

R-BED™ | Risk-Based Evaluability Design

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

Version 2.0 (supersedes v1.1) | 28 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

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 2.0 keeps that architecture unchanged and adds the layers an adopting organization needs to act on the evidence: five cross-cutting objectives (safety, security, privacy, robustness, resilience) applied as lenses across the dimensions; sharper definitions of the four deployment controls; a Consequence Register; pathways for intended-behavior harm and misuse alongside failure-mediated harm; additional quantitative tools with explicit limits; a seven-stage lifecycle with eight roles and levels of evidence independence; an extended System Evaluation Specification; and five dispositions governed by eight authorization rules.

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.

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).

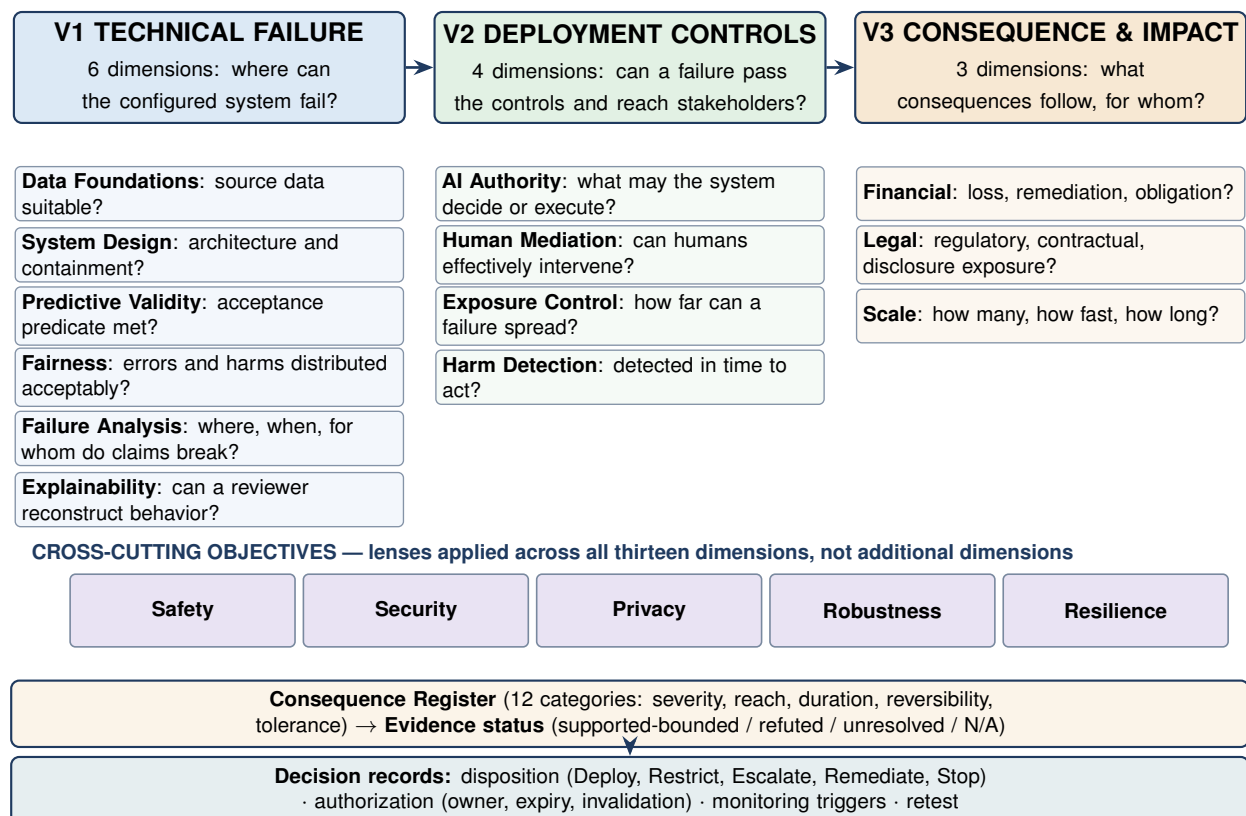


Figure 1. R-BED v2.0 architecture: thirteen diagnostic dimensions in three connected vertices (6 + 4 + 3), five cross-cutting objectives applied across them, and the Consequence Register and decision records that connect evidence to action. Arrows show a simplified failure-mediated pathway; not every harmful outcome requires a technical failure. Dimensions are locations for claims and evidence, not scores, and evidence from one vertex cannot substitute for another.

- **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.

What changed from v1.1

Area	Status	Summary
Architecture (6+4+3, 13 dimensions)	Unchanged	Dimension names and counts unchanged; V3 renamed “Consequence and Impact” (label only).
Deployment Controls	Changed	Four definitions strengthened (Section 3).
Cross-cutting objectives	New	Safety, security, privacy, robustness, resilience as lenses, with matrix (Section 4).
Consequence Register	New	Twelve categories (Section 5).
Mathematics	Changed	v1.1 identity preserved; ten extensions with stated assumptions (Section 8).
Lifecycle, roles, portfolio	New	Seven stages, eight roles, independence levels I0–I3 (Sections 7, 9).
SES	Changed	Eight fields kept; optional v2.0 blocks added (Section 10).
Dispositions	Changed	Deploy, Restrict, Escalate, Remediate, Stop; eight authorization rules (Section 11).
Coverage matrix, prior art	New	Twenty capabilities with honest status (Sections 14, 15).

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. 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.

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.
Implemented	A reference implementation of the record format and checks exists. It is optional tooling, separate from this specification.
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.
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.
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).
Authorization record	A named, dated, expiring record that a disposition of Deploy or Restrict was approved on stated evidence, with invalidation conditions and monitoring.

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 (6+4+3) is unchanged from v1.1.

3.1 Vertex 1 — Technical Failure (six dimensions)

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

3.2 Vertex 2 — Deployment Controls (four dimensions, strengthened definitions)

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
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.
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.

Control	Definition (what must hold)	Evidence that shows it	Does not count as evidence
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.
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 Vertex 3 — Consequence and Impact (three dimensions)

In v1.1 this vertex was labeled “Business Impact”. The three dimensions are unchanged; the label now states that the question includes consequences to people and society, recorded in the Consequence Register (Section 5), not only to the organization.

Dimension	Question / example of relevant evidence
Financial	What loss, remediation burden, or financial obligation can result? Scenario analysis, portfolio concentration, and recoverability.
Legal	What legal, regulatory, contractual, or disclosure exposure can result? Applicable obligation mapping, decision trail, and notices/rights.
Scale	How many decisions, people or systems can be affected, how quickly, and for how long? Volume, concurrency, propagation and reach analyses.

Consequence attributes are documented separately in the Consequence Register. They do not create new R-BED dimensions. A purely monetary representation is insufficient for unpriced harms.

Objective	Question the lens asks	Typical evidence	Specialized assessment (outside R-BED)
-----------	------------------------	------------------	--

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
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
Data Foundations	V1		○	●	●	
System Design	V1		●	○	○	○
Predictive Validity	V1	○			●	
Fairness	V1				●	
Failure Analysis	V1	●	○	○	●	○
Explainability	V1		○	●	○	○
AI Authority	V2	●	●	○		○
Human Mediation	V2	○	○	○		●
Exposure Control	V2	●	●	●	○	●
Harm Detection	V2	●	●	○	○	●
Financial	V3	○	○			●
Legal	V3	○	●	●		○
Scale	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 replaces the single “business impact” question of v1.1 with twelve categories, each assessed on the same attributes.

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 (unchanged from v1.1)

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.
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. A result such as 0 of 60 (upper bound 0.060) 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.

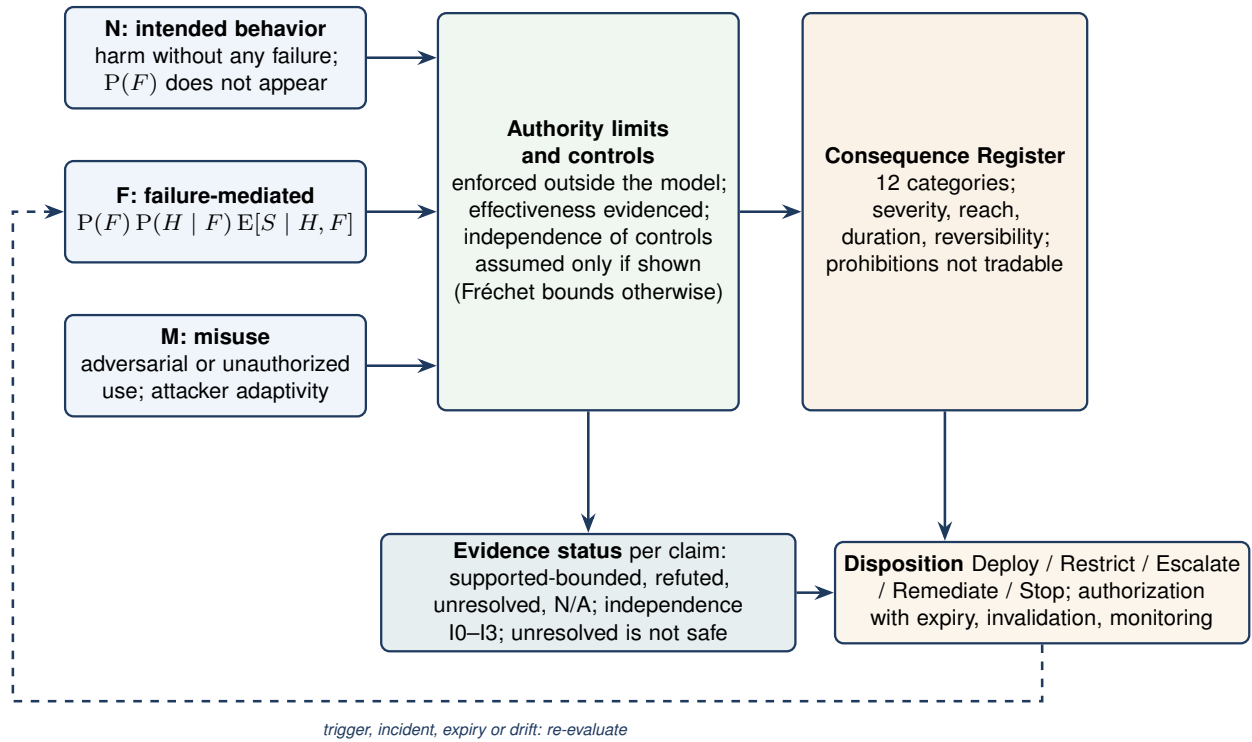


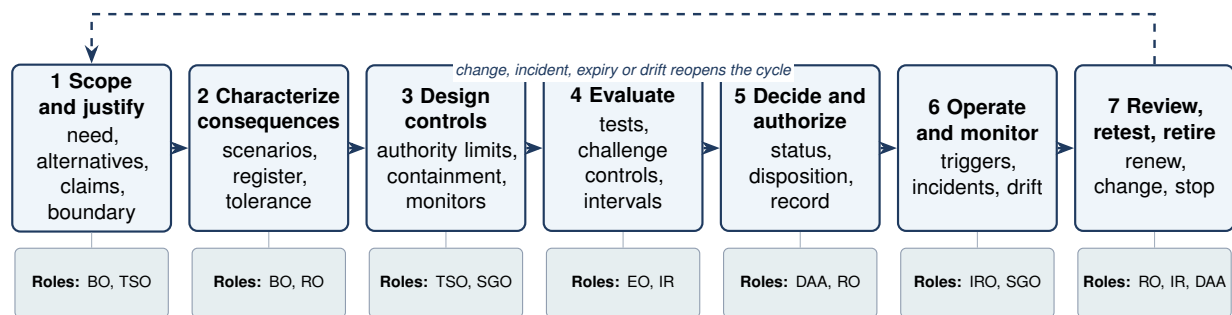
Figure 2. Pathways to consequence and to decision. Intended-behavior (N), failure-mediated (F) and misuse (M) pathways pass through authority limits and controls to the Consequence Register; evidence status and independence feed a disposition and its authorization record, which is reopened by triggers, incidents, expiry or drift.

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: A1 no independent containment, A2 an independent server-side limit check, A3 limit check plus reduced authority.

Config.	Description	$P(F A)$	$P(H F, A)$	$E[S H, F, A]$	$E[L_F A]$
A1	No independent containment	0.04	0.80	60	1.92
A2	Server-side limit check	0.04	0.01	70	0.028
A3	Limit check, reduced authority	0.03	0.005	60	0.009

The failure rate is unchanged between A1 and A2: the containment acts on $P(H | F)$, not on $P(F)$. Severity is higher in A2 because the harms that remain are the uncontained larger ones. Three cautions apply. The ordering $A1 > A2 > A3$ 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.



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 3. 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).

10 System Evaluation Specification (SES)

The SES is the bounded, versioned record of claims, scope, metrics, test evidence, status, decision and renewal triggers. The eight minimum fields of v1.1 are unchanged; v2.0 adds optional blocks that become expected when the stated consequence or disposition warrants them.

10.1 Minimum record (eight fields, unchanged)

Field	Minimum content
Claim and owner	ID, intended purpose, responsible decision maker, affected stakeholders, accepted boundary.
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.

Field	Minimum content
Retest obligations	Safeguard changes, independent verification, time-bound review and reevaluation triggers.

10.2 Optional v2.0 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.

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.

```
{ "spec_version": "2.0",
  "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": "C2", "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. (v1.1 "Re-work".)
Stop	Withdraw the system or capability: suspend pending resolution or retire. (v1.1 "Suspend" and "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 C2, C3.
K2	Legal and regulatory	Consumer-protection notice obligations triggered by an erroneous refund or denial. Moderate, reversible. Bears on C3.
K3	Privacy	Scenario S4 (a model-formed query or reply exposes another customer's account data). Moderate, <i>irreversible</i> ; therefore high-consequence by Section 5. Bears on C4 (below), not on C1.
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, upper bound 0.060).

Because K3 is high-consequence, the record must show a claim (C4), a scenario that reaches it, evidence for it that is not IO-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 C4 is Supported against its threshold and CTL3 is evidenced.

12.2 Completed record (hypothetical)

SES stage	Hypothetical record
Claims	C1: a retrieved note cannot override refund policy. C2: a server-side check rejects over-limit refunds regardless of model content. C3: no unauthorized refund is completed (categorical prohibition). C4: 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: C2 and C4, 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	C1: 3 of 20 scenarios produced an above-threshold tool call (95% interval 0.03–0.38). C2: 0 of 25 forced attempts completed (upper 95% bound 0.14); alert-to-intervention time not measured.
Evidence status	C1 <i>refuted</i> . C2 <i>unresolved</i> : the bound is wide and detection latency is untested. C3 <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	C1 remains refuted (the model still misdirects), but its consequence is contained. C2 <i>supported (bounded)</i> : the upper bound 0.06 is within its 0.10 pilot threshold, for this build, gateway and window only. C4 <i>supported (bounded)</i> on the same basis. C3 <i>unresolved</i> until production-like outcome data exist. The 0.06 bound would <i>not</i> support a stringent tolerance.
Residual risk	R1: an over-limit refund completes despite the authorization service, upper bound 0.06 per forced attempt (an upper bound, not an estimate). R2: 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.
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. Invalidation: any completed over-limit refund, model, tool or authorization-service version change, drill failure. Broader deployment awaits C3 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.
- **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.
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.

14 Coverage of twenty capabilities

The matrix states, for twenty capabilities, how fully R-BED specifies each in v1.1 and v2.0, 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	v1.1	v2.0	Other resources (author summary)
1	Falsifiable system-claim definition	Complete	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	Complete	Familiar from FMEA, fault-tree and bow-tie practice; R-BED applies it to configured AI systems.
3	Intended-behavior harm pathway	Not covered	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	Not covered	Partial	Security testing is a specialized discipline; R-BED requires the evidence and pointer, not the method.
5	Deployment-control challenge testing	Complete	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	Partial	Human-factors research and NIST AI RMF MAP/MEASURE address oversight; v2.0 sharpens the definition only.
7	Consequence recording (categories, severity, reversibility)	Partial	Complete	Impact assessment resources go deeper on method; R-BED supplies the register structure.
8	Stakeholder-inclusive impact assessment	Partial	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	Partial	AI Verify includes fairness checks; R-BED locates the claim but specifies no metrics or thresholds.
10	Uncertainty and rare-event evidence	Partial	Complete	Standard statistics; v2.0 adds bounds, trial counts and abstention.
11	Evidence status and independence of evidence	Partial	Complete	NIST AI RMF MEASURE 1.3 mentions independent reviewers; R-BED defines levels I0–I3.
12	Disposition and authorization record	Partial	Complete	Assurance and safety-case practice record decisions; R-BED specifies eight checks.
13	Lifecycle and role assignment	Not covered	Partial	Management-system standards cover this organizationally; R-BED names roles only.
14	Portfolio-level view	Not covered	Partial	Inventory is organizational (NIST GOVERN 1.6); R-BED summarizes SES records without a score.
15	Monitoring and re-evaluation triggers	Partial	Partial	NIST MANAGE 4 and ISO/IEC 42001 cover operations; R-BED records triggers, not operations.
16	Security testing (specialized)	Not covered	Partial	Specialized practice; R-BED routes to it.
17	Privacy impact assessment	Not covered	Partial	Specialized practice; R-BED routes to it and cannot establish legal compliance.
18	Organizational governance, policy, training	Not covered	Not covered	NIST GOVERN and ISO/IEC 42001. Outside R-BED.
19	Management-system certification	Not covered	Not covered	ISO/IEC 42001 is certifiable; R-BED is not a standard or certification.
20	Automated test toolkit or benchmark execution	Not covered	Not covered	AI Verify provides a toolkit; R-BED is not one.

Counted from the table: v1.1 has 3 Complete, 8 Partial and 9 Not covered; v2.0 has 7 Complete, 10 Partial and 3 Not covered. Six rows moved from Not covered to Partial (rows 3, 4, 13, 14, 16, 17, where R-BED now records or routes the capability without performing it) and four from Partial to Complete (rows 7, 10, 11, 12). No row got worse, and three rows (18–20) remain outside R-BED’s scope by design.

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 (v2.0) 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 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_v2.0.pdf. Companion NIST AI RMF crosswalk: https://rbed.ai/NIST_AI_RMF_RBED_Crosswalk_v2.0.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.

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 Migration from v1.1

v1.1	v2.0
Vertex “V3 Business Impact”	“V3 Consequence and Impact”; same three dimensions.
Disposition “Rework”	Remediate.
Disposition “Suspend” or “Retire”	Stop.
Deploy, Restrict	Unchanged in name; now require an authorization record (rules 1–2, 5).
Eight SES fields (none)	Unchanged. Optional blocks (Section 10). A v1.1 record migrates by relabeling; a Deploy or Restrict migrated from v1.1 is incomplete under v2.0 until an authorization record is added.

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>.
- 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.