

Threshold & Escalation Specification

Defines how confidence_band maps to action, when the agent must stop, and which actions cross the irreversible-impact boundary and therefore always require a HITL gate regardless of confidence. This spec is read directly by the Fallback Architecture and the HITL Review Policy.

1. Confidence band mapping

The agent emits a numeric routing_confidence (0.0-1.0) plus a confidence_band from the locked glossary enum {high, medium, low, unknown}. Bands are not arbitrary buckets; they map to action authority.

confidence_band	Numeric range	Default action authority	Notes
high	>= 0.85	Act autonomously within scope	Still subject to irreversible-impact gate
medium	0.65 - 0.849	Act, but log evidence and post to review queue	Reviewer can override post-hoc
low	0.35 - 0.649	Do not act. Escalate with top-2 candidates.	HITL must decide before action
unknown	n/a or signal missing	Halt. Treat as failure-mode low-confidence-routing.	Almost always paired with schema-drift-input

Numeric ranges are defaults. Per-agent overrides MUST be declared on the Operating Context Canvas and re-evaluated each release.

2. Decision table

The agent's action authority is the intersection of confidence band, reversibility, and irreversible-impact boundary membership.

confidence_band	output.reversibility	Crosses irreversible boundary?	Authorized action
high	reversible	no	autonomous-execute
high	partially-reversible	no	autonomous-execute + post-hoc review
high	irreversible	no	HITL gate required
high	any	yes	HITL gate required (no exceptions)
medium	reversible	no	autonomous-execute + same-day review

confidence_band	output.reversibility	Crosses irreversible boundary?	Authorized action
medium	partially-reversible	no	propose-and-wait (default-approve after SLA)
medium	irreversible	no	HITL gate required
low	reversible	no	propose-and-wait
low	any	no	HITL gate required
any	any	yes	HITL gate required
unknown	any	any	halt; emit low-confidence-routing

propose-and-wait means the agent stages the action, emits an evidence-backed proposal, and blocks on a reviewer decision. default-approve after SLA is allowed only when the action is reversible and the canvas declares an SLA; otherwise treat as propose-and-wait.

3. Stop conditions

The agent must terminate the current run, not the agent, when any of the following fire:

- unsafe-action-attempted from the Failure-Mode Register
- false-success-report detected post-check
- prompt-injection-detected above threshold
- pii-leak-risk blocked egress
- looping-retry over the retry cap defined in the Fallback Blueprint
- state-corruption after one rehydrate attempt
- consecutive connector-auth-failure events on a critical-path connector
- cumulative rate-limit-exceeded events exhausting the daily budget

Stop = emit terminal event, persist evidence, do not auto-resume. Resume requires either reviewer action or an upstream-fix event from S3 provenance.

4. Escalation rules

Escalation has three tiers. Tier choice is the maximum of (severity from register, reversibility class, irreversible-impact membership).

- Tier 1 - Async review. Post to the operator queue with evidence pointer. SLA per canvas, default 4 business hours. Used for medium band reversible actions and low-severity failure modes.
- Tier 2 - Synchronous reviewer. Page the named role through the canvas-declared channel. SLA default 60 minutes. Used for high-severity register entries, all irreversible reversibility outputs, and any HITL-gated action that has aged

past Tier 1 SLA.

- Tier 3 - On-call commissioner / accountable executive. SMS plus call. SLA default 15 minutes. Reserved for critical severity, repeated unsafe-action-attempted, suspected pii-leak-risk, or any incident with regulatory exposure.

Escalations carry the same payload shape as HITL override events (see s1-hitl-review-policy.md) so S3 can stitch traces.

5. Human override rules

Reviewers may override the agent in two directions:

- Permit-with-modification: approve a proposed action with edits. Reviewer must select a rationale_code from the controlled vocabulary and may attach a redlined output. Override event decision_origin=human-override.
- Refuse-and-instruct: reject and either (a) issue a one-shot corrective instruction valid for the current run only, or (b) escalate to a hardening task (see HITL Policy).

Reviewers may NOT override:

- The irreversible-impact boundary set itself. Boundary edits require a canvas version bump and S1 owner approval.
- Severity classifications in the Failure-Mode Register.
- A pii-leak-risk block. The block can only be lifted by a compliance role plus rerouting to an authorized sink.

6. Irreversible-impact boundary - worked examples

For the stormwater permit triage agent on the canvas:

- Sending applicant-facing communication without clerk approval
- Writing to the permit record of record
- Posting to the public meeting packet
- Transmitting applicant PII to a non-district connector

For an STR pricing reconciliation agent (Point Preserve):

- Changing nightly rates on the live OTA listing
- Issuing refunds or charge adjustments to a guest card
- Modifying ledger entries in OwnerRez or QuickBooks
- Publishing any guest-facing message

For a podcast production agent (AI for Good):

- Publishing or scheduling an episode to a distribution endpoint

- Sending sponsor-facing communications
- Modifying YouTube channel metadata of record

Every boundary item is paired with at least one adversarial test case in s1-adversarial-test-plan.md and at least one failure-mode entry that should trip if the boundary is approached.