RED BEAR: ENTERPRISE DECISION INTELLIGENCE · DATA SHEET · D · DELEGATION, POLICY AND AGENT AUTHORITY

Set the guardrail once, then widen the dial as it is earned

A rule about risk and a judgement about a person are not the same setting.

KEY POINTS

- Three separate objects: the Trust Dial belongs to the decider, the guardrail to the decision type, and role scope says which types the decider may touch at all
- One dial per person or agent, 0 to 10 in five bands; the band is the review rate; a move is one point at a time, by a person, with a reason, after the projection
- Nine fixed kinds of guardrail; policy as versioned code; enforcement is Block, Approve or Notify; a widening needs a second person
- An agent's ceiling is 9; every agent has an owner, one purpose, one dial and a kill switch

What it is

Every large organisation has a delegation matrix, and almost none is current. It was drafted after someone was surprised by a loss, approved by a committee, and filed. It records what a job title may sign for, and nothing about whether the person holding it has ever made a decision of that kind well, badly, or at all.

It goes stale because it holds two different kinds of statement in a single line. "Decisions above this value need a named approver" is a rule about the world. "This team can be left to price their own renewals" is a judgement about an actor. The first should change when risk changes; the second when evidence changes. Written as one row in one table, neither can move without reopening the other, so neither moves. The control plane keeps the two apart. **Guardrails on the decision type set the hard boundary. The dial on the decider sets how much room inside that boundary they get.**

The pain point

Leaders are drowning in approvals: senior people see every decision and give none proper attention. Delegation is a conversation, not a control: authority granted verbally, remembered differently, never adjusted. Trust is assumed, not earned: discretion set by seniority rather than by the record, so good deciders are held back and poor ones are not caught. Changing a policy is a leap in the dark: the owner cannot see what a threshold change will do until it has done it, so policy is set conservatively and never revisited.

And AI adoption stalls on authority. Nobody signs off an agent acting without a human, because there is no evidence it decides well and no way to stop it, so agents are confined to suggestions: most of the cost, little of the benefit.

How it works

The dial. Every person who decides on the organisation's behalf has one Trust Dial, and so does every AI agent: 0 to 10, five bands. *No Discretion* (0): everything is referred. *Recommend* (1–3): they propose and a human decides. *Supervised* (4–6): they act within narrow limits and every decision is reviewed afterwards. *Delegated* (7–9): wide limits, decisions sampled. *Autonomous* (10): reviewed by exception, still recorded, still inside guardrails. The band is the review rate. There is one dial per decider, not one per decision type: a new joiner and a twenty-year veteran in the same role have the same scope and different dials, and that difference is the point. A dial moves one point at a time: for a person as an act of coaching between director and manager, for an agent as a decision in its own right, owned by a person and recorded like any other. It cannot be committed without the projection having been displayed: the routing function re-run over the class's history at both settings, showing the before-and-after band mix and, in words, how many of these decisions would still need a person. The earned trust ledger scores each actor's record and recommends a setting with its evidence; it has no path to apply it.

The guardrails. Guardrails are set per decision type in nine fixed kinds: authority, policy, financial limits, risk appetite, compliance, segregation of duties, exceptions, human intervention and auditability. Policy is code, not prose: versioned, owned, tested, applied to people, processes and agents alike; every evaluation records the version it used. Enforcement has three strengths: *Block* prevents release, *Approve* requires a named person, *Notify* permits release and records. Conflicts are caught at authoring, exceptions carry a sunset date or are refused, and a widening requires activation by a second person with authority for that type. The scenario sandbox replays a named period under a proposed policy or dial through the same routing function as live routing and reports the difference; it applies nothing. Pre-authorised envelopes hold standing authority, so opportunities are not lost while an approval is assembled for a decision everyone would have agreed to.

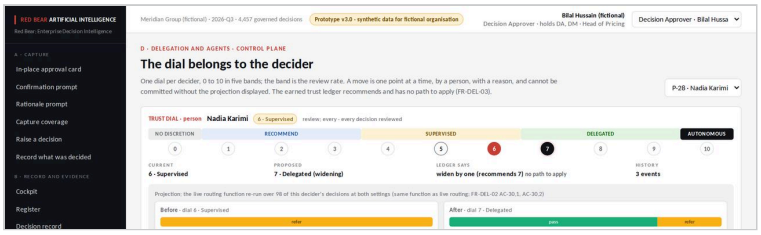
Agents, one class at a time. The agent gateway is an endpoint an agent calls before acting; it returns allow, hold or route, and the call itself is the capture event whatever the answer. An agent that acts and reports afterwards is being observed, not governed. Every agent has an identity, an accountable human owner, a defined scope and an off switch one person can reach. One agent, one purpose, one dial: a dial earned on refunds says nothing about contract variations, and a materially changed model starts again near the bottom of Supervised. Shadow mode records an agent's decisions without executing them and compares them with human decisions on the same class, so autonomy extends class by class, promoted by a person. Why measure rather than assume: Vaccaro, Almaatouq and Malone's 2024 meta-analysis in *Nature Human Behaviour* found human–AI combinations on average performed worse than the better of the two alone on decision tasks. Adding a

person to an agent's decision is not reliably an improvement, and neither is removing one. **An agent's ceiling is 9.** No agent holds Autonomous on a consequential decision type, and accountability never transfers to one.

What you get

Capability	What it does	Requirements
Decision rights, role scope and the rights matrix	Who may decide which class to what limit, current from HR and the directory; the five roles against person, committee and agent; team tiers that seed a dial and never hold one	FR-DEL-01, FR-DEL-06, FR-DEL-07, FR-DEL-08
Trust dials and the earned trust ledger	One dial per decider, projection before commit, reason recorded; trust scored from the record and recommended, never applied	FR-DEL-02, FR-DEL-03
Envelopes and interception	Standing authority as versioned policy; a decision held before it executes and released when it clears; an override is a separate, attributed event with a reason	FR-DEL-04, FR-DEL-05
Policy builder, lifecycle, sandbox and optimisation	Nine kinds, three strengths, versions, conflicts caught at authoring, exceptions with sunsets, second-person activation of a widening; replay a period under a proposed setting; test thresholds against risk scenarios	FR-DEL-14 to FR-DEL-17, FR-DEL-19
Override telemetry	Approvals over a breach, by policy first, presented as a question about the policy	FR-DEL-18
Agent gateway, registry, kill switch, shadow mode and lifecycle	Allow, hold or route before the act; owner, scope, model version, off switch; promotion class by class; the product's own AI governed the same way	FR-DEL-09 to FR-DEL-13

Screens



S-20 Control plane: The control plane: the dial for a decider, the before-and-after band mix, the ledger's recommendation, and the commit panel with its reason. Prototype render on synthetic data for fictional entities.

RETURN ON INVESTMENT

Delegation becomes reversible: a setting with a projection, not a restructure. **Risk appetite stops standing in for trust:** the boundary and the room inside it move independently. **Agents become adoptable:** authority extended on evidence, class by class, and withdrawable. Two of the eight benefits rest on this cluster alone: **confident AI adoption**, measured by the classes where agents decide within guardrails and their override and reversal rates, and **resilience as conditions change**, measured by the share of policies replayed before commit and the time from a condition changing to the policy changing. No measure carries a target.

Platform engines

Decision Intelligence for routing and the authority table (RBP FR-DI-02, RBP FR-DI-04, RBP FR-DI-05, RBP FR-DI-07); Presence for deciders, the agent ceiling and dial moves (RBP FR-PR-01, RBP FR-PR-03, RBP FR-PR-04); Decision Learning for calibration (RBP FR-DL-07); the AI gateway (RBP FR-AI-03, RBP FR-AI-05). Specified in the platform FSD.