

Beyond Pre-Deployment Validation

Toward Continuous, Consequence-Aware Evaluation of AI Systems

Vishal Kumar Srivastava, PhD | 29 October 2026 | Companion to R-BED™ v2.1

Status	A position paper. It proposes a method and a research agenda. It reports no empirical result and no data; the numerical examples are arithmetic on invented parameters. The method is not a standard, a certification or an endorsed practice.
---------------	--

Abstract

Evaluation of AI systems is often framed as a gate before deployment. Three facts argue against treating it only that way: the conclusions of an evaluation depend on a configuration and conditions that change; the safeguards that stand between a failure and a harm may fail silently; and monitoring effort is finite and must be directed. This paper states the case for *continuous, consequence-aware evaluation*, in which evaluation evidence has an owner, a configuration, an expiry and invalidation triggers; safeguards are shown to work under challenge; and monitoring effort is allocated by the consequence of failure rather than by its frequency. It describes a concrete proposal (R-BED v2.1 and its CSE Specification), classifies each component as a specified core element, proposed or hypothetical, states one testable hypothesis (H7) together with the conditions under which the proposal is expected to do worse than simpler alternatives, and lists what would have to be shown before the claim is supported. Continuous assurance is not continuous testing: the proposal concerns what should be rechecked, when and by whom, not that everything be rechecked always.

1 Problem

1. Evidence expires. An evaluation result is a statement about a model, prompt, tools, data, permissions and operating conditions at a time. A change to any of them can make the statement false without any test failing, because no test was rerun. Practices that ask for ongoing monitoring and revalidation after change exist for traditional models; whether and how they carry over to configured generative and agentic systems, where a prompt edit or a tool update is a material change, is not settled.

2. Safeguards are claims. A deployment control (an authorization limit, a human approval step, a filter) is credited in a risk argument. A documented control is not a working control. Whether it works under realistic failure is an empirical question that is rarely recorded in a form a decision can depend on.

3. Attention is allocated. Monitoring, independent validation and incident response are finite. If effort follows what is observed most often, it follows the frequent and visible; a rare, severe, irreversible consequence may receive the least. Whether allocating by consequence does better is an empirical question, and there are settings in which it would not.

2 Related practice (what this paper does not claim)

The proposal builds on, and does not replace, model risk management (validation, ongoing monitoring, outcomes analysis; in the United States the interagency guidance SR 26-2 of 17 April 2026 replaced SR 11-7 and, in its footnote 3, places generative and agentic AI outside its scope), the NIST AI Risk Management Framework 1.0 and its subsequent materials, ISO/IEC 42001 and 23894, observability platforms, LLM and agent evaluation frameworks, runtime policy enforcement, continuous-assurance

methods from other risk disciplines, and incident-driven evaluation. Each addresses part of the problem. The claim is not that any of them is deficient but that an explicit link from consequence to control evidence to the lifetime of a decision is not, to the author's knowledge, specified in one traceable record for configured generative and agentic systems. That belief is itself unverified; the author has not performed a systematic review.

3 Proposal

The proposal has three parts, each of which depends on the others.

- P1.** *Consequence-aware evaluation*[E]: a system claim is evaluated against failure pathways, the controls that may stop them and the consequences that follow, kept as separate questions, under a decision record with five dispositions and eight authorization rules.
- P2.** *Evidence-based control assurance*[P]: an assurance ladder L0 (present) to L3 (independently verified), with a control credited for a high-consequence claim only at L2 or above, where L2 means that measures met pre-set thresholds on conservative bounds under challenge. For agents, an authority envelope and trajectory metrics; for human oversight, seeded-error detection and automation-bias measures.
- P3.** *Continuous evaluation and authorization*[P]: an Evidence Validity Register pins each item of evidence to a configuration with an expiry and invalidation triggers; a change is assessed against declared dependencies, with criticality and magnitude of change determining no effect, review, retest or invalid; an invalid item reverts its claim to Unresolved; an authorization cannot outlive the evidence it depends on. Monitoring items receive monitoring priority tiers T1–T4 by rules, funded in a fixed order with a random audit floor.

Maturity labels: [E] specified core (specified, with reference code that passes its own tests), [P] proposed extension, [H] hypothesis. No [P] or [H] element has been empirically validated.

3.1 Quantitative frame

Let F be a failure, H harm and S its severity. The earlier framework's identity $E[L | A] = P(F | A)P(H | F, A)E[S | H, F, A]$ is retained. A monitoring-and-intervention policy π acts through detection (d), timely intervention (t) and containment (c), each conditional on the preceding step, so that the probability of an avoided harm contains the factor $d t c$ by the chain rule. The expected avoided loss of policy π_1 relative to π_0 is $E[L | A, \pi_0] - E[L | A, \pi_1]$, defined only for comparable exposures and horizons. Detection does not imply containment. Whether the result is described as causal depends on the design that produced the inputs: a designed experiment, with a concurrent control or injected failures, may support a causal statement for its conditions; an observational comparison supports only an association, because monitoring is applied where risk is thought to be high and increases the number of incidents detected. Net value must subtract monitoring cost, the cost of false interventions and any risk the monitoring adds; under invented but plausible parameters the net interval can include negative values, which is the correct reading of "not shown to pay for itself".

A zero-failure result bounds a rate; it does not show zero risk. With 60 trials and no failure, the two-sided 95% Clopper–Pearson upper bound on the failure rate is about 0.060 (one-sided, 0.049), which meets a 10% threshold and, on the two-sided bound, does not meet a 5% one; meeting a 1% threshold needs about 368 failure-free trials at two-sided 95% (299 one-sided). The trial counts are properties of the binomial, not findings.

4 Hypothesis H7 and how it could fail

H7. Under a fixed budget of monitoring and independent-validation effort, allocating by R-BED's ordinal consequence-aware tiers avoids more expected harm than allocation by observed failure frequency or by uniform allocation, in settings where severity and observed frequency diverge and where consequence classification is reasonably reliable; and R-BED does no better, or worse, in specified other settings.

The hypothesis is deliberately stated so that it can fail, and the Validation Plan names two situations in which the proposal is expected to do worse: (i) when the consequence class assigned by the user is wrong for a material share of pathways, because the method then concentrates effort on the wrong pathways, and (ii) when harm is dominated by frequent, low-severity events, because a frequency-based allocation targets precisely those. A third family of settings (stale evidence, an adaptive adversary, a budget too small for the highest tier) is uncertain. The primary design is controlled failure injection, with the same injected set faced by every allocation policy and severity assigned by an independent panel before injection; a randomized allocation in a real portfolio is the only design that would support a causal statement about real harm; and a permuted-tier negative control must show no benefit. A reference implementation of the analysis pipeline on synthetic data exists so that the code and the preregistered decision rule can be checked before real data. Its output depends only on invented parameters and is not evidence.

5 Threats and limits

- *Ordinal ratings can be wrong or gamed.* Tiers follow self-assessed ratings. The random audit floor, the evidence-age rule and independent verification for the top tiers reduce but do not remove this.
- *Declared dependencies.* Proportional invalidation depends on the owner's declared dependencies; undeclared dependencies are handled conservatively (retest), at a cost whose size is unknown.
- *Seeded failures may be easier than real ones,* so measured detection may be optimistic; harm may be delayed or unobserved.
- *Burden.* Whether organizations can maintain the registers at acceptable cost is a feasibility question.
- *Accumulated small changes and adaptive adversaries* are not modeled.
- *Defaults are not findings.* Cadences and thresholds in the specification are illustrative.
- *No comparison with other practices* has been made; the proposal makes no claim of superiority.

6 Research agenda

In order of dependency: (1) agreement among reviewers on consequence classes and tiers (a precondition for H7); (2) burden and feasibility pilots of the Evidence Validity Register; (3) failure-injection studies of control effectiveness and of oversight measures; (4) the H7 allocation study with the negative scenarios; (5) a randomized allocation in a real portfolio; (6) a designed comparison with other governance practices by teams that did not design R-BED; (7) replication. Each ends with a decision to retain, revise or remove the part tested, and null and negative results are reported with equal prominence.

References and source notes

- R-BED Public Framework v2.1, https://rbed.ai/RBED_Public_Framework_v2.1.pdf; R-BED CSE Specification v2.1, https://rbed.ai/RBED_CSE_Specification_v2.1.pdf; R-BED Validation Plan v2.1 (same author).
- NIST, *AI Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (2023), <https://doi.org/10.6028/NIST.AI.100-1>. NIST states that the framework is under revision; this paper refers to version 1.0.

- 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>. Guidance; not an enforceable standard.
- ISO/IEC 23894:2023; ISO/IEC 42001:2023 (paywalled; public summaries only).
- C. J. Clopper and E. S. Pearson, *Biometrika* 26(4), 1934, 404–413.

© 2026 Vishal Kumar Srivastava. All rights reserved except as stated in the R-BED Public Framework: this paper may be read, downloaded, cited and applied without charge, with attribution; republication or derivative versions require written permission. R-BED™ is claimed as a trademark by the author; no registered-mark status and no NIST endorsement is asserted. **Contact:** vshfrm@gmail.com