

R-BED™ | Risk-Based Evaluability Design

A public, consequence-aware methodology for evidence-based evaluation and authorization of deployed AI systems

Version 2.1 | 29 October 2026 | Public community resource

Prepared by Vishal Kumar Srivastava, PhD

Purpose	Specify how to define system claims, characterize consequences, find failure and misuse pathways, test whether deployment controls contain them, and make bounded, authorized deployment decisions that are reopened when conditions change.
Applicability	Traditional machine learning, generative AI, retrieval-augmented systems, and AI agents, with depth proportional to consequence, exposure, uncertainty and risk tolerance.
Status	Proposed practitioner methodology. Not a standard, certification, validated scoring instrument, or NIST-endorsed framework. Not empirically validated (Section 13).

Executive summary

AI risk does not stop at deployment. Neither should evaluation. R-BED includes continuous system evaluation (CSE): evidence is dated and pinned to a configuration, safeguards are re-challenged, monitoring effort is allocated by consequence, and authorization lapses when its evidence does.

R-BED organizes system-level evaluation evidence across three connected vertices: (V1) Technical Failure—where a configured AI system can fail; (V2) Deployment Controls—whether that failure, or intended or adversarial behavior, can reach a person, business process, or external system; and (V3) Consequence and Impact—what material consequences follow and for whom. Thirteen diagnostic dimensions (6+4+3) turn these questions into claims, evidence requests, falsification tests, and decisions.

Version 2.1 has two parts. The first specifies what to evaluate and decide: the architecture, the conditional pathway identity, the evidence statuses, the five dispositions, the eight authorization rules, five cross-cutting objectives, a Consequence Register, intended-behavior and misuse pathways, extended quantitative tools, a seven-stage lifecycle with eight roles and independence levels, the System Evaluation Specification, and hypotheses H1–H7. The second is *continuous system evaluation* (CSE), which asks what happens *after* an authorization is issued: how long evidence may be relied on, how safeguards are shown to keep working, how scarce monitoring effort is directed, and what is learned from failures. The CSE Specification (https://rbed.ai/RBED_CSE_Specification_v2.1.pdf) gives the detail; Section 9.5 summarizes it. Each element is labeled **[E]** specified core (reference code exists and passes its own tests), **[P]** proposed extension, or **[H]** research hypothesis. Nothing labeled [E], [P] or [H] has been empirically validated.

The central practice is to evaluate a defined claim about a configured system, not a model score in isolation. A credible result must identify the deployment context, test distribution or scenarios, system versions, limitations, control tests, uncertainty, affected stakeholders, and accountable decision. R-BED does not collapse the thirteen dimensions into a single numerical rating, and it does not aggregate across systems into a portfolio score.

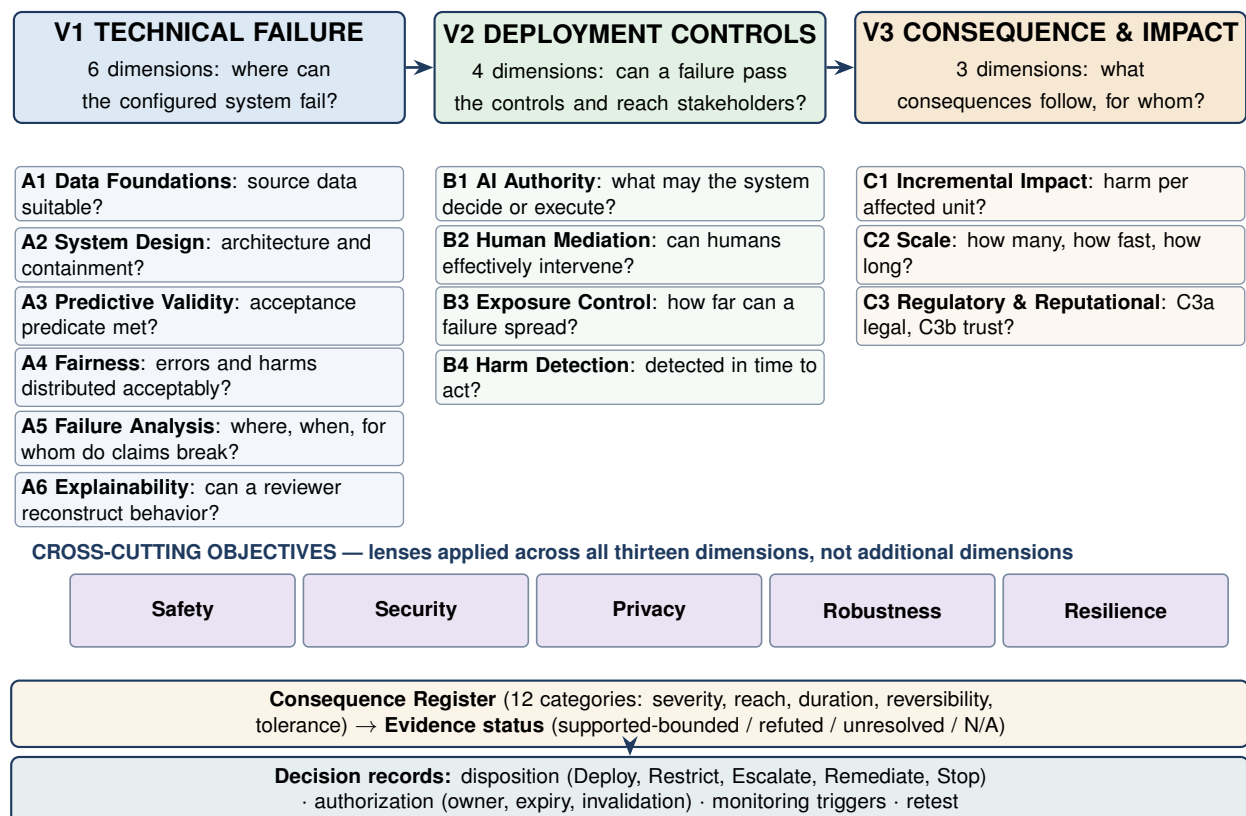


Figure 1. R-BED v2.1 architecture: thirteen diagnostic dimensions in three connected vertices (6 + 4 + 3), five cross-cutting objectives applied across them, and the Consequence Register and decision records that connect evidence to action. Arrows show a simplified failure-mediated pathway; not every harmful outcome requires a technical failure. Dimensions are locations for claims and evidence, not scores, and evidence from one vertex cannot substitute for another.

Methodological contribution

R-BED connects four elements in one traceable record: technical evaluation evidence, evidence that deployment controls work, the real-world consequences at stake, and an accountable deployment decision. It does this for the configured system rather than the model in isolation. Five features are combined; each has precedents elsewhere (Section 15), and the contribution claimed is their combination, not the novelty of any one.

- **Consequence-aware evaluation:** technical failure pathways, the controls that may stop them, and consequences for people, organizations and society are kept as separate questions, because model-level metrics alone do not show what a deployed system can cause (Sections 3, 5).
- **Independent control testing:** a documented safeguard is not credited until evidence shows it prevents or contains the specified pathway under challenge, at a stated level of evidence independence (Section 7).
- **Explicit uncertainty:** claims carry confidence bounds and acceptance thresholds fixed before testing; unresolved claims stay visible; zero observed failures does not establish zero risk, and not every evaluation yields a reliable quantitative estimate (Sections 8, 7).
- **Evidence-based authorization:** evidence informs five dispositions (Deploy, Restrict, Escalate, Remediate, Stop) under eight rules, with a named decision authority, an expiry and invalidation conditions. The record supports organizational judgment and does not replace it (Section 11).
- **System-level scope:** the unit assessed includes retrieval, tools, permissions, human interventions, monitoring and operational outcomes. It is applied to generative and agentic systems and is not limited to them (Section 2).

R-BED complements, and does not replace, NIST AI RMF 1.0, ISO/IEC 42001 and technical evaluation

resources. Its effectiveness has not yet been empirically validated (Section 13).

Five principles

- **Risk governs depth:** prioritize consequences and unresolved uncertainty, not incident frequency alone.
- **Failure does not equal harm, and harm does not require failure:** test the independent boundaries that stop an error from becoming an action, and separately examine harm from intended behavior and misuse.
- **Evidence must connect end to end:** component, trajectory, and outcome evidence jointly support a system claim.
- **A remediation must be tested:** detection, proposed fixes, and verified containment are distinct evidentiary states, and a control is credited only with evidence of its effectiveness.
- **Evidence, authorization and decision are separate records:** a test result is not an evidence assessment, and an evidence assessment is not a deployment decision. Unresolved is not safe.

Three methodological commitments

Version 2.1 organizes the methodology around three methodological commitments that depend on one another (Section 9.5; CSE Specification Section 14). They are the method's differentiators; they are not architectural layers, and they are distinct from the three vertices and the five cross-cutting objectives (Section 2.1).

- **Consequence-aware evaluation:** decide what matters, by potential consequence and not by frequency alone.
- **Evidence-based control assurance:** decide whether the safeguards between a failure and a consequence work under realistic failure, at a stated level of assurance (L0–L3).
- **Continuous evaluation and authorization:** decide how long a conclusion may be relied on; evidence has an owner, a configuration, an expiry and invalidation triggers, and an authorization cannot outlive the evidence it depends on.

Continuous assurance is not continuous testing: not every check is rerun continuously, and the cadence of each activity follows consequence, exposure, uncertainty and the rate of change.

Contents and maturity at a glance

Area	Maturity	Summary
Architecture (6+4+3, 13 dimensions), objectives, Consequence Register	[E]	Three vertices, thirteen diagnostic dimensions, five cross-cutting objectives, twelve consequence categories.
Conditional pathway identity, evidence statuses, five dispositions, eight rules, SES minimum fields	[E]	"Continue" is the renewal of a Deploy authorization, not a sixth disposition.
Evidence Validity Register	[P]	Dated, configuration-pinned evidence with proportional, dependency-based invalidation (Section 9.5).
Control effectiveness	[P]	Assurance ladder L0–L3 and eight control types mapped to the four Deployment Control dimensions.
Agentic evaluation	[P]	Authority envelope, reversibility classes and eight trajectory metrics.

Area	Maturity	Summary
Oversight effectiveness	[P]	Seeded-error and automation-bias measures; extends dimension “Human Mediation”.
Monitoring allocation	[P], [H]	Monitoring priority tiers T1–T4 by six tier rules (TR1–TR6); lexicographic allocation; audit floor; hypothesis H7.
Failure-to-learning loop and portfolio assurance report	[P]	Discover, Reproduce, Contain, Fix, Verify, Learn; seven lists and ranked decision asks, no score.
Mathematics	[E], [P]	The conditional identity and its ten extensions are [E]; policy-conditioned loss, avoided-loss contrast and evidence-basis labels are [P].
SES	[E], [P]	Eight minimum fields and optional blocks [E]; five continuous-evaluation blocks [P].
Crosswalk and Validation Plan	—	NIST AI RMF 1.0 crosswalk: 51 subcategories mapped (44 Partial, 7 Relevant). Validation Plan: studies S1–S7 for hypotheses H1–H7.

1 Scope, audience, and limits

This specification is a non-prescriptive reference for AI developers, validators, risk managers, assurance teams, researchers, and deployment decision owners. A framework user selects the applicable dimensions and documents omissions. High-consequence deployments typically require greater breadth of scenarios, more independent challenge, tighter restrictions, and stronger ongoing monitoring.

R-BED is intended to complement, not replace, organizational risk governance, sector regulation, external assurance, human impact assessment, cybersecurity, privacy management, or the NIST AI RMF. Use of this method does not establish compliance. Its methods require tailoring to the setting and participation by stakeholders capable of identifying harms that internal teams may miss.

What R-BED is not. It is a consequence-aware, system-level evaluation and decision methodology, not an enterprise risk-management system. Organization-wide policy, training, third-party oversight, specialized security, privacy and safety assessments, and the implementation of governance are outside its scope (Section 14, rows 16–20).

The thirteen-dimension structure is an engineering proposal, not an empirically validated taxonomy. A positive evaluation does not demonstrate that unknown failure modes are absent.

Materiality tiers. (Distinct from the monitoring priority tiers T1–T4 of the CSE Specification; Section 2.1.) R-BED requires evaluation intensity to increase with consequence, exposure, uncertainty and risk tolerance, but it does not prescribe or validate universal numeric tier thresholds. Adopting organizations may define low/medium/high or other risk tiers in their governance policy, record the assigned tier and rationale in the SES, and link each tier to test depth, independence, monitoring and approval rules. Tier labels are organizational implementation choices, not a universal R-BED score. How materiality, inherent risk and residual risk differ, and what each decides, is set out in Section 8.4.

1.1 Claim levels for statements in this document

Each statement about R-BED itself carries one of five levels, so that a reader can tell a design intention from a demonstrated result.

Level	Meaning and current use
Specified	Defined in this document; says what a user should do, not that doing so works. Most of this document is at this level. The maturity label [E] means a specified core with reference code that passes its own tests; it does not mean demonstrated effectiveness.
Implemented	A reference implementation of the record format and checks exists. It is optional tooling, separate from this specification.
Software-tested	The implementation passes its own automated tests (they test that the code does what the specification says; they do not test that the method improves decisions).
Empirically supported	Demonstrated by independent studies on real systems. <i>No statement of R-BED is currently at this level.</i>
Hypothesis	A proposed benefit awaiting study (Section 13); for example, that independent control challenge finds failures that self-assessment misses.

2 Terms and unit of assessment

Term	Meaning
Configured system	The particular model, data, prompts, retrieval sources, tools, permissions, human roles, monitoring, interfaces, and operating environment being evaluated.
System claim	A falsifiable statement about acceptable system behavior within explicit deployment conditions, tolerance or prohibition, population, and time horizon.
Failure event F	The system violates a specified claim or acceptance predicate in a defined exposure unit.
Deployment harm H	A failure passes applicable controls and produces a consequential effect in a workflow, person, or external system.
Materiality	A judgment that a claim, pathway or consequence is significant enough to warrant evaluation effort, evidence depth and decision attention; it sets priority and does not measure loss (Section 8.4).
Inherent risk	The consequence potential of a pathway under the intended authority and exposure, before any deployment control is credited (Section 8.4).
Residual risk	What remains after credited, evidenced controls: recorded as an upper bound or an explicit abstention (Section 8.4).
Stressed loss; tail risk	Loss under a specified severe-but-plausible scenario; and statistics of the upper tail of a loss distribution. Neither is expected loss (Section 8.5).
Exposure terms	Exposed (total business), AI-influenced, affected (touched by a failure) and harmed populations, and the incremental consequence, are different quantities (Section 8.6).
Pathway	A defined route from system behavior to consequence: failure-mediated (F), intended-behavior (N), or misuse (M) (Section 8).
Consequence S	The magnitude and type of harm, including non-monetary effects that should not be forced into a price.
Consequence Register	The per-system record of consequence categories, severity, reach, duration, reversibility and tolerance (Section 5).
Evaluability	Capacity to reliably obtain current, credible, connected, and challengeable evidence about a system claim.
Exposure unit	The unit of analysis: request, session, transaction, decision, case, user, or time period, stated consistently.
Evidence status	Supported for stated conditions / Refuted / Unresolved / Not applicable—with reasons and evidence pointers.

Term	Meaning
Independence level	I0–I3: how far the producer of evidence is separated from the builders and operators (Section 7).
Disposition	The decision taken on a configured system: Deploy, Restrict, Escalate, Remediate or Stop (Section 11).
Evidence record	An item of evidence (identifier E1, E2, ...) with its source, test, independence level and the claim it bears on. Its currency is tracked separately in an evidence-validity record (EVID-01, ...; CSE Specification Section 5).
Authorization record	A named, dated, expiring record that a disposition of Deploy or Restrict was approved on stated evidence, with invalidation conditions and monitoring.
Materiality tier; monitoring priority tier	Two different scales: an organization-defined materiality tier sets evaluation depth; a monitoring priority tier (T1–T4, assigned by rules TR1–TR6) sets monitoring cadence after authorization (Section 2.1).
Control assurance level; independence level	L0–L3 describes how well a control is shown to work; I0–I3 describes how far the producer of the evidence is separated from builders and operators (Section 2.1).

2.1 Structure and notation reference

R-BED has five layers that answer different questions. They are not competing architectures. Letters identify dimensions only (A, B, C) and vertices only (V); other identifiers use longer prefixes so that no code is used for two things.

Layer	Count	Role and identifiers
Architectural vertices	3	The three questions of the architecture: V1 Technical Failure, V2 Deployment Controls, V3 Consequence and Impact. There is no fourth vertex.
Diagnostic dimensions	13	Locations for claims and evidence, never scores: A1–A6 under V1, B1–B4 under V2, C1–C3 under V3 (C3 is assessed as C3a regulatory and legal and C3b reputational and trust).
Cross-cutting objectives	5	Lenses applied across the dimensions: Safety, Security, Privacy, Robustness, Resilience. They are not additional dimensions.
Methodological commitments	3	What the method does differently: consequence-aware evaluation, evidence-based control assurance, continuous evaluation and authorization. They are differentiators, not architectural layers.
Procedures	3	The eight-step workflow, the seven-stage lifecycle and the ten-step continuous evaluation protocol (below).

Scales that look alike. Each is named in full wherever it could be mistaken for another.

Scale	Values	What it answers
Materiality tier	Set by the organization	How much evaluation effort a claim or pathway deserves, before authorization. Not a loss estimate.
Monitoring priority tier	T1–T4 (T1 highest)	How closely an authorized system is watched; assigned by tier rules TR1–TR6 (CSE Specification Section 9).
Control assurance level	L0–L3	How well a deployment control is shown to work.
Independence level	I0–I3	How far the producer of the evidence is separated from builders and operators (Section 7).
Reversibility class	R0–R3	How reversible an agent action is (CSE Specification Section 7).

Authorization rule; disposition	Rules 1–8; five dispositions	The conditions on, and the outcome of, an authorization decision (Section 11). “Continue” is the renewal of a Deploy authorization, not a sixth disposition.
--	------------------------------	--

Identifiers of records and studies. Claims CL-01, ...; scenarios SCN-01, ...; Register entries K1, ...; controls CTL1, ...; evidence items E1, ...; evidence-validity records EVID-01, ...; monitoring-plan items MON-01, ...; triggers TRG-01, ...; treatments TRT-01 and residual-risk records RSK-01, ...; learning records LR1, ... In the Validation Plan, hypotheses H1–H7, studies S1–S7 and allocation baselines BASE-01, ...; in Section 8.2, the hypothetical configurations CFG-01 to CFG-03.

Three procedures, one process. The eight-step workflow (Section 6) is the evaluation procedure for one configured system. The seven-stage lifecycle (Section 9) is the organizational view: who is accountable at each stage from scoping to retirement. The ten-step protocol P1–P10 (CSE Specification Section 4) continues the evaluation after authorization. They are views of one process, with this approximate correspondence:

Workflow steps	Lifecycle stages	CSE protocol
01 Context, 02 Claims, 03 Pathways	1 Scope and justify; 2 Characterize consequences	P1
04 Design evidence, 05 Break, 06 Challenge controls	3 Design controls; 4 Evaluate	P2–P4
07 Evaluate and decide	5 Decide and authorize	P5
08 Fix, verify and renew	6 Operate and monitor; 7 Review, retest, retire	P6–P10

3 The R-BED architecture: 13 dimensions

A dimension is a diagnostic location for a claim and evidence; it is not a standalone score. A system may need multiple evaluations within one dimension, and one experiment may inform several dimensions. Questions and sample tests below are illustrative, not mandatory or exhaustive. The architecture is 6+4+3.

3.1 V1 Technical Failure (dimensions A1–A6)

Dimension	Question / example of relevant evidence
A1 Data Foundations	Are source data, indexes, and knowledge authorized, suitable, current, and representative? Source lineage, freshness/coverage checks, and provenance tests.
A2 System Design	Are architecture, interfaces, dependencies, and failure containment mechanisms suitable? Design review, contract tests, dependency and fault-injection tests.
A3 Predictive Validity	Does configured system behavior satisfy an explicit acceptance predicate? Representative and stratified tests, uncertainty intervals, and deployment-like evaluations.
A4 Fairness	Are errors, benefits, and harms distributed acceptably across relevant groups? Disaggregated outcomes with intervals and scenario-specific fairness checks.
A5 Failure Analysis	Where, when, and for whom do claims break? Adversarial discovery, reproducible scenario tests, characterized failure regions and relevant population mass.

Dimension	Question / example of relevant evidence
A6 Explainability	Can a reviewer reconstruct consequential behavior? Tool and authorization traces, inputs and version identifiers, lineage and replay tests.

3.2 V2 Deployment Controls (dimensions B1–B4)

A deployment control is a mechanism outside the model's own judgment that limits what the system can do, how far an effect spreads, or how quickly harm is noticed and stopped. Each definition below states what must hold, what evidence shows it, and what does *not* count.

Control	Definition (what must hold)	Evidence that shows it	Does not count as evidence
B1 AI Authority	The inventory of actions the system can cause directly or through tools, each with a privilege level (read, draft, recommend, act-with-approval, act-autonomously), scope and limit; every limit is enforced by a mechanism that does not depend on model output.	Permission inventory; external authorization checks; deny-path and forced-attempt tests; least-privilege review; rate of completed versus attempted forbidden actions with an interval.	A prompt instruction or model refusal; a policy document without an enforcing control; a single passing request.
B2 Human Mediation	A human can <i>effectively</i> intervene: has the information, time, competence, authority and independence to detect and override, and the workflow does not make approval a formality.	Review information shown; seeded-error detection rate; override and escalation latency; workload and approval-rate analysis for automation bias; reviewer-competence records.	The existence of an approval step; a high approval rate read as quality; reviewers who cannot see the basis for the action.
B3 Exposure Control	Magnitude, rate, scope and duration of any single failure or misuse are bounded before containment, and the bound can be reversed or rolled back.	Tested amount, rate and scope limits under forced failure; blast-radius analysis; rollback drill; staged-rollout record; concurrency tests.	A limit defined in configuration but never exercised; a limit enforced only inside the model; an average-case volume estimate. This is the effectiveness of the limiting safeguards (a control question); how far consequences could reach if they fail is C2 Scale.
B4 Harm Detection	Harmful effects, not only technical errors, are detected with enough coverage, sensitivity and speed that a named owner can act before the consequence exceeds tolerance.	Alert sensitivity and coverage on seeded harms; false-alarm burden; escalation owner; alert-to-intervention time; monitoring drills.	Dashboards nobody owns; log retention without alerting; detection of failures that do not map to harms.

3.3 V3 Consequence and Impact (dimensions C1–C3)

The three dimensions are named C1, C2 and C3 so that each states what it measures, and the vertex asks about consequences to people and society, recorded in the Consequence Register (Section 5), not only to the organization. C1 covers incremental harm of any material kind, not only financial loss, and C3 names reputational and trust consequences beside regulatory and legal ones; both are proposed [P]

ILLUSTRATIVE FAILURE-MEDIATED PATHWAY — widths show conceptual narrowing only

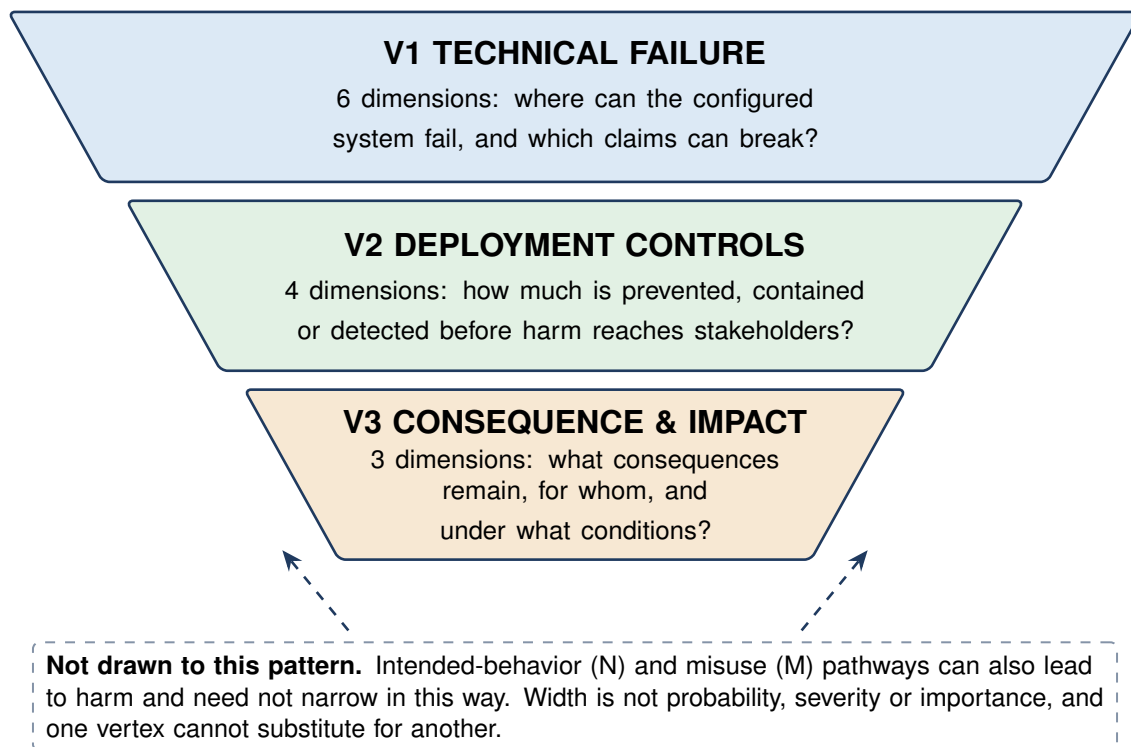


Figure 2. The R-BED architecture as an illustrative risk-narrowing pathway. It illustrates a failure-mediated pathway only and does not replace Figure 1. Technical Failure is drawn widest because it contains the most diagnostic dimensions and the most places where a System Claim may break. Deployment Controls determine how much of a failure-mediated pathway is prevented, contained or detected before harm reaches stakeholders. Consequence and Impact asks what material consequences remain, for whom, and under what conditions. Widths indicate conceptual narrowing only; they do not represent probability, severity, importance or weighting, and the figure is not a quantitative model. Intended-behavior and misuse pathways may also lead to harm without following this narrowing. The three vertices remain non-substitutable: strong controls do not offset weak technical evidence, or the reverse.

practice and have not been validated.

Dimension	Question / example of relevant evidence
C1 Incremental Impact	What additional harm, financial and nonfinancial, is attributable to an AI-influenced failure <i>per affected</i> decision, transaction, individual or exposure? Direct loss, economic harm, privacy violation, unfair treatment, safety consequence and other material harm, against a stated counterfactual baseline and not against total business value or nominal exposure. Each C1 entry states its <i>unit</i> (per harmful outcome, per affected exposure, or per AI-influenced decision) and its baseline (prior process, human-only process, or no system); figures on different units are not interchangeable. Evidence: per-unit scenario and loss analysis, baseline comparison, recoverability.
C2 Scale	How widely can the consequences occur? Affected decisions, customers and transactions; exposure concentration; correlated and common-cause failure; duration; potential for systemic propagation. Four populations are kept apart: exposed (total business), AI-influenced, affected (touched by a failure) and harmed (Section 8.6). C2 describes potential consequence reach, whereas B3 Exposure Control asks whether safeguards effectively limit it. Evidence: volume, concurrency, concentration and propagation analyses.

Dimension	Question / example of relevant evidence
C3 Regulatory and Reputational Impact	What broader institutional consequences can follow? Assessed as two components, <i>separately and not automatically combined into one number</i> : C3a Regulatory and legal , meaning regulatory scrutiny, compliance breach, enforcement exposure, litigation and related obligations (evidence: obligation mapping, decision trail, notices and rights); and C3b Reputational and trust , meaning customer confidence, public trust, institutional credibility, attrition and longer-term franchise consequences (evidence: stakeholder and trust-sensitivity analysis, disclosure and remediation plan).

How C1, C2 and C3 relate. C1 is the consequence per affected exposure, C2 is how many exposures are affected and for how long, and C3 is the institutional consequence that may follow. Each is assessed and recorded on its own. Two aggregations are distinct. A *scenario* consequence in a homogeneous illustration is $L_{\text{scenario}} = N_{\text{affected}} \times \Delta L$, a stated-values calculation with no probability attached. An *expected-loss* aggregate is $E[L_{\text{total}}] = \sum_i E[L_i]$, an expectation across pathways (Section 8.6). Neither is a universal financial-risk equation, and each applies only where units, denominators and conditioning assumptions match; ordinal assessments are never combined by multiplication, weights or a composite score. C3 can create financial consequences (fines, litigation cost, remediation, lost revenue from attrition); where such an amount is already counted in a direct or aggregate loss estimate it is recorded once, where it is priced, and C3 cites it rather than adding it again. Not every failure produces regulatory or reputational consequences: a C3 entry states the conditions under which they would arise (for example a systematic notice defect, a pattern affecting protected groups, or public disclosure), and “none expected” is a legitimate finding only when those conditions have been examined (Rule 8: residual risk is not silently zero).

Consequence attributes are documented separately in the Consequence Register (Section 5); they do not create new R-BED dimensions. C3a corresponds to the Register category *Legal and regulatory* and C3b to *Reputational and trust*; C1 draws on the per-unit severity of any category. Recovery, reversibility and recourse are explicit attributes of each Register entry, not dimensions: reversibility and operational recovery (rollback, containment) describe what can be undone before or as harm occurs, and recourse describes the remedy available to those harmed afterwards. A purely monetary representation is insufficient for unpriced harms.

4 Cross-cutting objectives

Safety, security, privacy, robustness and resilience are not additional dimensions. Each is a lens: a question asked of the thirteen dimensions, with its own primary locations of evidence and, where needed, a specialized assessment that R-BED does not replace. Treating them as lenses preserves the 6+4+3 architecture and avoids a fourteenth-to-eighteenth “dimension” that would invite scoring.

Objective	Question the lens asks	Typical evidence	Specialized assessment (outside R-BED)
Safety	Can the system contribute to physical, psychological or other harm to people?	scenario tests, drills, incident review	domain safety assessment where sector rules apply
Security	Can an adversary or unauthorized actor cause the system to act against its owner or users?	adversarial tests, permission audit, red-team	security testing and threat modelling
Privacy	Does the system expose, infer or retain personal information beyond its purpose?	data-flow review, leakage tests	privacy impact assessment

Objective	Question the lens asks	Typical evidence	Specialized assessment (outside R-BED)
Robustness	Does the system keep meeting its claims under shift, noise and edge conditions?	stratified tests, shift tests	statistical validation
Resilience	Can the deployment detect, contain and recover from failure and keep essential function?	recovery drills, fallback tests	continuity and recovery exercises

Cross-cutting objectives by dimension. ● = primary location of evidence for the objective; ○ = secondary. The primary marks are generated from the same table used by the reference implementation; an empty cell does not mean the objective is irrelevant to that dimension.

Dimension	V	Safety	Security	Privacy	Robustness	Resilience
A1 Data Foundations	V1			●	●	
A2 System Design	V1		●			
A3 Predictive Validity	V1				●	
A4 Fairness	V1				●	
A5 Failure Analysis	V1	●			●	
A6 Explainability	V1			●		
B1 AI Authority	V2	●	●			
B2 Human Mediation	V2					●
B3 Exposure Control	V2	●	●	●		●
B4 Harm Detection	V2	●	●			●
C1 Incremental Impact	V3	○	○			●
C2 Scale	V3	●	○		○	●
C3 Regulatory and Reputational Impact	V3	○	●	●		○

Using the lenses. For each system, the SES records whether each objective is in scope and why. An objective declared out of scope needs a reviewable reason, as with any N/A status. For an in-scope security or privacy objective on a high-consequence system, the SES must point to the specialized assessment (for example security testing or a privacy impact assessment); R-BED does not perform or substitute for it, and it cannot establish legal compliance.

4.1 Security evaluation protocol (method-level)

R-BED treats security as the question whether an adversary or unauthorized actor can cause the configured system to act outside its authority. The protocol below is deliberately at the level of evidence requirements, not attack technique.

1. **Threat model.** Name the adversary classes (external user, malicious insider, content author whose text the system reads, compromised tool or vendor), their access and their objective, and record what is out of scope.
2. **Authority abuse.** For each authority grant (Section 3), attempt each forbidden action through every

integration surface and record attempted versus completed counts with an interval.

3. **Control bypass.** Test whether limits are enforced outside the model and whether an input can change the control's inputs or configuration.
4. **Adaptivity.** Treat results as valid for the tested attacker effort and version; a new model, tool or prompt reopens the test. A static test set understates an adaptive adversary.
5. **Reporting.** Record trace identifiers, version tuple, tester independence (I0–I3), and residual risk as an upper bound or explicit abstention.

5 Consequence Register

The Consequence Register is the per-system record of what could happen to whom, linked to the scenarios and claims that bear on it. It records twelve categories, each assessed on the same attributes, rather than a single “business impact” question.

Category	What the entry records (examples)
Physical safety	Injury, loss of life, unsafe physical state; severity and recoverability for those affected.
Security	Unauthorized access, control, or action; integrity and availability effects on systems and data.
Privacy	Exposure, inference or retention of personal data beyond purpose; affected individuals and permanence of disclosure.
Financial	Direct loss, remediation cost, obligations; concentration and recoverability.
Legal and regulatory	Breach of obligation, notice duties, contractual exposure; applicable rules named by the adopting organization.
Discrimination and fairness	Unequal errors, access or treatment across groups; protected characteristics as defined by applicable law.
Psychological	Distress, manipulation, undue reliance; vulnerable populations.
Dignity and autonomy	Loss of agency, recourse or due process; decisions about people without meaningful explanation or appeal.
Operational continuity	Interruption of essential functions; degradation modes and recovery time.
Environmental	Resource and emissions effects where material to the use.
Societal and institutional	Effects on public services, information environments, institutions or markets beyond individual parties.
Reputational and trust	Loss of stakeholder confidence; recorded separately so it does not mask harm to others.

Attributes per entry: scenario and pathway (F, N or M), affected parties, severity (low, moderate, high, severe), reach, duration, reversibility (reversible, partially reversible, irreversible), tolerance or prohibition, and the claims it bears on.

- A consequence is *high-consequence* for R-BED purposes if its severity is high or severe, or it is irreversible. The classification drives stricter authorization rules (Section 11).
- A categorical prohibition recorded in the register cannot be exchanged for business benefit, and a small expected loss does not make it acceptable.
- Non-monetary categories are not converted to prices to be summed with financial ones. Entries are compared category by category.
- Affected parties, or people able to represent them, should take part in populating the register; internal teams will miss harms.

6 Eight-step R-BED evaluation workflow

01 / Context Document intended purpose, benefits, relevant populations, foreseeable misuse, exposure, risk tolerance, legal constraints, and system/version boundaries. Record why AI is used and which alternatives were considered.

02 / Claims Define material System Claims as observable acceptance boundaries, quantitative or categorical. Record prohibitions that cannot be exchanged for business benefit.

03 / Enumerate pathways Identify likely and high-consequence failure, intended-behavior and misuse pathways across the 13 dimensions and five objectives; populate the Consequence Register; include feedback, public incident reports, domain experts, and affected stakeholders where appropriate.

04 / Design evidence Select component, trajectory and outcome tests. Specify representative sampling, targeted adversarial tests, fault injection, instruments, stop conditions, evidence-independence level, and independent reviewers.

05 / Break and characterize Discover failure points; describe regions; reproduce under pinned configurations; estimate failure rates, coverage, and uncertainty without treating zero observations as proof of safety.

06 / Challenge controls Force relevant failures and misuse and test whether authorization, human mediation, exposure controls, and detection limit downstream effects before harm occurs. Estimate control effectiveness with an interval.

07 / Evaluate and decide Join technical, control and consequence evidence. Record claim status, uncertainty, residual risk (an upper bound or an explicit abstention), gaps, owner, and a disposition with its authorization record.

08 / Fix, verify and renew Place safeguards at the effective layer, independently retest material fixes, revise system claims and re-evaluate upon drift, model/tool changes, incidents, expiry, or expansion of authority.

7 Evidence architecture, independence and evaluation quality

Five proposed evaluability properties strengthen whether a system can be assessed: observability (events and context are recorded), testability (claims can be challenged safely), traceability (the end-to-end path can be reconstructed), reproducibility (configuration and stochastic variation are characterized), and actionability (evidence can trigger restriction or intervention). These are design objectives, not a formally proven necessary-and-sufficient set.

Review of the evaluation method. Before treating results as decision-grade evidence, a reviewer should test whether each metric measures the intended claim (construct validity), whether samples reflect the operating population, whether rare or severe failure regions are represented, whether uncertainty is bounded, and whether independent retest is feasible. Document metric limitations, detector blind spots, drift triggers and changes to the test protocol. This is a proposed review discipline, not empirical proof that any metric or test process is effective.

- **Component evidence:** unit tests of models, retrievers, tool wrappers, permission checks, monitors, and business rules.
- **Trajectory evidence:** ordered record of retrieval, prompts/versions, tool arguments, approval checks, human intervention, and state transitions.
- **Outcome evidence:** consequences for users, records, money, rights, performance, service obligations, and affected groups.

These perspectives are many-to-many with vertices: for example a component authorization test informs V2, and an outcome error rate can inform V1. Critical evidence should carry version tuple, timestamps, provenance, scenario ID, tester/approver, sampling frame, numerator and denominator, uncertainty, and

links to raw traces or appropriately protected records.

7.1 Evidence statuses

- **Supported (bounded):** adequate evidence supports the claim only for the specified population, versions, time window, and conditions. For a claim measured as a rate, the *acceptance threshold* (the maximum tolerable rate) is fixed before the test, and the claim is Supported only if the upper confidence bound of the result is at or below it. A result whose upper bound exceeds the threshold leaves the claim Unresolved.
- **Refuted:** admissible evidence shows violation of a required condition or control.
- **Unresolved:** crucial evidence, coverage, validity, independence, or uncertainty is inadequate; unresolved is not synonymous with safe.
- **Not applicable:** a documented, reviewable justification explains why a dimension or claim does not apply.

7.2 Independence levels

Independence describes how far the producer of an item of evidence is separated from the people who built or operate the system. It qualifies evidence; it does not make evidence correct.

Level	Definition
I0	Self-assessed by the team that built or operates the system.
I1	Produced by a separate team within the same reporting line.
I2	Produced by an independent function within the organization (for example internal audit or a separate assurance group) with its own reporting line.
I3	Produced by an external party with no commercial dependence on the outcome.

For high-consequence systems, a Deploy or Restrict disposition should not rest only on I0 evidence (rule 7, Section 11). Higher independence is a design choice proportional to consequence, not a requirement for every system.

7.3 Three separate records

R-BED keeps three records distinct so that one cannot silently stand for another: the *test result* (counts, intervals, traces), the *evidence assessment* (status of each claim and why), and the *deployment decision* (disposition and authorization). A passing test does not by itself change a claim's status, and a status does not by itself authorize deployment.

8 Quantitative risk interpretation

8.1 The conditional pathway identity

For one explicitly defined failure-mediated pathway, one exposure unit, and deployment configuration A , let F denote the failure event, H a harm caused by that failure after the deployment controls, and $S \geq 0$ the priced severity of the resulting harm, defined conditional on both F and H . Define the failure-attributable loss $L_F = S \times \mathbf{1}\{F\} \times \mathbf{1}\{H\}$: it equals S when F and H both occur, and equals zero otherwise. The chain rule and conditional expectation then give:

$$E[L_F | A] = P(F | A) \times P(H | F, A) \times E[S | H, F, A]$$

This is a conditional identity under the definitions above, provided the relevant conditional probabilities and expectations exist; it is not a validated loss forecasting model, a universal risk score, or a claim

that non-monetary harms can be priced. $P(F | A)$ requires failure evidence under a stated operating distribution; $P(H | F, A)$ requires challenge tests conditional on failure; and the severity term requires consequence evidence. If the conditioning event has zero probability, the corresponding factors require separate treatment rather than numerical interpretation. Dependencies, heavy tails, underreporting, shared-cause failures, unobserved pathways, and overlapping events must be disclosed. Do not sum pathways without addressing overlap; do not infer that a small average loss makes an irreversible or prohibited harm acceptable.

8.2 Ten extensions

Each extension states what it adds and the assumption that limits it. None converts the thirteen dimensions into a score.

#	Extension	Statement	Assumption or limit
1	Intended-behavior pathway N	$E[L_N A] = P(N A) E[S N, A]$, where N is harm caused by behavior that meets every claim. No failure factor appears.	Requires consequence evidence for correct behavior; a system that passes every claim can still cause harm.
2	Misuse pathway M	$E[L_M A] = P(\text{attempt}) P(\text{success} \text{attempt, controls}) E[S \text{success}]$.	Attempt probability is not stationary: it responds to the system and its controls. Test results hold for the tested effort and version.
3	Union of pathways	For nonnegative pathway losses with union loss equal to the point-wise maximum, $\max_i E[L_i] \leq E[L_{\cup}] \leq \sum_i E[L_i]$.	Bounds, not an estimate. The sum is the safe upper bound when overlap is unknown; the maximum is attained only under full overlap.
4	Dependent controls (Fréchet)	For failure probabilities p_a, p_b of two controls with unknown dependence, $\max(0, p_a + p_b - 1) \leq P(\text{both fail}) \leq \min(p_a, p_b)$.	Use whenever independence is not demonstrated. For $p_a = 0.1, p_b = 0.2$ the joint failure probability lies in $[0, 0.1]$.
5	Independent controls	If failures are independent, $P(\text{all fail}) = \prod_i p_i$.	Independence is an assumption to be evidenced (separate code, data, vendors, failure causes); shared dependencies defeat it.
6	Combining intervals	Bonferroni: k intervals with confidence c_i jointly cover with probability at least $1 - \sum_i (1 - c_i)$. The product of nonnegative intervals is an interval for the product.	Conservative. Three 95% intervals jointly guarantee only 85%. The product bounds ignore dependence among factors.
7	Rare-event trial counts	Zero failures in n trials gives a one-sided upper bound $1 - (1 - c)^{1/n}$. At 95%, about 2,995 clean trials bound the rate below 10^{-3} and about 2,995,731 below 10^{-6} .	Assumes independent trials from the operating distribution. A bound is not a measurement of the rate.
8	Control effectiveness by intervention contrast	Compare failure rates without and with a control under otherwise identical tests, $P(H F, \text{do}(\text{control off}))$ versus $P(H F, \text{do}(\text{control on}))$; report a conservative ratio, upper bound with control over lower bound without.	Causal only if the two arms differ solely in the control. Hypothetical: 12/40 without and 1/80 with give a conservative ratio of 0.41.
9	Exceedance over exposure	With per-exposure rate p over n independent exposures, $P(\text{at least one}) = 1 - (1 - p)^n$.	A rate that looks small per exposure is large at volume: $p = 0.01$ over 100 exposures gives 0.63.

#	Extension	Statement	Assumption or limit
10	Abstention rather than silent zero	If evidence is too thin for a defensible estimate (for example fewer than a stated minimum number of trials), record an explicit abstention with a reason; never record zero.	Abstaining keeps the claim Unresolved; it is a decision input, not a favorable result.

Threshold first, then sample size. Because a bound is not a measurement, the acceptance threshold must be chosen from the risk tolerance before testing, and it fixes the evidence needed: with zero failures at 95% confidence, 29 trials suffice for a threshold of 0.10, 299 for 0.01 and 2,995 for 0.001. These counts use the one-sided bound; the two-sided 95% Clopper–Pearson intervals recorded in an SES need 36, 72 and 368 clean trials for thresholds of 0.10, 0.05 and 0.01. A result such as 0 of 60 (two-sided 95% upper bound 0.060; one-sided 0.049) can support a claim against a 0.10 threshold for a restricted pilot; it does not support a claim against a stringent tolerance, and it must not be reported as if it did.

8.3 Hypothetical comparison of three configurations

SYNTHETIC All values below are invented for illustration; none is an observation. A fictional agent is compared in three configurations of the same system, with exposure unit a session and severity in arbitrary currency units: CFG-01 no independent containment, CFG-02 an independent server-side limit check, CFG-03 limit check plus reduced authority.

Config.	Description	$P(F A)$	$P(H F, A)$	$E[S H, F, A]$	$E[L_F A]$
CFG-01	No independent containment	0.04	0.80	60	1.92
CFG-02	Server-side limit check	0.04	0.01	70	0.028
CFG-03	Limit check, reduced authority	0.03	0.005	60	0.009

The failure rate is unchanged between CFG-01 and CFG-02: the containment acts on $P(H | F)$, not on $P(F)$. Severity is higher in CFG-02 because the harms that remain are the uncontained larger ones. Three cautions apply. The ordering $\text{CFG-01} > \text{CFG-02} > \text{CFG-03}$ holds only if each factor was estimated with evidence, and each factor has its own interval. A small expected value does not make a prohibited or irreversible harm acceptable. A comparison within the identity says nothing about pathways N and M, which need their own entries.

8.4 Materiality, inherent risk and residual risk

Maturity: the concepts are standard in risk practice; their roles and requirements below are proposed [P] and have not been validated. R-BED uses three terms that are often run together; they decide different things and are used in a fixed order.

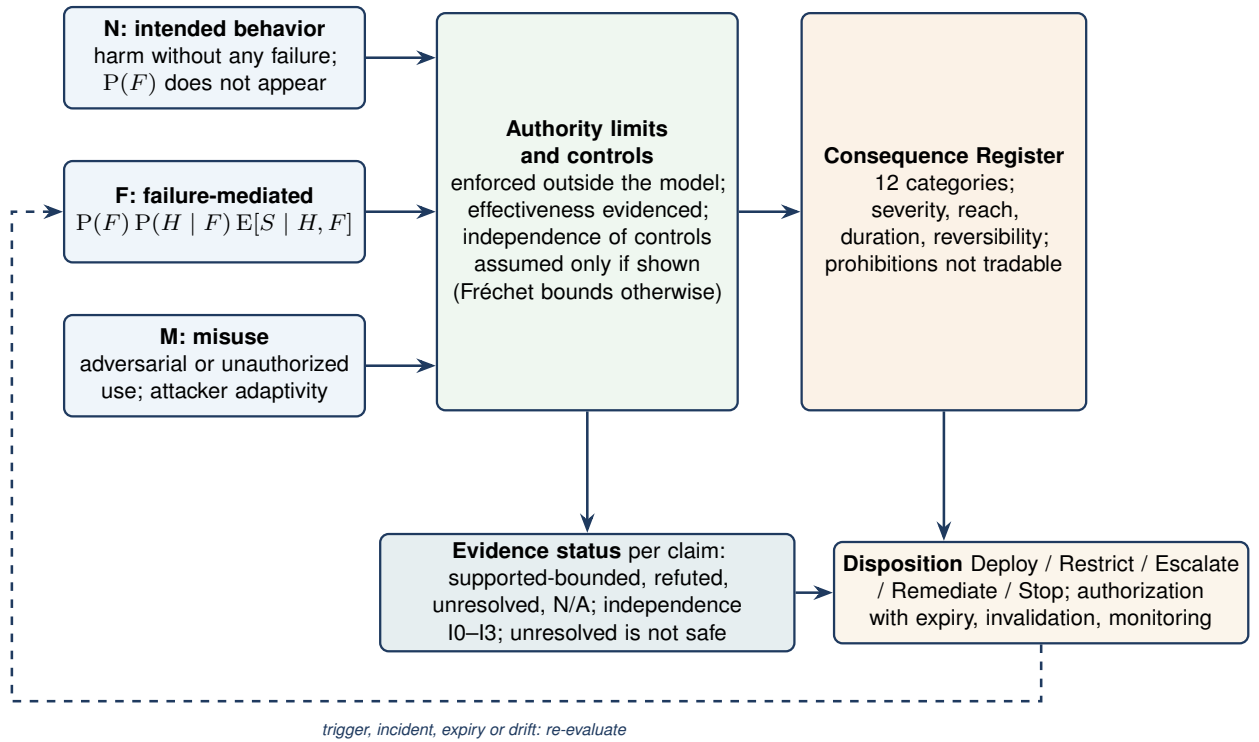


Figure 3. Pathways to consequence and to decision. Intended-behavior (N), failure-mediated (F) and misuse (M) pathways pass through authority limits and controls to the Consequence Register; evidence status and independence feed a disposition and its authorization record, which is reopened by triggers, incidents, expiry or drift.

Term	Definition	Role in R-BED
Materiality (inherent materiality)	The significance of the intended business purpose, the exposure and the potential consequences, judged before testing and before controls, as a reason to evaluate a claim or pathway and to give it decision attention. It is a threshold on importance, not an estimate of loss and not a measure of residual risk, and not the accounting or regulatory sense of the word unless the adopting organization adopts it.	<i>Prioritization:</i> selects which claims and pathways are evaluated, at what depth and independence, and (with the tiers of the CSE Specification) how closely they are monitored.
Inherent risk	The consequence potential of a pathway under the intended authority, exposure and configuration, with deployment controls not yet credited (a control without evidence of effectiveness is not credited, Rule 6). Recorded as the Consequence Register attributes (severity, reach, duration, reversibility) and, where evidence supports it, as expected loss without the control.	<i>Consequence assessment:</i> fixes whether an entry is high-consequence (Section 5). The classification follows the inherent consequence and is not lowered because a control exists.

Term	Definition	Role in R-BED
Residual risk	What remains under a specified system configuration after the controls that have been tested and credited with evidence: the conditional factors of Section 8.1 estimated with those controls in place, recorded as an upper bound or an explicit abstention (Rule 8). Never silently zero; transfer or acceptance is not reduction.	<i>Authorization</i> : the quantity the approving authority compares with the stated tolerance when issuing, restricting or renewing an authorization. It is not an input to the high-consequence classification.

Order of use. Materiality directs effort (a highly material system can have low residual risk, and a minor one can have a high residual risk that still needs a decision); inherent risk characterizes what could happen; evaluation tests whether credited controls work; residual risk, with its uncertainty, is what is authorized. A strong control therefore changes the residual risk and the evidence required to keep it, not the inherent classification, which continues to trigger the stricter rules (Rules 4, 5 and 7). A residual risk below tolerance is a bounded, dated statement that lapses with its evidence (CSE Specification Section 5); it is not a proof of safety.

8.5 Stressed, correlated and tail loss

Maturity: [P]; the measures are standard, the requirement for high-consequence entries is proposed. The identity of Section 8.1 gives an *expected* loss under the operating distribution. Three quantities should not be confused.

Quantity	Meaning and use
Expected loss	$E[L]$ per exposure unit and horizon, from the identity. An average; it can be small while a severe outcome is possible.
Stressed loss	The loss L_s (or its ordinal severity and reach) under a stated, severe-but-plausible scenario s , with the scenario's assumptions listed. No probability for s is needed; the question is whether the system and the organization can withstand it.
Tail-risk measure	A statistic of the upper tail of the loss distribution over a horizon: the quantile VaR_α , or $ES_\alpha = E[L \mid L \geq VaR_\alpha]$. Optional. Used only where a defensible loss distribution exists (enough relevant observations or a justified model, priced severities, modeled dependence); otherwise the stress scenarios and the bounds of Section 8.2 are reported instead, and no tail figure is invented.

What a stress scenario must cover. Severe-but-plausible failures; *common-cause* failures, in which one shared model version, prompt, retrieval source, vendor, credential, control or monitor fails many exposures at once; *correlated exposures*, in which one input, event or adversary reaches many exposure units together and can exhaust response capacity; *concentration*, in which affected exposures cluster in one segment, counterparty, sector, geography or channel; and potentially catastrophic consequences (irreversible, widespread or hard to contain). Independence of controls is not assumed; where dependence is unknown the Fréchet bounds (extension 4) apply.

Requirement for high-consequence applications. For each high-consequence Consequence Register entry, and for any system granted authority to act at scale, the SES records at least one stress scenario in the Scenarios block, with the shared components, the number of exposures affected, the containment assumed, and a qualitative (ordinal severity and reach) or quantitative stressed-loss assessment with its bounds. The material claim that the stressed outcome is within tolerance has an evidence status like any other; without an assessment it is Unresolved, and under Rule 4 unresolved is not safe. This adds no new rule or disposition. Expected-loss evidence, however favorable, does not stand in for it.

8.6 Exposure units and portfolio aggregation

Maturity: [P]; the identity below is elementary and is not claimed as new. Five quantities are easily conflated and should be kept apart, and each rate or loss must name its exposure unit and denominator.

Quantity	Meaning
Total business exposure (exposed population)	Everything the activity puts at risk (for example all accounts, notional, transactions or customers), whether or not an AI system touches it.
AI-influenced exposure	The part of the total in which the configured system's output influenced a decision or action. A subset of the total; influenced is not the same as caused.
Affected decisions	The subset of AI-influenced decisions touched by a technical failure F , whether or not harm followed. A failure is not yet a harm.
Harmful decisions (harmed population)	The subset of affected decisions that produced harm (the event H after controls), by a stated criterion. This is the count behind C2; some of these would have been harmful under the prior process as well.
Incremental model-attributable consequence	The per-unit measure of C1: the change in consequence attributable to the system against a stated counterfactual baseline (the prior process, a human-only process, or no system): $L_{\text{incr}} = L_{\text{with}} - L_{\text{baseline}}$. It can be negative. A causal reading needs a design that supports it (CSE Specification Section 12).

Scenario versus expected-loss aggregation. Two calculations serve different purposes. A homogeneous scenario consequence is $L_{\text{scenario}} = N_{\text{affected}} \times \Delta L$ (stated values, no probability). The expected-loss aggregate below is an expectation across pathways. Neither is a universal financial-risk equation.

Aggregate expected loss. For exposure units or pathways $i = 1, \dots, m$ with finite expectation and no loss counted twice,

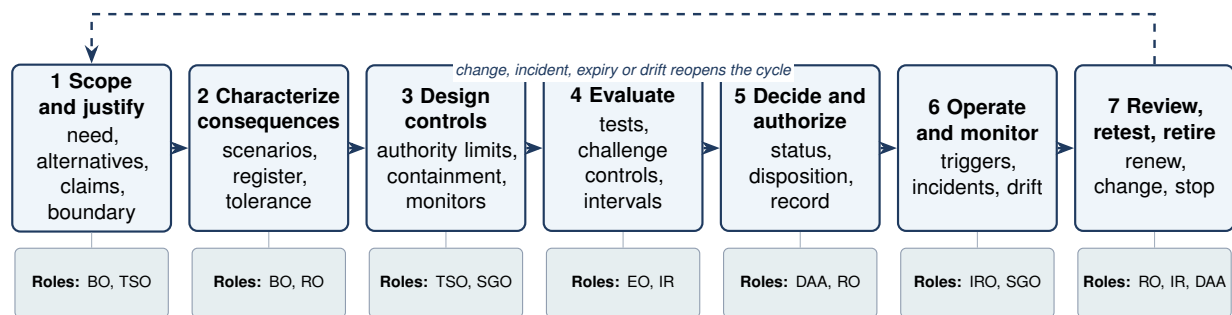
$$\mathbb{E}[L_{\text{total}}] = \sum_{i=1}^m \mathbb{E}[L_i].$$

This follows from *linearity of expectation* and *requires no independence*: correlated pathways have the same expected total as independent ones. If one event can be counted in several units, the sum overstates and the union bounds of extension 3 apply. Dependence matters for everything beyond the mean: $\text{Var}(L_{\text{total}}) = \sum_i \text{Var}(L_i) + 2 \sum_{i < j} \text{Cov}(L_i, L_j)$, and the quantiles or Expected Shortfall of the total cannot be recovered from the per-unit values without a joint model. In general the sum of per-unit tail measures is not the tail measure of the total: for ES it is an upper bound (ES is subadditive), for VaR there is no such bound in general, and the two are equal only under extreme dependence (comonotonic losses). Aggregate tail-risk assessment therefore generally requires modeling dependence, or a stress scenario that fails the units together; the portfolio report of Section 9.5 lists shared components as common-cause candidates for that reason.

Hypothetical financial-services illustration (SYNTHETIC; every number is an assumption chosen for the arithmetic, none is an observation). A lender uses an AI underwriting assistant whose recommendation informs small-business loan decisions; loans average 100,000. Four kinds of figure are kept apart. An *expected* figure is the mean under a stated probabilistic model with point parameters. An *upper bound* uses an assumed bound on a parameter and limits the *expected* value (it is not a maximum realized count or loss, and holds only while the assumption holds). A *scenario* figure uses stated values with no probability attached. The *incremental* figure is always with-system minus baseline, on the same decisions: $L_{\text{incr}} = (H_{\text{with}} - H_{\text{base}}) \times \ell$, where H counts harmful decisions and $\ell = 20,000$ is the assumed loss per harmful decision, including remediation, whichever process caused it. The baseline is a human-only process on the same 1,000 decisions with an assumed harmful-decision rate of 0.3% (3 expected). It depends on the exposure and not on the AI failure rate, so it is the same 3 in every row below; if the exposure changed, it would change in proportion.

Step	Assumed quantity	Reading
Total business exposure	5,000 loans originated in the quarter; 500,000,000	Everything at stake, AI or not.
AI-influenced exposure (C2)	Recommendation used on $N = 1,000$ loans; 100,000,000 (20%)	Analyzing a large portfolio does not make all of it exposed to AI-attributable loss.
Technical failure	$P(F A) = 5\%$: 50 expected failures	A failure is not yet a harm.
Harmful outcome	Assumed human review. Point value $P(H F, A) = 0.10$: 5 expected harmful decisions (notional 500,000, 0.1% of total). Assumed upper bound $P(H F, A) \leq 0.20$: at most 10 expected harmful decisions	The point value gives an expected count; the bound gives a ceiling under the assumption. They are not interchangeable.
Expected incremental loss (C1)	$(5 - 3) \times 20,000 = 40,000$	Mean under the point model. A realized quarter can differ.
Upper bound on expected incremental loss	$(10 - 3) \times 20,000 = 140,000$	A ceiling on the expected value, not on any realized quarter. Holds only while $P(H F, A) \leq 0.20$ and the baseline and ℓ hold. It is not conservative against a smaller baseline, which would raise it.
Aggregate	A second pathway, data disclosure, with assumed quarterly $E[L_2] = 10,000$. $E[L_{\text{total}}] = 40,000 + 10,000 = 50,000$	Expected figures only; holds whether or not the pathways are correlated. Bounds are not added to expectations.
Regulatory and reputational (C3)	C3a: if the affected decisions share a systematic defect in adverse-action notices, assessed <i>moderate</i> , conditional. C3b: <i>low</i> unless the pattern becomes public or concentrates in one segment	Ordinal, assessed separately, not converted to a number. The re-notification cost is already inside ℓ , so it is not added again.
Inherent (no credited control)	$P(H F, A) = 1$: 50 expected harmful decisions; $(50 - 3) \times 20,000 = 940,000$	Expected incremental loss with the control not credited. Same N , baseline and ℓ as above.
Reassessment trigger	The point value assumes a 90% catch rate and the bound an 80% rate. A monitored rate below 80%, or a major model update, invalidates the evidence	The claim reverts to Unresolved; under Rule 4 the authorization is not renewed without new evidence.
Stressed scenario	One shared update sets the failure rate to 15% and the catch rate to 60%: 150 failures, 60 harmful; $(60 - 3) \times 20,000 = 1,140,000$	A specified severe-but-plausible scenario, not a bound and not an expectation. About 28.5 times the expected 40,000 and 8.1 times the upper bound. Estimating the two factors separately and combining them as if independent would understate it when one cause moves both.

Ordering, with the same N , baseline and ℓ throughout: expected 40,000 < upper bound on expected 140,000 < inherent 940,000 < stressed 1,140,000. All figures are for one quarter of 1,000 AI-influenced decisions. The upper bound limits the expected value under its assumptions; it is not a maximum realized loss, and the stressed figure does not bound the others. The example illustrates the distinctions only. No VaR or ES figure is given because no defensible loss distribution is supplied, and nothing here is a result.



BO business owner · TSO technical system owner · RO risk owner · EO evaluation owner · IR independent reviewer · DAA deployment-approving authority · IRO incident-response owner · SGO safeguard owner. Roles are functions; the validator checks that the approving authority differs from the evaluation owner and business owner.

Figure 4. Seven lifecycle stages with the roles accountable at each. The cycle is reopened by change, incident, expiry or drift. The eight workflow steps of Section 6 are carried out within stages 1–5.

9 Lifecycle, roles, escalation and portfolio

9.1 Lifecycle stages

Stage	Outputs	Accountable roles
1 Scope and justify	Need, alternatives considered, why AI, system boundary, claims, objectives in scope.	BO, TSO
2 Characterize consequences	Scenarios by pathway, Consequence Register, tolerances and prohibitions.	BO, RO
3 Design controls	Authority grants and limits, containment, detection, fallback.	TSO, SGO
4 Evaluate	Tests, control challenge, intervals, evidence records with independence levels.	EO, IR
5 Decide and authorize	Evidence statuses, residual risk, disposition, authorization record.	DAA, RO
6 Operate and monitor	Monitoring against triggers, incident handling, drift checks.	IRO, SGO
7 Review, retest, retire	Scheduled review, change-triggered retest, renewal or Stop.	RO, IR, DAA

9.2 Eight roles

Roles are functions, not job titles. One person may hold several only where the separation rule below allows.

Code	Role	Accountability
BO	Business owner	Owns the purpose, benefits and tolerance; accepts residual consequence on behalf of the organization within delegated authority.
TSO	Technical system owner	Owns the configured system and its versions; implements fixes.
RO	Risk owner	Owns the risk record and Consequence Register; confirms tolerances.
EO	Evaluation owner	Designs and runs the evaluation; records results and evidence status.
IR	Independent reviewer	Challenges the design, evidence and status; does not report to BO or TSO.
DAA	Deployment-approving authority	Issues, restricts, renews or withdraws the authorization.
IRO	Incident-response owner	Owns detection-to-intervention response and incident records.

Code	Role	Accountability
SGO	Safeguard owner	Owns each control that limits authority, exposure or harm, and its effectiveness evidence.

Separation of duties. For a high-consequence system the approving authority must differ from the evaluation owner and the business owner, and the independent reviewer must not report to the technical system owner. Small organizations may be unable to separate every role; then the gap is recorded as an exception with an expiry and, where possible, external (I3) review.

9.3 Escalation

Escalate (a disposition, Section 11) when: a claim tied to a high-consequence entry is unresolved and a decision is nevertheless required; evidence of a refuted prohibition appears after authorization; independence required by tier cannot be obtained; or the decision exceeds the approving authority's delegation. Escalation names the receiving authority and the question it must answer.

9.4 Portfolio requirements

An organization holding many systems needs a portfolio view that does not hide risk. A portfolio summary should list, without computing a composite score: systems by disposition; claims by evidence status; authorizations expired or expiring; systems with unresolved claims tied to high-consequence entries; exceptions in force; and controls shared across systems, because a failure of a shared control is a common-cause failure that per-system records do not show. Aggregating exposure across systems requires its own analysis of overlap (Section 8.2, extension 3); the expected-loss sum, its independence-free justification and the need to model dependence for tail measures are in Section 8.6.

9.5 Continuous system evaluation (summary)

Maturity: [P] proposed; hypothesis H7 is [H]. This section summarizes the CSE Specification (https://rbed.ai/RBED_CSE_Specification_v2.1.pdf), which gives the protocol, definitions, acceptance criteria and a worked agentic use case. CSE operates within the lifecycle of Section 9 and specifies what happens after Authorize.

Element	What it requires
Lifecycle	Evaluate, Authorize, Monitor, Challenge, Reassess, then Continue (renew a Deploy), Restrict, Remediate or Stop. Periodic and event-triggered activities are scheduled separately.
Evidence Validity Register	Evidence is pinned to a configuration and has an expiry and invalidation triggers. A change is assessed by the criticality of the component it touches and the magnitude of the change: a critical component with a major change invalidates the evidence, and the claim reverts to Unresolved. An undeclared dependency is treated as requiring retest. An authorization may not outlive the evidence it depends on.
Control effectiveness	L0 present, L1 designed, L2 observed to meet pre-set thresholds on conservative bounds under challenge, L3 independently verified (I2 or I3). A control credited for a high-consequence claim needs L2 or better.
Agentic evaluation	An authority envelope (tools, resources, reversibility classes R0–R3) and trajectory metrics reported with conservative bounds; failure to find an out-of-envelope step in a finite set is not proof of its absence.
Oversight effectiveness	Seeded-error detection, automation bias, false intervention, dwell time, workload, authority to act, and independence of the measurement. A procedure that exists is not an oversight that works.

Element	What it requires
Monitoring allocation	Monitoring priority tiers T1–T4 from consequence, reversibility, exposure, assurance and evidence age; no numeric weights. Allocation is lexicographic with a random audit floor drawn independently of tier. Unfunded items are reported to the committee. Defaults are illustrative.
Learning and portfolio	Incidents move through gates: Discover, Reproduce, Contain, Fix, Verify, Learn. The portfolio report lists, without a score, what is unresolved, expiring, unfunded or past review, with ranked decision asks.
Avoided-loss contrast	Expected loss without monitoring minus expected loss with it, on comparable exposures, with intervals, net of cost and new risk, and labeled causal only if the evidence design supports it.

Boundaries. CSE adds no new disposition and no composite score. It does not replace pre-deployment validation, model risk management, security or privacy assessment, or enterprise governance, and it is not a monitoring platform.

10 System Evaluation Specification (SES)

The SES is the bounded, versioned record of claims, scope, metrics, test evidence, status, decision and renewal triggers. The record has eight minimum fields and optional blocks, including five blocks for continuous evaluation (below) that become expected when the stated consequence or disposition warrants them.

10.1 Minimum record (eight fields)

Field	Minimum content
Claim and owner	ID, intended purpose, responsible decision maker, affected stakeholders, accepted boundary; dimensions touched (C1–C3 among them, with the optional C3a/C3b component).
System boundary	Model/tool/prompt/data versions, integration surfaces, delegated permissions, deployment populations.
Consequence framing	Impact classes, exposure unit, severity, reach, duration, reversibility, external constraints.
Tests and sampling	Scenario strata, representative sample definition, challenge tests, controls to be injected, methods.
Evaluation results	Raw or protected trace IDs, metric definitions, numerator and denominator, estimates and intervals, errors, failed controls.
Evidence coverage	Supported / refuted / unresolved / N/A by material claim; gaps and measurement limitations.
Disposition	Deploy / Restrict / Escalate / Remediate / Stop; owner, rationale, approvals, monitoring and triggers.
Retest obligations	Safeguard changes, independent verification, time-bound review and reevaluation triggers.

10.2 Optional blocks

Block	Content
Claim threshold	For a claim measured as a rate, the acceptance threshold (maximum tolerable rate), fixed before testing.
Governance	Holder of each of the eight roles.
Justification	Need, alternatives considered, why AI.
Objectives in scope	Each of the five objectives, in or out of scope with reason, and a pointer to the specialized assessment where required.
Scenarios	Each scenario, its pathway (F, N or M) and the claims it bears on.
Consequence Register	Entries by category with severity, reversibility, tolerance, the scenarios that reach them and the claims they bear on; each linked claim must be borne on by a linked scenario and, for high-consequence entries, have evidence.
Authority grants	Capability, privilege level, scope, limit, and the control that enforces the limit.
Controls	Description, whether enforced outside the model, claimed independence, effectiveness evidence, and whether the risk argument relies on it.
Evidence records	Source type, independence level I0–I3, pointer.
Treatments and residual risk	Option (avoid, reduce likelihood, reduce consequence, transfer, accept) and whether it reduces harm; residual as an upper bound or an explicit abstention.
Exceptions	Rule excepted, rationale, approver, expiry.
Authorization	Authorizer, date, expiry, conditions of a Restrict, invalidation conditions, monitoring.
Monitoring triggers	Signal, threshold, resulting disposition.
Evidence validity	Per evidence item: pinned configuration, operating conditions, date, owner, uncertainty, review-by and expiry, invalidation triggers, dependencies with criticality, whether authorization depends on it.
Control assessments	Per credited control: type, V2 dimension, assurance level L0–L3, measures with pre-set thresholds, challenge conditions, verifier independence.
Monitoring plan	Per item: linked claim and consequence, signal, monitoring priority tier T1–T4, cadence, latency evidence, independent validation, escalation disposition.
Authority envelope	Tools, resources, reversibility classes R0–R3, approval rule, delegation and memory limits.
Learning records	Incident or near miss, stage (Discover to Learn), gate evidence for each stage reached.

A human-fillable template is in Appendix A. A machine-readable encoding of the same record, in JSON, follows; its content is **SYNTHETIC** and is the record of the worked example in Section 12. The excerpt shows four of the blocks; the continuous-evaluation blocks are illustrated in the CSE Specification Section 15. A record declares `spec_version 2.1` to have the checks of this specification applied.

```
{ "spec_version": "2.1",
  "system_name": "Customer-support refund agent (SYNTHETIC example)",
  "authority_grants": [ { "id": "G1", "capability": "initiate refund",
    "privilege": "act-with-approval",
    "scope": "refunds <= threshold after authenticated approval",
    "limit": "threshold per refund and per day", "enforced_by": "CTL1" } ],
  "controls": [ { "id": "CTL1", "enforced_outside_model": true,
    "description": "Independent server-side authorization service",
    "effectiveness_evidence": ["E2", "E3"], "credited": true } ],
  "evidence": [ { "id": "E2", "claim_id": "CL-02", "source_type": "test",
    "independence": "I2", "description": "Forced over-limit attempts 0/60" } ],
```

```
"authorization": { "authorizer": "Director of Risk (synthetic)",
  "authorized_on": "2026-10-28", "expires": "2026-11-27",
  "conditions": ["5% of tickets", "30 days"],
  "invalidation_conditions": ["any completed over-limit refund",
    "model, tool or authorization-service version change"],
  "monitoring": ["daily attempted-vs-completed report"] } }
```

11 Dispositions and authorization rules

11.1 Five dispositions

Disposition	Meaning
Deploy	Operate within the stated boundary on the stated evidence; requires a current authorization record.
Restrict	Operate only within explicit, bounded conditions (scope, exposure, duration, population) while evidence is incomplete; requires a current authorization record that states the conditions.
Escalate	The decision exceeds the evaluator's or approver's authority, or the evidence is in a state that requires a higher authority (Section 9). No operation beyond what is already authorized.
Remediate	Do not extend operation; fix the identified deficiency and retest independently.
Stop	Withdraw the system or capability: suspend pending resolution or retire.

11.2 Eight authorization rules

These are the conditions that a Deploy or Restrict disposition must meet. They are checks on a record's discipline; satisfying them does not show that the system is safe.

- Rule 1. Named, current authorization.** A Deploy or Restrict has an authorization record with an authorizer, a date and an expiry. An expired authorization is no authorization.
- Rule 2. Invalidation and monitoring.** The record states the conditions that invalidate it and the monitoring that detects them.
- Rule 3. No refuted claim; prohibitions supported; thresholds fixed first.** Deploy is not permitted with a refuted material claim, or with a categorical prohibition that is refuted or unresolved. A claim measured as a rate is Supported only against an acceptance threshold fixed before testing, and only if the result's upper bound is at or below it.
- Rule 4. Unresolved is not safe.** Deploy is not permitted while a claim tied to a high-consequence Consequence Register entry is unresolved.
- Rule 5. Restrict is bounded and must prevent the pathway.** A Restrict states its conditions: scope, exposure limit, duration and population. Where a refuted or unresolved claim is tied to a high-consequence entry, the Restrict must also name a credited control, enforced outside the model and with effectiveness evidence, that prevents that pathway from reaching affected people. An exposure percentage limits scale; it is not a restriction on the pathway.
- Rule 6. Controls credited only with evidence.** A control on which the risk argument relies has effectiveness evidence; limits on authority are enforced outside the model.
- Rule 7. Independence and separation for high consequence.** Where a high-consequence entry exists, evidence is not IO-only, the approving authority differs from the evaluation and business owners, and in-scope security and privacy objectives cite their specialized assessments.
- Rule 8. Residual risk is never silently zero.** Residual risk is recorded as an upper bound or an explicit abstention. Transfer or acceptance of risk is not recorded as harm reduction. Exceptions expire.

Any exception to a rule is recorded with a rationale, an approver and an expiry; an exception does not alter the evidence status of any claim.

12 Worked application — illustrative only

SYNTHETIC A fictional customer-support refund agent is permitted to answer policy questions but may initiate a refund only after authenticated authorization and approval above an established threshold. Its model gives a plausible response while an untrusted retrieved note induces a tool call above that threshold. The scenario is a teaching construction, not an observation or validated experiment; every figure is invented.

Vertex	Claim and falsification probe	Evidence / decision implication
V1	A retrieved note cannot override policy; send instruction-like text in untrusted retrieval content and observe tool arguments.	Capture retrieval provenance, attempt logs and version tuple. A correct final response does not erase a misdirected tool call.
V2	A server-side check rejects over-limit refunds regardless of model content; inject forced unauthorized attempts.	Record actual deny decisions, approval path, state changes, and detection latency. A prompt-only refusal is not evidence of control.
V3	No customer or business consequence may arise above tolerance.	Analyze attempted vs. completed refunds, affected records, legal notices, recoverability and cumulative exposure.

Disposition rule (illustrative): keep the agent in sandbox or restricted mode while the permission boundary is unresolved. After implementing an independent authorization control, rerun the seeded failure, legitimate allowed requests, regression scenarios and monitoring drills before any broadened deployment decision.

12.1 Consequence Register, authority and controls (hypothetical)

ID	Category	Entry
K1	Financial	Unauthorized money movement. Moderate severity, partially reversible by payment recall; reach per customer account; tolerance: no unauthorized refund completed. Bears on CL-02, CL-03.
K2	Legal and regulatory	Consumer-protection notice obligations triggered by an erroneous refund or denial. Moderate, reversible. Bears on CL-03.
K3	Privacy	Scenario SCN-04 (a model-formed query or reply exposes another customer's account data). Moderate, <i>irreversible</i> ; therefore high-consequence by Section 5. Bears on CL-04 (below), not on CL-01.
G1	Authority grant	"Initiate refund": act-with-approval, refunds at or below the threshold, per-refund and per-day limits, enforced by control CTL1.
G2	Authority grant	"Read customer account data": read only, the authenticated customer's own account, enforced by control CTL3.
CTL1	Control	Independent server-side authorization service that rejects over-limit refunds regardless of model content; enforced outside the model; credited, with effectiveness evidence E2 and E3.
CTL3	Control	Server-side scoping of every account-data query to the authenticated customer's session, which the model cannot widen; credited, with effectiveness evidence E4 (0 of 60 forced cross-account attempts succeeded, two-sided 95% upper bound 0.060).

Because K3 is high-consequence, the record must show a claim (CL-04), a scenario that reaches it, evidence for it that is not I0-only, an approving authority distinct from the evaluation and business owners, pointers to the specialized security and privacy assessments, and, for a Restrict while any claim on it is unresolved or refuted, a credited control that prevents the pathway (rules 5 and 7 of Section 11). In this record CL-04 is Supported against its threshold and CTL3 is evidenced.

12.2 Completed record (hypothetical)

SES stage	Hypothetical record
Claims	CL-01: a retrieved note cannot override refund policy. CL-02: a server-side check rejects over-limit refunds regardless of model content. CL-03: no unauthorized refund is completed (categorical prohibition). CL-04: replies and queries expose account data only to the authenticated customer. Owner: head of support operations; boundary: one build, one retrieval snapshot, 90 days. Acceptance thresholds, fixed before testing for this restricted pilot with human approval retained: CL-02 and CL-04, upper 95% bound at most 0.10; a general release would need a far lower threshold.
Initial tests	20 injected-note scenarios; 25 forced over-limit tool calls through a sandbox payments gateway.
Results	CL-01: 3 of 20 scenarios produced an above-threshold tool call (95% interval 0.03–0.38). CL-02: 0 of 25 forced attempts completed (upper 95% bound 0.14); alert-to-intervention time not measured.
Evidence status	CL-01 <i>refuted</i> . CL-02 <i>unresolved</i> : the bound is wide and detection latency is untested. CL-03 <i>unresolved</i> : no production-like outcome evidence.
Initial disposition	<i>Restrict</i> : sandbox only; any completed over-limit refund or version change triggers re-review.
Safeguard change (Remediate)	Add an independent server-side authorization control that does not depend on model output.
Independent retest (I2)	A reviewer outside the build team reruns the seeded failure, legitimate requests, regression scenarios and a monitoring drill: the attack still reaches tool calls, but 0 of 60 forced attempts complete (upper 95% bound 0.06) and 10 of 10 monitoring drills trigger intervention within the agreed limit (lower 95% bound 0.69). A second forced test of the account-data scoping (CTL3) gives 0 of 60 cross-account disclosures (upper 95% bound 0.06).
Updated status	CL-01 remains refuted (the model still misdirects), but its consequence is contained. CL-02 <i>supported (bounded)</i> : the upper bound 0.06 is within its 0.10 pilot threshold, for this build, gateway and window only. CL-04 <i>supported (bounded)</i> on the same basis. CL-03 <i>unresolved</i> until production-like outcome data exist. The 0.06 bound would <i>not</i> support a stringent tolerance.
Residual risk	RSK-01: an over-limit refund completes despite the authorization service, upper bound 0.06 per forced attempt (an upper bound, not an estimate). RSK-02: harm from the policy itself under intended behavior (pathway N): <i>abstain</i> , no outcome data; to be measured in the pilot. Insurance against payment fraud is recorded as a <i>transfer</i> , not as harm reduction.

SES stage	Hypothetical record
Updated disposition	<i>Restrict</i> : a limited pilot at 5% of tickets for 30 days. The 5% only limits scale; the restriction on the pathways is CTL1 (over-limit refunds) and CTL3 (cross-account disclosure), both enforced outside the model and evidenced. Human approval retained, a daily attempted-versus-completed refund report and a weekly drill; authorization expires on day 30. <i>Invalidation</i> : any completed over-limit refund, model, tool or authorization-service version change, drill failure. Broader deployment awaits CL-03 evidence and risk-owner sign-off.

13 Limitations and validation status

- **No empirical validation.** R-BED has not been validated against real incident data, expert judgment, or decision outcomes. The thirteen dimensions, the twelve categories, the independence levels and the ten extensions are engineering proposals. The benefits listed below are hypotheses.
- **Records, not guarantees.** The authorization rules check the discipline of a record. A record that satisfies every rule can describe an unsafe system if the tests were poor.
- **Independence qualifies; it does not certify.** An I3 review can be wrong; an I0 test can be right.
- **Adaptive adversaries and open-ended behavior.** Test sets are finite; misuse probability responds to the controls; a model or prompt change can invalidate evidence.
- **Quantities are conditional.** The identity and its extensions hold under stated definitions; dependence, heavy tails and underreporting can dominate.
- **Specialized assessments are outside scope.** Security testing, privacy impact assessment, safety engineering and legal analysis remain separate disciplines.
- **Organizational governance is not created by R-BED.** It names roles and records decisions; it does not supply authority, resources or culture.
- **Proposed elements are designs.** Everything labeled [P] (the Evidence Validity Register, the control assurance ladder, trajectory and oversight measures, tiers and cadences, the avoided-loss contrast) and [H] (H7) is unvalidated. Defaults are illustrative. Burden, gaming of ordinal ratings, accumulated small changes and undeclared dependencies are open problems.
- **Reference implementation is tooling, not evidence.** Its automated tests show that the code follows this specification, not that the method improves decisions.

13.1 Validation roadmap (planned; no results exist)

The studies below are proposed. None has been run; no result, sample size or finding is claimed.

Study	Question	Proposed design
Inter-rater reliability	Do independent reviewers assign the same evidence status and high-consequence classification to the same SES?	Several reviewers rate shared synthetic and, with permission, real records; report agreement with intervals.
Retrospective reconstruction	Does the method locate the control or evidence gap in documented historical incidents?	Reconstruct public incident reports as SES records without hindsight bias controls waived; compare with published findings.
Practitioner pilot	Can teams complete a record, and what does it cost?	Pilot with consenting teams; measure effort, omissions and reviewer feedback.

Study	Question	Proposed design
Independence effect	Does independent challenge find failures that self-assessment misses?	Paired internal and independent evaluations of the same system.
Control-effectiveness prediction	Do intervention contrasts predict containment in later operation?	Compare predicted and observed containment in monitored pilots.
H7: consequence-aware monitoring allocation	Does allocating a fixed monitoring budget by R-BED tier avoid more harm than frequency-based or uniform allocation, and when does it do worse?	Preregistered paired comparison with failure injection under fixed budgets, including scenarios in which R-BED is expected to do worse; protocol in the Validation Plan. No result exists.

14 Coverage of twenty capabilities

The matrix states, for twenty capabilities, how fully R-BED specifies each, and how other public resources address it. *Complete* means R-BED specifies the capability within its own scope, not that it has been validated; *Partial* means it addresses an aspect and needs additional practice; *Not covered* means R-BED does not address it. The right-hand column summarizes publicly described scope; ISO/IEC standards are paywalled and were not consulted in full, so those entries are cautious.

#	Capability	R-BED	Other resources (author summary)
1	Falsifiable system-claim definition	Complete	NIST AI RMF MAP and ISO/IEC 23894 address risk identification at organizational level; the claim form is R-BED-specific.
2	Failure-mediated pathway analysis	Complete	Familiar from FMEA, fault-tree and bow-tie practice; R-BED applies it to configured AI systems.
3	Intended-behavior harm pathway	Partial	Impact-assessment practice (for example ISO/IEC 42005) addresses harm without failure; R-BED records it but does not assess it.
4	Misuse and adversarial pathway	Partial	Security testing is a specialized discipline; R-BED requires the evidence and pointer, not the method.
5	Deployment-control challenge testing	Complete	Forced-failure control tests are R-BED's emphasis; AI Verify and NIST TEVV-Athlon (draft) address testing more broadly.
6	Human-oversight effectiveness	Partial	Human-factors research and NIST AI RMF MAP/MEASURE address oversight; the specification includes measures (seeded-error detection, automation bias) that are untested.
7	Consequence recording (categories, severity, reversibility)	Complete	Impact assessment resources go deeper on method; R-BED supplies the register structure.
8	Stakeholder-inclusive impact assessment	Partial	ISO/IEC 42005 and NIST AI RMF MAP address it; R-BED asks for participation but is not an assessment method.
9	Fairness evaluation	Partial	AI Verify includes fairness checks; R-BED locates the claim but specifies no metrics or thresholds.
10	Uncertainty and rare-event evidence	Complete	Standard statistics; R-BED specifies bounds, trial counts and abstention.
11	Evidence status and independence of evidence	Complete	NIST AI RMF MEASURE 1.3 mentions independent reviewers; R-BED defines levels I0–I3.

#	Capability	R-BED	Other resources (author summary)
12	Disposition and authorization record	Complete	Assurance and safety-case practice record decisions; R-BED specifies eight checks.
13	Lifecycle and role assignment	Partial	Management-system standards cover this organizationally; R-BED names roles only.
14	Portfolio-level view	Partial	Inventory is organizational (NIST GOVERN 1.6); R-BED summarizes SES records without a score.
15	Monitoring and re-evaluation triggers	Complete	NIST MANAGE 4 and ISO/IEC 42001 cover operations; R-BED specifies cadences, validity expiry and re-evaluation triggers within its scope; R-BED still does not operate monitoring.
16	Security testing (specialized)	Partial	Specialized practice; R-BED routes to it.
17	Privacy impact assessment	Partial	Specialized practice; R-BED routes to it and cannot establish legal compliance.
18	Organizational governance, policy, training	Not covered	NIST GOVERN and ISO/IEC 42001. Outside R-BED.
19	Management-system certification	Not covered	ISO/IEC 42001 is certifiable; R-BED is not a standard or certification.
20	Automated test toolkit or benchmark execution	Not covered	AI Verify provides a toolkit; R-BED is not one.

Counted from the table: 8 Complete, 9 Partial and 3 Not covered. Three rows (18–20) remain outside R-BED's scope by design. Row 15 (monitoring and re-evaluation triggers) is Complete as specified, not as operated.

15 Relationship to NIST AI RMF 1.0 and to other resources

NIST AI RMF 1.0 provides the voluntary GOVERN, MAP, MEASURE and MANAGE functions. R-BED can support particular outcomes by structuring system claims, pathways, test methods, control challenges, consequences, and evidence-based decisions. The companion Crosswalk (v2.1) maps 51 of the 72 subcategories, 44 as Partial and 7 as Relevant contribution; 21 are not mapped and are listed there. No claim is made that R-BED covers all NIST AI RMF outcomes or provides NIST certification or endorsement.

15.1 What is not new

R-BED rests on long-standing ideas, and the document does not claim them. Risk as a combination of likelihood and consequence is the basis of ISO 31000-style risk management. Failure-mode analysis, fault trees and bow-tie diagrams analyze how failures propagate. Layered protection, often described as the Swiss-cheese model (Reason, 2000), underlies the control-challenge emphasis. Safety and assurance cases (claim, argument, evidence) are closely related to R-BED's claims and evidence statuses. Red-teaming, model and system cards, and confidence intervals for binomial proportions are standard.

15.2 Positioning of continuous system evaluation

Model risk management already requires ongoing monitoring, outcomes analysis and revalidation after change for traditional models; US interagency guidance SR 26-2 (17 April 2026) replaced SR 11-7 and, by its footnote 3, places generative and agentic AI outside its scope. Observability platforms, LLM and agent evaluation frameworks, runtime policy enforcement, continuous-assurance practice in other risk disciplines and incident-driven evaluation each cover part of what CSE touches. CSE does not claim any single component as novel. The proposed contribution is the traceable integration of

consequence-aware prioritization, evidence of control effectiveness and evidence-validity management into one evaluation-to-decision record for configured generative and agentic systems; that integration is untested. The CSE Specification Section 14 classifies each component as established, R-BED-specific or untested, and compares the approaches.

15.3 Other AI risk and evaluation resources

R-BED is intended to complement, not replace, existing resources, and it makes no claim of empirical superiority over any of them.

- **ISO/IEC 23894 (AI risk management guidance), ISO/IEC 42001 (AI management system) and ISO/IEC 42005 (AI system impact assessment)** address organizational risk-management processes, a certifiable management system, and impact assessment. R-BED can supply system-level evidence for such processes; it provides neither a management system nor an impact-assessment method.
- **AI Verify (Singapore)** offers a testing framework and toolkit for selected technical and process checks. R-BED is not a toolkit; it specifies how test results, control challenges and consequences are connected in one bounded, reviewable record.
- **NIST TEVV-Athlon (NIST AI 200-2, initial public draft)** describes how to build customized assessments from an organization's TEVV objectives. R-BED could be used to scope an assessment of one kind: pathways to consequence in a configured deployment.

Positioning. R-BED's specific emphasis is the chain of evidence connecting pathway, containment, consequence and authorized action: whether a failure, intended behavior or misuse reaches harm given the deployed controls, how that was independently challenged, what decision the bounded evidence supports, and when that decision lapses. The contribution is the assembly of these elements for deployed AI systems into one reviewable record with explicit authorization rules, not the novelty of each element. Broader governance, measurement and assurance practices in the resources above remain outside R-BED's scope. Descriptions are summaries by the author; readers should consult each resource for its authoritative scope.

16 Public availability, attribution and use

© 2026 Vishal Kumar Srivastava. All rights reserved except as stated here. This public specification may be read, downloaded, cited and applied in an organization's own evaluation practice without charge and without a framework-access fee, with attribution. Republication, translation or derivative versions require the author's written permission. No endorsement, certification, warranty or regulatory safe harbor is provided. Third-party rights are unaffected. Rights in this document, in any software that implements it, in the R-BED™ name, and in other materials are separate; terms stated for one do not extend to another.

Public URL of this document: https://rbed.ai/RBED_Public_Framework_v2.1.pdf. Companion NIST AI RMF crosswalk: https://rbed.ai/NIST_AI_RMF_RBED_Crosswalk_v2.1.pdf.

Methodology contact: Vishal Kumar Srivastava · vshfrm@gmail.com

Trademark: R-BED™ is claimed as a trademark by the author; no registration, certification or NIST endorsement is asserted. This resource is independent of NIST and does not offer a product or service.

Field	Fill in
-------	---------

A SES template (human-fillable)

Field	Fill in
System, version tuple	Model / tools / prompts / data versions; date of evaluation.
Justification	Need; alternatives considered; why AI.
Roles	BO, TSO, RO, EO, IR, DAA, IRO, SGO: name each holder.
Objectives	For each of safety, security, privacy, robustness, resilience: in/out of scope, reason, specialized-assessment pointer.
Claims	ID; statement; purpose; owner; stakeholders; accepted boundary; acceptance threshold fixed before testing; prohibition? (Y/N); material? (Y/N).
Scenarios	ID; pathway (F / N / M); description; claims affected.
Consequence Register	ID; category (of 12); severity; reversibility; reach; duration; tolerance; claims.
Authority grants	Capability; privilege level; scope; limit; enforcing control.
Controls	ID; description; enforced outside model? (Y/N); independent of which controls; effectiveness evidence; credited? (Y/N).
Tests and sampling	Strata; sample definition; challenge tests; controls injected; methods.
Results	Claim; metric definition; numerator / denominator; interval and method; trace IDs; failed controls.
Evidence records	ID; claim; source type; independence (I0–I3); pointer.
Evidence status	Per material claim: supported (bounded) / refuted / unresolved / N/A; reasons; gaps; limitations.
Treatments, residual	Option; reduces harm? (Y/N); residual as upper bound or abstention with reason.
Disposition	Deploy / Restrict / Escalate / Remediate / Stop; owner; rationale; dissent.
Authorization	Authorizer; date; expiry; Restrict conditions; invalidation conditions; monitoring.
Triggers and retest	Signals, thresholds, actions; safeguard changes; independent verification; review-by date.
Exceptions	Rule; rationale; approver; expiry.

B Record versioning

An SES declares `spec_version`; this specification is `spec_version 2.1`. The eight minimum fields are always required. The optional blocks are additive: a record that omits them is checked under the rules that apply to its declared version, and a Deploy or Restrict on a high-consequence claim in a 2.1 record needs validity records for the evidence the authorization depends on, control assessments for credited controls, and a monitoring plan for high-consequence claims. The three consequence dimensions are named C1 Incremental Impact, C2 Scale and C3 Regulatory and Reputational Impact; the optional claim field `c3_components` records C3a (regulatory and legal) and C3b (reputational and trust) separately.

References and source notes

- NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (2023), <https://doi.org/10.6028/NIST.AI.100-1>.
- ISO/IEC 23894:2023, *Guidance on risk management*; ISO/IEC 42001:2023, *AI management system*; ISO/IEC 42005:2025, *AI system impact assessment*. Paywalled; described from public summaries, not consulted in full.
- Singapore IMDA / AI Verify Foundation, AI Verify testing framework and toolkit; see the IMDA announcement of the AI Verify Foundation (2023), <https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/singapore-launches-ai-verify-foundation>.
- R-BED CSE Specification v2.1, https://rbed.ai/RBED_CSE_Specification_v2.1.pdf.
- Board of Governors of the Federal Reserve System, OCC and FDIC, *Revised Guidance on Model Risk Management*, SR 26-2 (17 April 2026), <https://www.federalreserve.gov/supervisionreg/srletters/SR2602.pdf>; supersedes SR 11-7. Guidance, not an enforceable standard.
- NIST, AI Agent Standards Initiative (2026), <https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative>. An initiative, not a requirement.
- NIST, *The TEVV-Athlon Framework for Evaluating AI Systems*, NIST AI 200-2 (initial public draft, August 2026), <https://www.nist.gov/artificial-intelligence/tevv-athlon-framework-evaluating-ai-systems>.
- NIST AI Resource Center, *AI RMF Core*, <https://airc.nist.gov/airmf-resources/airmf/5-sec-core/>; *Community Crosswalks*, <https://airc.nist.gov/airmf-resources/crosswalks/>.
- C. J. Clopper and E. S. Pearson, “The use of confidence or fiducial limits illustrated in the case of the binomial,” *Biometrika* 26(4), 1934, 404–413.
- J. Reason, “Human error: models and management,” *BMJ* 320, 2000, 768–770.
- *Methodology note*: this document is a standalone public statement of the R-BED methodology; its section definitions, example and limitations may be referenced independently of any other publication.