Decision Traceability

Purpose

Decision traceability is the practice of connecting every consequential claim in a model output back to specific source evidence and a specific decision_origin. It is what separates a defensible output from a confident-sounding guess. In GROW, an output without a trace is not an output — it is a draft.

This spec defines (a) the four-way classification of decisions, (b) the rule for flagging unsupported conclusions, and (c) a worked backward trace of a real-shape agent output.

Classification of Decisions

Every consequential statement in an output is one of four kinds. The classification maps cleanly onto the C2 decision_origin enum from s1-hitl-review-policy:

Decision kind	What it is	C2 decision_origin
Retrieved data	The statement is a faithful restatement of a record from a source.	agent (when accepted as-is)
Model inference	The statement is a generalization, synthesis, or interpretation produced by the model.	agent (with reduced source confidence)
User input	The statement is provided by a human in the loop and propagated.	human-override
Tool output	The statement is the result of a deterministic tool call (calculator, query, API).	agent (high confidence if tool is authoritative)

When HITL pathways activate, the mapping shifts:

- A reviewer accepts and signs: human-override event is recorded; the underlying classification of the statement is preserved.
- A reviewer rewrites the statement: human-override replaces the classification with user input.
- The system declines to answer and falls back to a templated response: fallback.
- The system routes to a senior reviewer: escalation.

These events live in decision_trace[] per the provenance schema, with the C2 shape: {event_id, timestamp, agent_id, decision_origin, evidence_pointer, rationale}.

Flagging Unsupported Conclusions

A statement is unsupported if any of the following holds:

1. No evidence_pointer resolves to a source on the inventory whose content materially supports the statement.

2. The supporting source has confidence_band: low or unknown and the statement is asserted without qualification.

3. The statement is a model inference that bridges across sources whose merge is not authorized by a recorded transformation in the lineage map.

4. The supporting source is stale beyond its update_frequency window and has not been re-validated.

When any of these hold, the system MUST:

- Mark the statement inline with a visible qualifier ([unsupported], [low confidence], or [stale source]).
- Reduce the record's confidence_band to the strictest applicable level.
- Emit a decision_trace event with decision_origin: fallback if the system declines to assert the claim, or escalation if it routes to a human.

Silent assertion of an unsupported claim is a reliability defect and a s2-scoring-system failure.

Worked Backward Trace

The synthetic output below is the closing paragraph of an automated STR Revenue & Occupancy Snapshot drafted Monday morning for Point Preserve. The trace decomposes each claim.

> "Point Preserve booked 27 confirmed reservations in May 2026 against a target of 24, yielding a gross of $94,210 before platform fees. May occupancy ran 86%, ahead of the 30A market average we estimate at 74%. We recommend holding base rates for June and revisiting the cleaning-fee floor after Tropical Storm Brett."

Backward trace:

Claim	Kind	Source(s)	Evidence pointer	confidence_band	Notes
"27 confirmed reservations"	Retrieved data	ownerrez-bookings	s3://prov/.../sources/ownerrez/bookings_2026-05.json	high	Direct count of status=confirmed rows.
"gross of $94,210"	Tool output	ownerrez-bookings + reconciliation tool	s3://.../tools/reconcile_result_0xC11.json	high	Reconciled with airbnb-csv-dump; deltas < $50.
"May occupancy ran 86%"	Tool output	ownerrez-bookings	s3://.../tools/occupancy_calc.json	high	Deterministic ratio over calendar nights.

Claim	Kind	Source(s)	Evidence pointer	confidence_band	Notes
"30A market average we estimate at 74%"	Model inference	None on inventory	—	unknown — flagged	No authoritative comp source listed. Inline [unsupported]; trace event decision_origin: escalation.
"Hold base rates for June"	Model inference	ownerrez-bookings trend + prior memos	s3://.../retrieval/prior_memos_top3.json	medium	Acceptable because labeled as recommendation, not as fact.
"Revisit the cleaning-fee floor after Tropical Storm Brett"	User input	Doug's morning note	s3://.../inputs/morning_note_2026-05-28.md	high	decision_origin: human-override event recorded.

The trace produces five outcomes:

1. The 30A market average claim is rewritten or removed before publication; the agent does not get to assert it.

2. The record's top-level confidence_band is set to medium (downgraded from high because of the inference bridge).

3. decision_trace[] contains at least three events: one agent, one escalation, one human-override.

4. evidence_pointer is set to the directory containing all per-claim pointers.

5. The output as shipped carries an inline footnote or sidebar listing the per-claim sources by source_id, satisfying the publication requirement in s3-governance-retention-policy.

Operating Rules

1. Every consequential statement is one of the four kinds. "Mixed" is not a kind — decompose.

2. Inline qualifiers ([unsupported], [low confidence], [stale source]) are non-optional when the conditions hold; they are not a style choice.

3. The trace lives with the output. A trace stored separately from the artifact it justifies is not a trace; it is filing.

4. Backward traceability is the GROW-default. Forward propagation (this output feeds the next system) is handled in s3-lineage-map-spec.