

Responsible AI Platform: Selection, Architecture, and Tradeoffs

- 2026 shift: from principles documents to enforced platform controls
- Audit evidence now generated as a byproduct of building, not retrofitted
- EU AI Act general application and Article 50 transparency apply from 2 August 2026
- GPAI obligations apply since August 2025; Annex III high-risk deferred to 2 December 2027
- Three decisions: what to buy, how to wire it, which tradeoffs to record
- No platform is uniformly responsible — responsibility is properties you fund and prove



Why Platform Engineering Now: Regulation, Market and Failure Modes

SAFER AI
STRONGER SYSTEMS
BRIGHTER OUTCOMES



Regulatory baseline:
EU AI Act phased application,
NIST AI RMF 1.0



Transparency duties
enforce from 2 August 2026;
high-risk deferred to 2 December 2027



22+ governance vendors across policy, observability, guardrails, AI security



Failure modes:
unenforced model cards, audits without evidence pipelines



Shared vocabulary:
risk tiering, registry, guardrails, evals, provenance, approval gates

From Principles to Enforceable Capabilities

- Capability map: tiering, lineage, inventory, evals, guardrails, oversight, monitoring, evidence
- A real control blocks a deployment or emits audit-ready evidence
- Map each capability to a principle and testable exit criteria
- Maturity ladder: manual review, tool-assisted, policy-as-code, continuously verified
- Gap analysis: owner, evidence, enforcement point, failure behaviour, review cadence



Build, Buy, or Compose: Selection Criteria and Vendor Evaluation

ALIGN RISK, CAPABILITY, AND VALUE FOR AI GOVERNANCE



Market splits into six buyable segments; most enterprises need two to four



Score on enforcement vs advisory, exportable evidence, neutrality, TCO



Highest-variance test: shadow AI, embedded SaaS AI, agents, MCP tools



Demand one control mapped to two frameworks with shared evidence



Lock-in risk is real: many 2026 vendors will be acquired or pivot by 2028



Test the unhappy path in your own environment before the demo

Evaluate Across Six Vendor Segments

	 Segment 1	 Segment 2	 Segment 3	 Segment 4	 Segment 5	 Segment 6
Enforcement vs Advisory						
Exportable Evidence						
Neutrality				 shadow-AI discovery		
TCO						



MAKE A MORE
INFORMED CHOICE

Reference Architecture: Control Plane, Data Plane, Enforcement Points

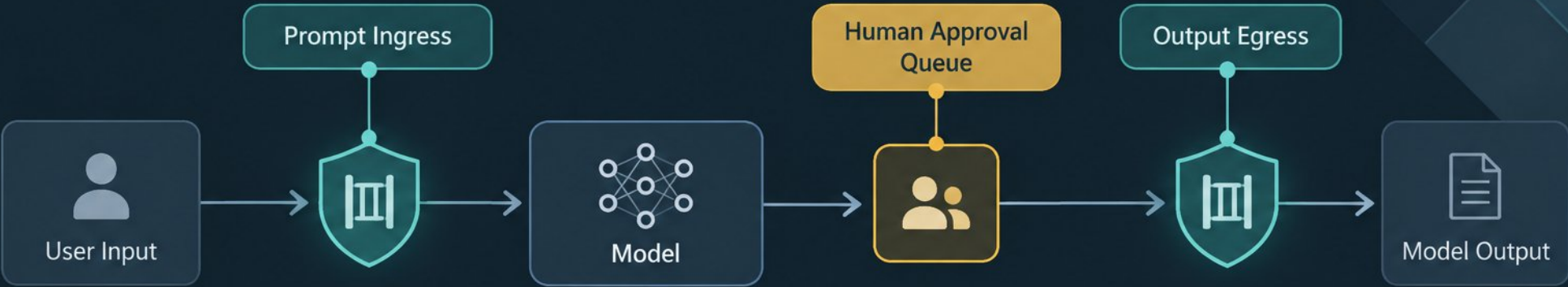
Governed, secure, and observable AI systems from policy to runtime



- 1 Six logical layers, from intake and tiering to the audit evidence store
- 2 Control plane authors, versions, and shadow-tests policy off the request path
- 3 Data plane enforces inline: prompt ingress, retrieval, tools, output egress, agent actions, MCP
- 4 Policy-as-code lifecycle: author, dry-run, shadow, promote, version, roll back
- 5 Gateway enforces virtual keys, budgets, SSO/RBAC, PII redaction, per-decision records
- 6 Integrate with catalog, CI/CD, registries, secrets, SIEM, ticketing, and IdP

Controls in Depth: Guardrails, Evaluation and Human Oversight

MULTIPLE LAYERS. HIGHER ASSURANCE. SAFER OUTCOMES.



Evaluation stack:
benchmarks, domain golden sets, adaptive red teaming, release gates, regression monitoring

Prompt injection is unsolved:
layer deterministic filters, classifiers, least-privilege tools, human approval

Authoritative enforcement on output side;
pre-inference filtering is secondary

Metrics that matter:
FP/FN per risk tier, adaptive attack success rate, not aggregate accuracy

Every control emits:
decision, policy version, actor identity, model destination, timestamp, signature

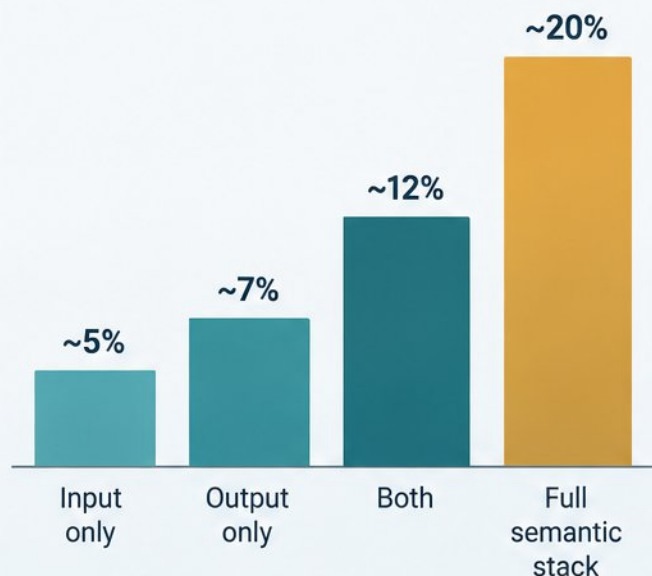
Cost and Latency of Control: Findings You Can Budget Against

Practical tradeoffs. Measurable impact. Safer systems.

MEASURE
DESIGN
DEPLOY
PROTECT

Guardrail Overhead on a ~250 ms Base

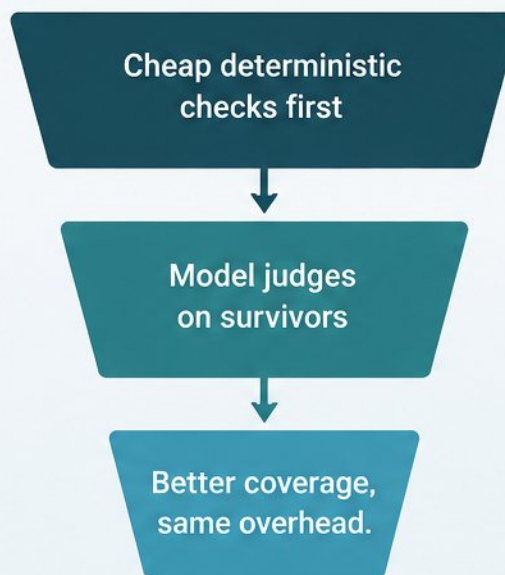
Incremental latency by guardrail configuration



Guardrail overhead on ~250 ms base: input ~5%, output ~7%, both ~12%, full semantic stack ~20%.

Cascade Funds Recall

Cheap checks first, model judges on survivors – better coverage, same overhead.



Guardrail overhead on ~250 ms base: input ~5%, output ~7%, both ~12%, full semantic stack ~20%.



Cascade funds recall: cheap deterministic checks first, model judges on survivors – better coverage, same overhead.



Placement is the biggest lever: blocking output checks serialize two model passes.



Budget at the UX boundary, not the model boundary; anything over ~50 ms inline gets disabled.



Purpose-built decision models beat fine-tuned SLMs: no labeling bill, no standing serving cost.



Run the numbers on your own traffic – a guardrail you can afford to leave on is the only kind that protects you.

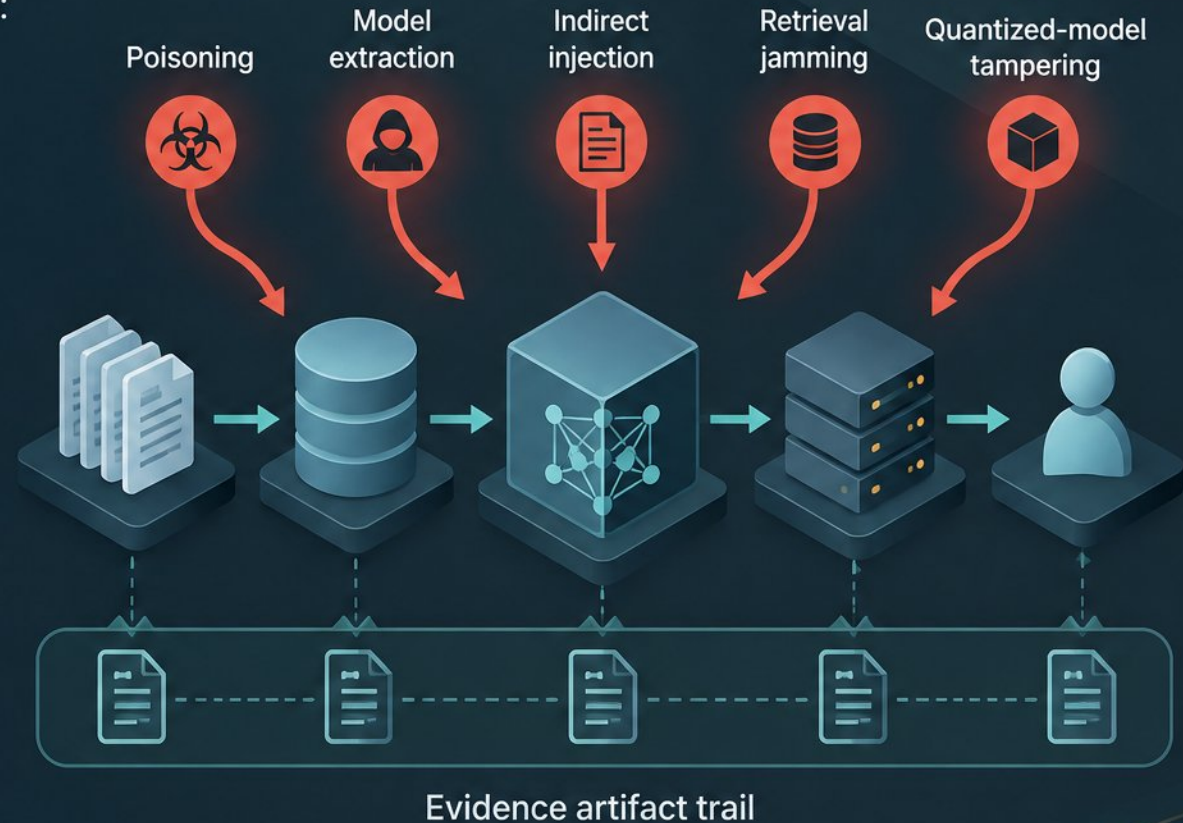
COST

LATENCY

SAFER AI
THROUGH
PRACTICAL
TRADEOFFS

Security, Privacy and the AI Supply Chain as Platform Primitives

- Treat models as software supply chains: provenance, version pinning, signing, attestation
- AI-BOMs: CISA/G7 minimum elements, CycloneDX ML-BOM, SPDX AI profiles
- Mark undisclosed closed-model fields explicitly — never leave silent blanks
- Lineage, consent, licensing, retention and deletion propagate across datasets
- PII minimization, residency rules, and cross-border routing constraints
- Updated threat model: poisoning, model extraction, indirect injection, MCP risks
- Incident response: kill switches, rollback authority, vendor audit rights



Tradeoffs That Decide the Design



- Depth of control vs. latency and cost: which checks stay switched on?

Depth of control



Latency and cost



- Precision vs. recall: who bears the cost of each error type?

Precision



Recall



- Centralization vs. autonomy: one policy surface or federated controls?

Centralization



Autonomy



- Model neutrality vs. deep integration; open weights vs. managed APIs

Model neutrality



Deep integration



- Build vs. buy under time pressure — interim accepts need expiry dates

Build



Buy



- Decision records: options, chosen tradeoff, accepted risk, review trigger

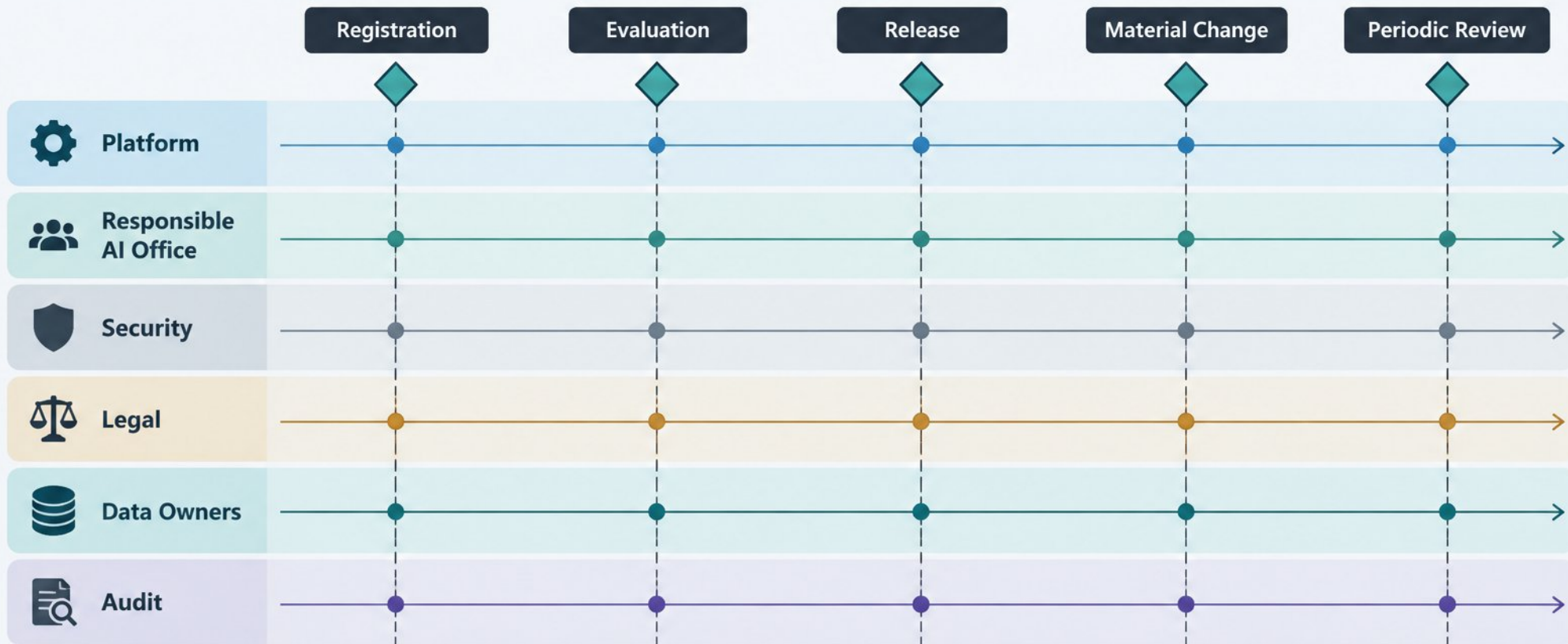
Options



Review trigger

Operating Model: Roles, Decision Rights and Evidence Ownership

CLEAR ROLES. CONTROLLED PROGRESSION. AUDITABLE OUTCOMES.



- AI accountability is plural: business, governance, risk and legal owners each accountable
- Every activity has exactly one accountable role — shared accountability is the top failure mode
- Gates at registration, evaluation, release, material change and periodic review
- Three lines of defense: platform telemetry becomes independent audit evidence
- Metrics: gate cycle time, control coverage, override rate, evidence currency

KEY TAKEAWAY

Defined roles, decision gates and evidence ownership enable trustworthy and accountable AI at scale.

Adoption Roadmap and Common Pitfalls

GOVERN
ENABLE
SUSTAIN



Stage it: inventory and tiering, then evaluation gates, then runtime enforcement, then continuous verification



Sequence by pressure: start where enforcement produces evidence cheaply and regulation bites hardest



Expect 40–60% more AI systems than you thought, including shadow AI and embedded SaaS features



Anticipate pitfalls: tool sprawl, unowned controls, untestable policies, metrics counting activity, not risk



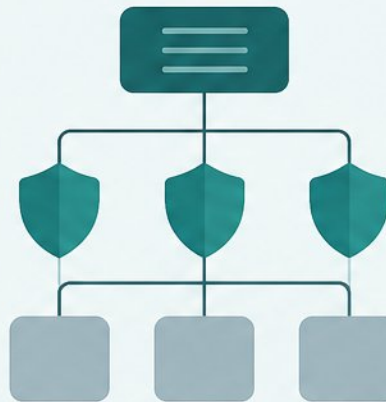
Sustain momentum: board reporting, risk burn-down, and treating retirement as a governance win

FROM
AWARENESS
TO VALUE

Workshop Part 1: Scoring Candidate Configurations Against Your Risk Tiers

- 1 Exercise 1: score two candidate stacks against your own risk tiers
- 2 Registry-led stack with separate guardrails versus gateway-centric with fused telemetry
- 3 Use the capability map as your checklist, not vendor feature lists
- 4 Discovery test: how much shadow AI, agents and embedded SaaS does it surface?
- 5 Test the unhappy path: silent model updates, unapproved agent actions, evidence export
- 6 Deliver a defensible recommendation: rationale, unclosed gaps, compensating controls
- 7 Peer critique: defend scoring with the same checklist and failure scenarios

Registry-led stack with separate guardrails



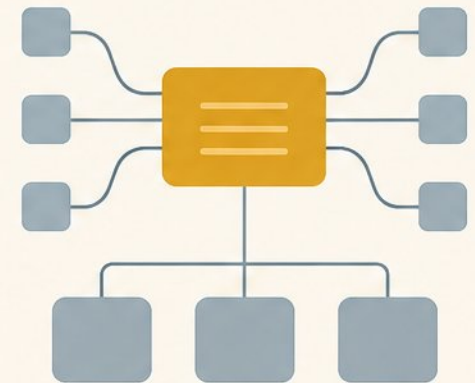
Discovery test:
how much shadow AI,
agents and embedded
SaaS does it surface?



Evidence
artifact trail



Gateway-centric with fused telemetry

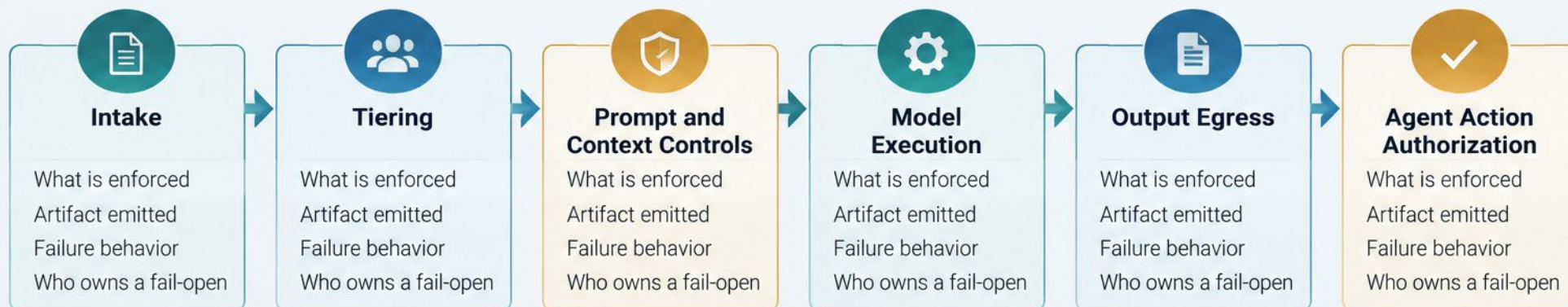


Discovery test:
how much shadow AI,
agents and embedded
SaaS does it surface?



Workshop Part 2: Enforcement-Point Architecture for a High-Risk Use Case

GOVERN > BUILD > DEPLOY > MONITOR
SAFER AI. STRONGER OUTCOMES.



Key Exercises

- Exercise 2: sketch enforcement points from intake and tiering to output egress and agent action authorization**
 - For each point state: what is enforced, artifact emitted, failure behavior, and who owns a fail-open
 - Name the weakest link explicitly — often the boundary you never instrumented
- Exercise 3: three accepted tradeoffs, evidence to monitor, and the trigger forcing reconsideration**

Compare Use Cases



High-Risk Regulated Case

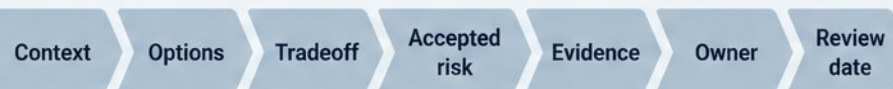


Low-Risk Internal Productivity Case

Contrast the high-risk regulated case against a low-risk internal productivity case.



Decision Record



Committing Back at Work: Owners, Evidence and Review Triggers



One decision record per team: configuration, tradeoffs, owners, review dates



Three commitments each: one enforceable control, one automated evidence, one recorded tradeoff



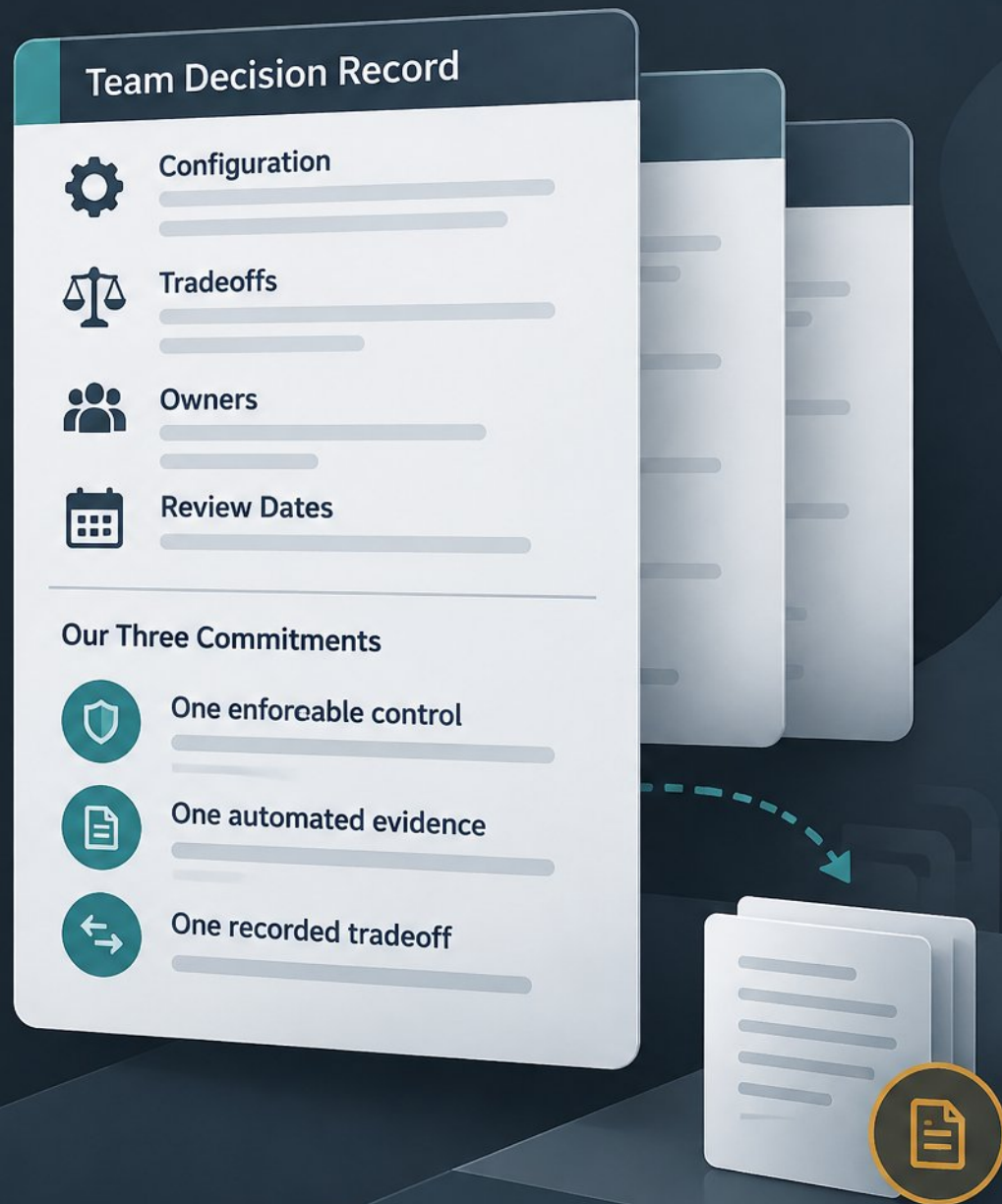
First measurable outcomes: tiering coverage, gate cycle time, evidence exportability



Set review triggers now: model version, vendor, jurisdiction or use-case change



A platform is responsible by properties you funded, enforced and can prove



PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/c2750a98/public