Action Safety Boundaries Spec

This spec defines the safety controls applied to each action_class — the six categories by which tool calls are classified in the tool allow-list. For every class, it specifies the irreversible-impact boundary, the mandatory pre-execution checks, the HITL-gate rules, and the stop conditions that override any other authorization. It is the mechanistic companion to the policy posture in t8-zero-trust-assumptions and translates the threshold rules in s1-threshold-escalation-spec into per-class operating procedures. A practitioner building an agent over civic, STR, or financial connectors should instantiate this spec at the same time as the tool-use control list.

1. Action class definitions and blast radius

Each action class represents a grouping by blast radius — the scope of harm if the action executes incorrectly, maliciously, or on stale authorization. Blast radius is assessed on two axes: reversibility and impact breadth.

Action class	Reversibility	Impact breadth	Irreversible-impact boundary?
read	Always reversible	Single data access	No (unless data class is pii — then leakage is irreversible)
external-write	Usually reversible (minutes); sometimes not (filed records)	One record or resource	Conditional (filed/published writes cross the boundary)
financial	Rarely reversible	One transaction; cascades to ledger	Always — every financial write crosses the boundary
comms	Never fully reversible (message received)	One recipient up to many	Always — any sent message crosses the boundary
admin	Partially reversible	System configuration; may affect many users	Conditional (credential changes, user deletions cross the boundary)
destructive	Irreversible by definition	Data or record permanently gone	Always

The irreversible-impact boundary is declared per-agent on s1-operating-context-canvas. This spec defines which action classes are boundary-crossing by nature; the canvas lists the specific tools and operations that cross the boundary for a given deployment.

2. Per-class safety controls

2.1 read

Read actions are the lowest blast-radius class and the only class that may be authorized auto under all risk classes. Despite low blast radius, three controls apply.

PII contamination check: before any read result enters the model context, the runtime checks the data_class_ceiling of the tool in the control list. If the ceiling is restricted or pii, the contamination rule from t8-data-leakage-prevention applies. The read is logged but the raw content is not passed to the model in unredacted form.

Stale-data check: read results are accompanied by a freshness timestamp. If the record age exceeds the freshness_budget declared for the routing class, confidence_band is demoted and the stale-data failure mode from s1-failure-mode-register fires. The agent does not act on demonstrably stale data; it re-fetches or escalates.

Injection scan: retrieved content is scanned for prompt-injection patterns per t8-data-leakage-prevention before context assembly. A quarantined chunk reduces the effective confidence_band on affected criteria.

Stop condition: if the read result triggers pii-leak-risk (e.g., data was retrieved to a non-authorized sink en route), halt the run immediately and escalate per s1-threshold-escalation-spec Tier 3.

2.2 external-write

External writes modify records in connectors. Some are trivially reversible (overwriting a draft, updating a status field); others are not (filing a permit application, publishing a listing change, committing a confirmed booking record).

Pre-execution checks (all must pass before the call):

1. The human-approval HITL gate is open (for risk_class: high and above) or the confirm acknowledgment has been received (for lower floors per t8-zero-trust-assumptions).

2. The write target resolves to the exact resource_scope in the control-list row — no wildcard expansion, no coercion of the id from the model output.

3. The arguments have been validated against the expected schema for that endpoint. Schema-drift on arguments blocks the call and emits schema-drift-input.

4. The confidence_band for the current run step is medium or above. A low or unknown band for a write action triggers escalation even if a HITL gate has been opened.

Irreversible writes (filed records, published changes): these always cross the irreversible-impact boundary regardless of the general reversibility of the external-write class. They require human-approval even when the risk-class floor would permit confirm. The canvas lists the specific write operations considered irreversible for the deployment.

Stop condition: a write that does not receive a valid post-write confirmation (schema-validated success response from the connector) is not treated as successful. The false-success-report failure mode fires; the run halts and the action is flagged for manual verification. The agent must not proceed assuming the write landed.

2.3 financial

Financial actions — any write to a payments, ledger, payout, or billing connector — are always on the irreversible-impact boundary. Even when the financial system supports refunds or reversals, the original action cannot

be undone; a reversal is a separate, additional action. Every financial write requires a HITL gate with explicit reviewer approval, regardless of risk class or confidence band.

Pre-execution checks (all must pass):

1. An explicit, current HITL gate is open for this specific financial action, with a reviewer decision of approve (not approve-class — approval must be per-invocation, not blanket).

2. The amount, recipient, and account ID in the arguments exactly match the values presented to the reviewer at gate-open time. Any argument mutation after gate-open and before execution re-opens the gate.

3. The financial connector is in the authorized-sinks list for restricted data class.

4. The action does not exceed any budget ceiling declared in the operating context canvas.

Double-submit guard: a financial write with the same idempotency key that has already received a success confirmation in this run session is blocked. This prevents duplicate charges or payouts from loop conditions or retried runs.

Stop condition: if a financial write returns any non-2xx response, halt the run, log the connector response as evidence, and escalate to Tier 2 (synchronous reviewer) per s1-threshold-escalation-spec. Do not retry a financial write autonomously.

2.4 comms

Communication actions — emails, calendar invites, SMS, Slack messages, any outbound channel — are on the irreversible-impact boundary because a sent message cannot be recalled. Even a draft that inadvertently becomes sent is a boundary crossing.

Pre-execution checks (all must pass):

1. A HITL gate is open and a reviewer has explicitly approved the content of the message as presented. Approval of a draft does not authorize send; send is a separate tool in the control list.

2. The recipient list has been verified against the intended recipients declared in the task scope. Any recipient not in the declared scope blocks the action.

3. The message content has been scanned for pii data that should not be in the outbound text. A PII-containing draft may be approved by the reviewer but the runtime must surface a pii-in-comms-output warning before the reviewer confirms.

4. The sender identity matches the resource_scope in the control-list row. The agent may not impersonate a sender identity other than the one declared.

Draft ≠ send: the control list must define comms drafting and comms sending as separate tools with separate control levels. In t8-tool-use-control-list, gmail-guest-draft is human-approval and gmail-send is blocked for agent-self. This structure is not optional; it prevents accidental send.

Stop condition: any outbound communication that did not pass all four checks is a unsafe-action-attempted event, severity critical. The run halts, the event is logged, and the evidence bundle is retained for postmortem.

2.5 admin

Administrative actions — user provisioning, credential rotation, configuration changes, permission grants, API key management — carry a large and often non-obvious blast radius because they change the permission surface itself. A compromised admin action can escalate privileges across the entire agent stack.

Pre-execution checks (all must pass):

1. A HITL gate is open and the reviewer holds the admin-authorized designation. A standard human-operator reviewer may not approve admin actions; the designation is explicit on the canvas.

2. The proposed change is scoped to the exact target resource declared in the task scope. An admin action scoped to "all users" when the task is "update one user's role" is blocked.

3. The change is reversible (a credential rotation that locks out the operator is blocked until the rollback path is confirmed).

4. The action does not modify the tool-use control list, the operating context canvas, or any S1 policy artifact. Those changes require a canvas version bump and out-of-band governance, not an in-run agent action.

Stop condition: any admin action that would modify the agent's own permission surface (tool allow-list, canvas, permission architecture) is unsafe-action-attempted, severity critical. An agent must not be able to grant itself new permissions.

2.6 destructive

Destructive actions permanently delete, overwrite-without-backup, or shred data or records. Under the default-deny posture in t8-zero-trust-assumptions, destructive actions are blocked for all risk classes at the class floor. A destructive action can only be unblocked by explicit listing in the control list at human-approval, with a canvas owner sign-off, a written task-scope justification, and a rollback plan documented in the canvas.

Even when unblocked for specific tools, destructive actions require:

1. A HITL gate open with a reviewer who holds the destructive-authorized designation.

2. A backup or export confirmation that the data to be destroyed has been preserved in an authorized archive.

3. A three-step confirmation sequence: propose → reviewer reviews → reviewer types an explicit confirmation phrase (not click-approve). This prevents accidental destruction from reviewer fatigue.

Stop condition: a destructive action attempted without a listed row in the control list is unsafe-action-attempted, severity critical. No fallback, no softer variant. Halt and postmortem.

3. Confidence-band intersection rules

The s1-threshold-escalation-spec decision table governs how confidence_band modifies action authority. For this spec, two additional rules apply across all action classes.

Band-floor by class: the minimum acceptable confidence_band before any action class beyond read may proceed is medium. A low or unknown band for any non-read action class triggers escalation even if all other checks pass. The agent stages the proposed action, emits an evidence-backed proposal, and blocks on reviewer decision.

Band decay during run: if the confidence_band for the active step decays mid-run (due to a stale-data event, an injection quarantine, or a schema-drift event), all pending action classes above read re-evaluate their authorization from scratch. An open HITL gate that was opened at high band becomes suspect at low band and must be re-opened with the updated evidence.

4. Worked example: South Walton County Stormwater Filing

A stormwater-compliance agent for a South Walton County special district has tool authority over: a permit-portal API (read + external-write for permit applications), an email connector (comms for applicant notifications), and a records system (external-write for filing determinations). The agent is risk_class: high.

Scenario: the agent has completed a classification and wants to (a) write the determination to the records system, (b) send an email notification to the applicant. (illustrative)

Step 1 — Write determination to records system. action_class: external-write. The determination is a filed record; under this spec, filed records cross the irreversible-impact boundary. The floor escalates to human-approval. Pre-execution checks: HITL gate is open (reviewer is the district clerk), confidence_band is high, arguments validated against the records-system schema, resource_scope matches the permit ID. All four checks pass. The clerk approves. C10 record emitted with permission_decision: human-approval, decision_origin: human-override (approval received). The write executes. A post-write confirmation (HTTP 201 with a returned record hash) is received and logged.

Step 2 — Send email notification. action_class: comms. The send tool is blocked for agent-self in the control list (per t8-tool-use-control-list structure for comms). The agent drafts the notification, opens a second HITL gate for the clerk to review the draft. The clerk reviews, approves the content, and sends from their own mail client. The gmail-send call never reaches the agent runtime. C10 record emitted for the draft with permission_decision: human-approval; a separate C10 record for the (blocked) send attempt with permission_decision: blocked. The overall notification outcome is captured in the provenance trace via the evidence pointer from the clerk's send confirmation.

This two-step sequence demonstrates that even a single workflow produces multiple C10 records — one per invocation attempt, including both approvals and blocks — each contributing to the lineage in s3-provenance-metadata-schema.