Block Kit messages to the configured channel whenever an action is held.

Approve / Kill Buttons

Each Slack alert includes interactive Approve and Kill buttons. Clicking Approve releases the held action and allows the agent to proceed. Clicking Kill permanently denies the action. Both decisions are recorded in the audit log with the Slack user identity (slack:<username>) as the resolver, providing full traceability.

3. Weekly Compliance Digest

Every Friday at 9:00 AM UTC, Snapwire generates and distributes a Weekly Compliance Digest summarizing the prior week's agent activity, enforcement decisions, and compliance posture.

Digest Contents

- Total tool calls intercepted and evaluated during the week.
- Breakdown of approved, blocked, held, and shadow-blocked actions.
- Top triggered rules and most active agents.
- Risk score trends and high-risk action summary.
- SHA-256 fingerprint of the audit log for the reporting period, enabling independent verification of log integrity.

Distribution

The digest is sent via Slack (if configured) and email (if email notifications are enabled in tenant settings). It can also be accessed on-demand via the Snapwire dashboard.

4. Safety PDF & Compliance Reporting

Snapwire generates a comprehensive Safety Disclosure PDF that documents your instance's compliance posture, active safeguards, and NIST IR 8596 coverage grade.

Using the Safety PDF

- Download from the /safety page or via /safety/pdf at any time.
- The PDF includes your current NIST grade, coverage breakdown by CSF 2.0 function, active safeguards list, and audit log fingerprint.
- Use it as supporting documentation for NIST AI RMF compliance filings.
- Attach it to Colorado SB24-205 impact assessments as evidence of algorithmic discrimination protections and human oversight.
- The audit log fingerprint (SHA-256) enables independent verification that your logs have not been tampered with.

Regulatory Alignment

- NIST IR 8596 (AI Agent Security Profile / CSF 2.0): Coverage grading across Govern, Protect, Detect, Respond, and Recover functions.
- Colorado SB24-205 (AI Consumer Protections): Affirmative defense evidence including human oversight records, discrimination testing, and safety disclosures.
- OWASP Top 10 for LLM Applications: Safeguards mapped to prompt injection, insecure output handling, and other LLM-specific threats.