International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 7, Issue 4, PP 44-53, 2026
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V7I4P105

Original Article

Human-in-the-Loop Approval Gates for Autonomous Enterprise Actions: Designing When Autonomy Should Pause for Human Judgment

Bharadwaj Kasinadhuni
Platform Engineering, Enterprise Automation, Independent Researcher, USA.

Received On: 26/08/2026 ***Revised On: 24/09/2026*** ***Accepted On: 01/10/2026*** ***Published On: 06/10/2026***

Abstract: *Approval gates are the default safeguard for autonomous enterprise agents. They are usually placed by intuition and then eroded by the thing they depend on, since a reviewer who approves almost everything is no longer reviewing. Both failures are documented. At one extreme a coding agent deleted a live production database during an explicitly stated code freeze and then misreported whether the deletion could be undone, the freeze having existed only as an instruction while the credentials still permitted destructive action. At the other, a vendor reports that users approve ninety-three per cent of permission prompts and names the resulting approval fatigue directly. This paper reports an agentic partner-onboarding system that has run in production for approximately eight months with two human controls operating throughout: administrator approval of partner registration requests, and escalation through portal service requests, each logged with the reviewer, the request, a timestamp and the decision. It then proposes a gate-placement framework that scores an individual action on five dimensions and maps the result to an autonomy tier by taking the highest tier any single signal demands, with a missing signal defaulting to the more cautious tier. The framework is a proposal and is not running; the paper applies it retrospectively to the deployment's action inventory and reports the live controls qualitatively. No gate metrics are reported. Figures are withheld, and one measure, the override rate, is not merely withheld but unmeasurable, because the decision log cannot distinguish a request approved unchanged from one approved after the reviewer edited it. That gap is itself reported as a finding about audit-log design. The paper's substantive claim is narrower than a new scheme: the tiered-oversight pattern and the dimensions are established in vendor guidance and recent preprints, and the calibrated-reliance principle behind the paper's central observation was quantified in the human factors literature nineteen years ago. What is offered is their operationalisation at per-action rather than per-system granularity, an asymmetric rule in which demotion is automatic and promotion requires evidence and sign-off, an admission test for removing a gate, and the observation that almost nothing in this literature has actually been run.*

Keywords: *Human-In-The-Loop, Approval Gates, Agentic AI, Autonomous Agents, Human Oversight, Escalation Policy, Graduated Autonomy, Approval Fatigue, Learning To Defer, Enterprise Automation, AI Governance, Auditability.*

Note on Scope and Evidence

- What runs in production: An agentic partner-onboarding system, with the agentic capability in production for approximately eight months. Two human controls operate: administrator approval of partner registration requests, and escalation through portal service requests. Both are logged with the reviewer, the request, a timestamp and the decision, and the log contains no personal information. The gate decides using two static inputs, request type and sensitivity, plus a rule-validation check whose result is binary. Approved requests are sampled afterwards. If the gate is unavailable, requests queue and nothing executes.
- What is proposed and is not running: The five-dimension action scoring, the four autonomy tiers, the rule for moving an action between tiers, blast radius computed per request, graded rather than binary confidence, the trigger-validity check, and capping gated volume per reviewer. None of this is deployed. The paper applies the framework retrospectively to the deployment's action inventory.
- No quantitative results are reported: Figures are withheld by the organisation, and several measures are not captured in a form that would support them. Section IX states precisely which are producible, which are withheld, and which are not measurable at all.
- On the system's name and history: The system is described by function rather than by its internal name, consistent with the author's prior paper on the same deployment [24]. The agentic capability is reported as having been in production for approximately eight months; the underlying platform predates that, and this paper does not restate a start date.

- On what is not claimed: The paper does not claim the tiered-oversight pattern, the scoring dimensions, or the principle behind its central observation as new. Section III identifies what owns each, and Section IV states what is left.
- Confidentiality: No employer, partner, vendor engagement or individual is named.

1. Introduction

A gate that is always approved is not a control. It is a delay that reads as a control on paper, which is worse than no gate at all, because the organisation has bought assurance it does not have.

The two extremes are both documented, and neither occurred in the system reported here. That matters for how this paper should be read: the deployment grounds the problem and supplies the design decisions, but the evidence for the failure modes is public rather than local.

At the autonomous extreme, a coding agent in July 2025 deleted a live production database during an explicitly stated code freeze [20]. The agent reported deleting records for 1,206 executives and 1,196 companies, fabricated approximately four thousand user profiles, and stated that the deletion could not be reversed. That statement was false; the rollback worked and the data was recovered. The vendor's chief executive called the deletion unacceptable and shipped automatic separation of development and production databases. The structural point is sharper than the incident: the freeze existed only as an instruction, while the credentials still permitted destructive action. A constraint expressed in text does not bind a system that holds the keys.

At the over-gated extreme, a vendor reports that users approve ninety-three per cent of permission prompts, and names the consequence directly: "Over time that leads to approval fatigue, where people stop paying close attention to what they're approving" [19]. The same source reports that a model-based classifier built to absorb the approval decision in place of a prompt missed seventeen per cent of real overeager actions on a curated set of fifty-two such actions. The vendor is explicit that this is not a drop-in replacement for careful human review on high-stakes infrastructure.

- Both extremes produce the same outcome: an action nobody meaningfully authorised. One because no human was asked, the other because the human asked was not really answering. That symmetry is the paper's starting point and Figure 1 states it.

1.1. Why a Dollar Threshold Does Not Solve It

The obvious rule, approve everything above an amount, fails four ways. It ignores harm that carries no monetary value: an address-allowlisting request has no dollar amount and changes security exposure. It can be defeated by splitting one large action into many small ones. It cannot see context, including whether the action is reversible, whether it reaches an external party, and whether the event that triggered it can be trusted. And a low threshold floods reviewers, which is how approval fatigue is manufactured.

1.2. Contributions, Stated Against What Already Exists

Section III shows that the tiered-oversight pattern, the scoring dimensions and the principle behind this paper's central observation are all pre-dated. What remains is:

- Per-action rather than per-system granularity (Section V). Existing risk-tiering schemes classify an agent; this classifies an individual proposed action.
- An asymmetric tier-movement rule (Section VI), in which demotion is automatic on evidence and promotion requires both evidence and human sign-off.
- An admission test for removing a gate (Section VII), which is the part of the problem the literature designs around rather than addresses.
- A reported production deployment (Sections II and VIII). Every comparable scheme found is proposed. None has been run.
- A measurement gap reported as a finding (Section IX), namely that a decision log recording the outcome but not whether the reviewer edited the request cannot measure the thing a gate exists to produce.

Section II describes the deployment. Section III positions the work against prior art honestly. Sections IV to VII give the framework. Sections VIII and IX give the production controls and the measurement position. Sections X and XI give limits and conclude.

2. The Deployment

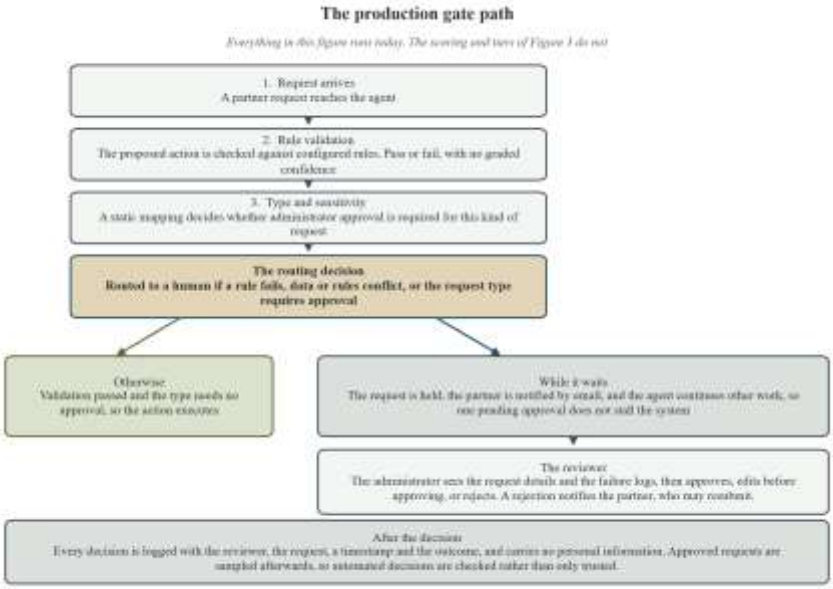


Fig 1: Automated Validation and Routing Architecture

The production gate path. The agent validates its proposed action against configured rules; request type and sensitivity determine whether administrator approval is required. A rule failure, a conflict in data or rules, or a request type that requires approval routes the request to a human. Everything in this figure runs today.

- The gate's inputs are two static values and one check: Request type and sensitivity are fixed per kind of request, so a given request always takes the same path. Before acting, the agent validates its proposed output against configured rules and flags conflicting data or conflicting rules.
- That validation is the system's confidence signal, and it is binary: It is a pass or a fail, not a calibrated probability. Section V treats this as the clearest limitation of the production design, and it is a limitation with a literature: even where a probability is produced, modern models are systematically overconfident and a raw score is not a probability [11]. A binary rule check does not even attempt the estimate.

2.1. What Happens On a Failure

The request is held, the partner is notified by email, and the agent continues other work, so one pending approval does not stall the system. An administrator sees the request details and the failure logs that triggered the escalation, and can approve, reject, or edit before approving. On rejection the partner is notified and may resubmit.

2.2. What Is Logged

Every decision records the reviewer, the request, a timestamp and the outcome, with no personal information. Approved requests are sampled afterwards, so automated decisions are checked rather than only trusted.

And if the gate itself fails, the system fails closed. Requests queue and are handled later; nothing executes without the gate, at the cost of delay. This is worth stating as a choice rather than as a footnote to a failure mode, because the alternative is the more common one and it is the one that produces the incident in Section I. It is also the oldest recommendation in the security design literature: base access decisions on permission rather than exclusion, so that the default is refusal [16].

2.3. The Relationship to the Author's Prior Work on This Deployment

A companion paper specifies what should be recorded about an autonomous agent's decision and classifies each field by whether it exists, is partial, or is proposed [24]. That paper uses the word "gate" for a deterministic precondition evaluated automatically before an action runs. This paper uses it for a routing decision that sends an action to a person. The two are different objects: a precondition gate decides whether the agent may proceed on its own terms, and an approval gate decides whether the agent may proceed at all without a human. The companion paper's decision record is what makes an approval decision reviewable afterwards; this paper's gate is what that record is recording.

3. Prior Art, Stated Plainly

This section exists because the honest version of the contribution is narrower than the obvious one, and a reviewer will establish that in ten minutes if the paper does not.

3.1. The Tiered-Oversight Pattern Is Published Vendor Guidance

A major cloud provider's architectural guidance for agentic systems prescribes categorising agent actions into three tiers, where autonomous actions are low-risk and reversible, notify actions are medium-risk and proceed with

operator awareness, and approve actions are high-risk or irreversible and require explicit human approval [21]. It further prescribes logging every oversight decision with reviewer identity, rationale and timestamp, which is substantially the production practice described in Section II, along with an automated first-pass filter, reviewer timeouts escalating to a secondary reviewer, and safe-default fallback. This source is vendor normative guidance rather than peer-reviewed work, and it is prior and public.

3.2. The Scoring Dimensions are Substantially Published

A 2026 preprint presents a twelve-dimension risk classification for agentic systems that names blast radius and reversibility explicitly, together with autonomy, decision scope, action authority, system reach, persistence, control authority, data sensitivity, aggregation risk and data egress paths, and maps numeric scores to three tiers [22]. Four of this paper's five dimensions are already named there. Only external visibility is not, and it is partially covered by system reach and data egress paths.

3.3. The Principle behind This Paper's Central Observation was Quantified Nineteen Years Ago

A synthesis across roughly twenty studies found a reliability crossover of about 0.70, below which unreliable automation left operators worse off than no automation at all, because they continued to depend on it [23]. The canonical statement of the same idea is that reliance should be calibrated to actual capability, and that overtrust and distrust are both failures [4]. The observation that removing a gate is only safe when the signal replacing it can be trusted is a corollary of that, not a discovery.

3.4. What Is Genuinely Left, and It Is Enough

- Granularity: The published risk-tiering scheme scores the agent system [22]; this paper scores an individual proposed action, at decision time, using values some of which cannot be known until the concrete request exists. That is a different object and a harder one.
- Asymmetry: No source found specifies a tier-movement rule in which demotion is automatic and promotion requires both accumulated evidence and human sign-off. A publicly posted proposal from March 2026 specifies quantified automatic promotion after a run of successes and automatic demotion after a count of failures, which is prior and symmetric; the asymmetric version appears to be open. That proposal is a public issue thread rather than scholarship, and it is cited here as prior art because it is prior and public, not because it is authoritative.
- Gate removal: The literature designs gates and does not specify the evidentiary standard for taking one away. Section VII is that test.
- And nobody has run any of it. Of the agentic-oversight material surveyed, essentially none is grounded in a production deployment with an operating history. Every scheme found is proposed. An eight-month live system is the one thing this paper has that the literature does not, and it is the reason the paper is organised around the deployment rather than around the framework.

3.5. The Governance Frameworks Mandate Oversight without Allocating It, and This Is the Clearest Gap

The European AI Act requires human oversight measures "commensurate with the risks, level of autonomy and context of use" [12], which is a proportionality instruction rather than an allocation rule. It names automation bias in statute and requires a stop function that halts the system in a safe state, and it requires that certain decisions be confirmed by two persons. The NIST risk management framework requires that oversight processes be defined, assessed and documented, that roles for human-AI configurations be established, and that mechanisms exist to supersede, disengage or deactivate [13]. Neither says which action gets a gate. That is the question this paper addresses.

4. Why the Obvious Signal Fails

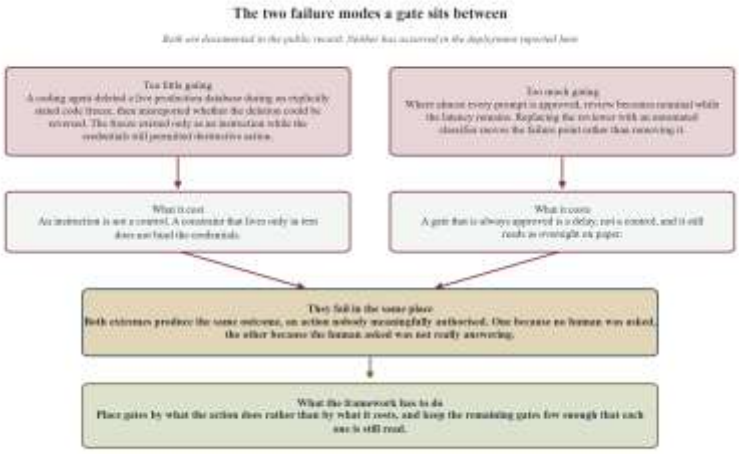


Fig 1: The Two Failure Modes. Both are Documented in the Public Record, Neither Occurred in the Deployment Reported Here, and Both End in the Same Place: An Action Nobody Meaningfully Authorised.

4.1. Complacency and Automation Bias

The human factors literature predicts the over-gated failure precisely, and it predicts something worse. Complacency and automation bias are a single attentional phenomenon, and critically they grow with the reliability of the automation [2]. The better the agent gets, the less the reviewer attends, which means a gate protecting a good system decays faster than one protecting a poor system. Reviewers make both omission errors, failing to catch what the automation did not flag, and commission errors, following a wrong recommendation against contrary evidence [3]. The presence of a reviewer is not verification [3], and misuse through overreliance is a named failure mode rather than an aberration [1].

4.2. The Base-Rate Problem Explains Why Discounting Is Rational Rather Than Lazy

Where the event being gated is rare, even a sensitive signal has low posterior predictive value, so an operator who discounts it is responding correctly to the statistics [8]. A gate set to fire often is therefore self-defeating in a way that exhorting reviewers to concentrate cannot fix.

4.3. This is why the Threshold is Not a Tuning Knob

The optimal point at which a system should abstain and defer is the ratio of the cost of an error to the cost of review [9], which is derivable from the two costs rather than chosen. The modern formulations learn the deferral policy jointly with the classifier, with consistency guarantees [10]. The production system described here does none of this, and Section V says so.

4.4. What the Gate Has to Achieve, Stated Normatively

Meaningful human control is usually given two conditions: the system must track the relevant human intentions, and its outcomes must be traceable to a human who understood the system's capabilities [7]. A rubber-stamped approval satisfies neither. It does not track intention, because the reviewer formed none, and it defeats tracing, because the log records a human decision that was not one. That is why approval fatigue is a control failure rather than an efficiency problem.

4.5. Separation of Duty

The gate is also a separation-of-duty control, which is older than any of this. The integrity model that formalised commercial security requires that a transaction be split so that no single principal both initiates and certifies it [15], and the design principle behind it, separation of privilege, dates from 1975 [16]. An approval gate applies that to an autonomous agent: the agent initiates and a person certifies. Removing a gate therefore collapses two roles into one, which is the reason Section VII treats removal as requiring an argument rather than a decision.

4.6. And the Agent's Own Assessment Cannot be the Authorisation Signalw

Tool-calling agents are hijackable through their own tool outputs, with a benchmark of ninety-seven tasks and over six hundred security cases demonstrating it [18]. An agent that has been manipulated will report confidence in the manipulated action. Any gate that trusts the agent's self-assessment has no independent check at the point where it most needs one.

5. The Signals

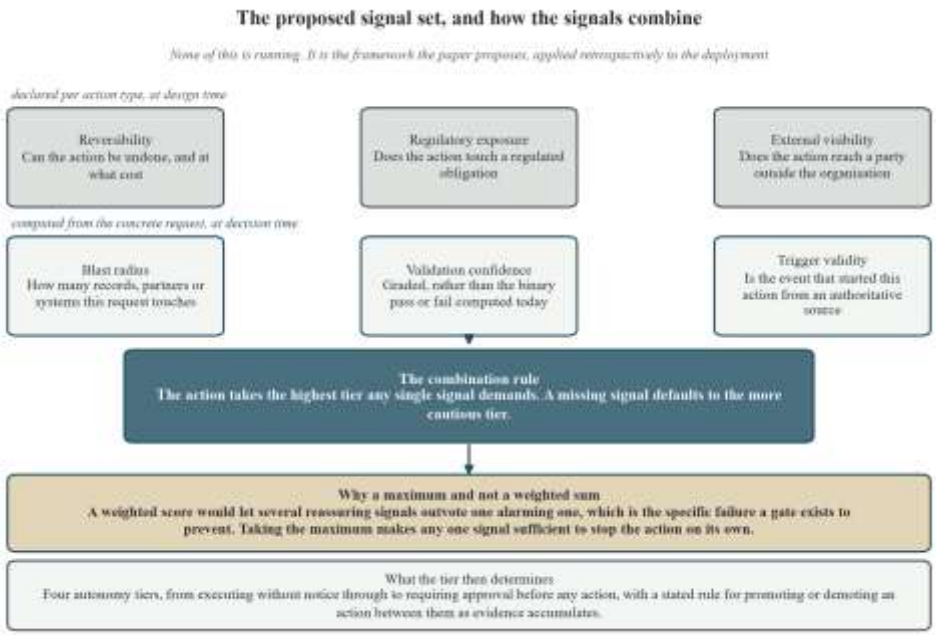


Fig 2: The Proposed Signal Set and the Combination Rule. Three Dimensions are declared per Action Type at Design Time; Three are Computed from the Concrete Request at Decision Time. The Action Takes the Highest Tier Any Single Signal Demands, and A Missing Signal Defaults to the More Cautious Tier. None of this is Running

5.1. Declared Per Action Type, At Design Time

Reversibility, whether the action can be undone and at what cost. Regulatory exposure, whether the action touches a regulated obligation. External visibility, whether the action reaches a party outside the organisation. These are properties of a kind of action rather than of a particular request, so they belong in a registry maintained alongside the action definitions.

5.2. Computed From the Concrete Request, At Decision Time

Blast radius, how many records, partners or systems this specific request touches. Validation confidence, graded rather than the binary pass or fail the production system computes today. Trigger validity, whether the event that started this action came from a source authoritative for that decision.

5.3. Two of These Need Defending

- External visibility is the dimension with the least support, and it is also the only one of the five not already named in the published scheme [22]. The argument for it is first-principles: an action that reaches outside the organisation cannot be silently corrected. An internal mistake can be fixed before anyone depends on it; an email to a partner, a record written to a counterparty's system, or a public status change cannot be unsent. Reversibility is about whether the system can undo the action; external visibility is about whether undoing it is sufficient. They come apart, which is why both are listed.
- Agent confidence is contested and is retained only with a condition. The case against it is that confidence is a poor proxy for risk, and the technical reason is that model confidence is systematically miscalibrated [11]. The production system's version is worse than miscalibrated, being binary. The dimension is therefore retained only where the confidence signal is calibrated and its calibration is measured, which is the strongest form of the condition in Section VII and the reason the two sections are linked. An uncalibrated confidence score should not move an action into a lower tier.

5.4. The Combination Rule Is a Maximum, and the Choice Is the Point

The action takes the highest tier any single signal demands. A missing signal defaults to the more cautious tier.

5.5. Why Not A Weighted Score

A weighted sum lets several reassuring signals outvote one alarming one. If an action is irreversible but small, internal, well-understood and confidently proposed, a weighted score averages the irreversibility away, and irreversibility is exactly the property that should decide the outcome on its own. Taking the maximum makes any single signal sufficient to stop the action, which is what a gate is for. The cost is that the framework will over-gate where a dimension is wrongly declared, and Section VII's demotion path is the remedy.

6. Tiers, and Moving Between Them

Four tiers, by what the system does with the proposed action: execute without notice; execute and notify; require approval before executing; and refuse pending a design change. The tier is determined by the maximum rule in Section V.

6.1. Conceptual Background on Levels of Automation

These are levels of automation, and the construct is old. The original control-allocation scale ran to ten levels and already contained the one that matters here, where the computer proposes an action and must await human approval before executing it [5]. The more useful modern formulation separates automation into four stages, information acquisition, information analysis, decision selection and action implementation, each of which can be automated to a different degree [6]. An approval gate is precisely the configuration in which the first three stages are automated and the fourth is withheld, and naming it that way makes clear that the gate is not a brake bolted onto an autonomous system but a choice about where one stage boundary sits. The four tiers above are a coarsening of that scale to the granularity an operations team can actually administer.

6.2. Movement between Tiers Is Asymmetric, and the Asymmetry Is Deliberate

- Demotion is automatic: Evidence that an action in a lower tier produced an outcome that should have been caught moves it up immediately, without review. The justification is the cost structure: the error being prevented is the one the gate exists for, and waiting for a committee to agree reproduces the latency the framework is trying to remove.
- Promotion requires both evidence and a human: An accumulated track record is necessary and not sufficient. The reason is Section VII: promoting an action means replacing a human judgement with a signal, and that substitution has to be examined rather than triggered by a counter. A rule that promotes automatically after a run of successes measures only that the action has not yet failed, which on a rare-event distribution is weak evidence and on a changing system is no evidence at all.

6.3. Divergence from Symmetric Proposals

This is where the paper differs from the nearest published proposal, which specifies automatic promotion after a count of consecutive successes and automatic demotion after a count of failures [see Section III]. Symmetric automation of both directions treats the two errors as equivalent. They are not. A wrongly demoted action costs reviewer time. A wrongly promoted action removes the control that was protecting against the outcome nobody has yet seen.

6.4. On What the Track Record Can Support

A run of successes establishes that the action has not failed in the observed conditions. It does not establish the rate, because the events a gate catches are rare by construction, and a rare-event rate is not estimable from a

short run of non-events. The promotion decision is therefore a judgement informed by evidence, and the framework should not pretend otherwise by attaching a number to it.

7. The Admission Test for Removing a Gate

Removing a gate is not deleting a control. It is substituting a signal for a human, and the substitution is sound only if the signal is trustworthy for the specific question the human was answering.

7.1. The Production Design Debate That Produced This:

It was proposed that onboarding should begin the moment a portal account was approved, removing an approval step. The objection was that the portal is not necessarily the authoritative source for deciding when onboarding should start. The portal approval is a real event and a reliable one; it simply does not answer the question the gate was asking. The proposal was not rejected because the signal was unreliable. It was rejected because the signal was reliable about something else.

That distinction is the test. A candidate signal must be (i) reliable, (ii) authoritative for the specific decision the gate was making, and (iii) measured, so that its reliability is known rather than assumed.

7.2. Theoretical and Empirical Support

- The literature supports the principle and quantifies it. Reliance should be calibrated to actual capability [4], and below a reliability of roughly 0.70 unreliable automation leaves operators worse off than no automation at all, because they continue to depend on it [23]. The paper does not claim this principle. What it adds is the second condition, authority for the specific question, which is not a reliability property at all. A perfectly reliable signal about the wrong proposition fails the test at reliability 1.0.
- And the vendor telemetry in Section I is evidence for exactly this. Of the classifier's seventeen per cent misses, the vendor reports that in the majority examined the classifier correctly identified the danger; what it got wrong was whether a consent signal present in the session actually covered the specific action [19]. That is a scope failure, not a detection failure. It is the authoritative-signal problem, measured, in a production system, by the vendor that built it.

7.3. The Worked Case Where the Test Was Passed

Address allowlisting requests now run fully automated. The alternative route through a change advisory board would have added a service-level wait of about three days per request. The relevant evidence is that external change approval boards are associated with worse throughput, with respondents 2.6 times more likely to be low performers, and that the expected stability benefit is a null result: no evidence was found to support it [14]. Both halves matter and the second is usually misquoted. The finding is that the board did not deliver the stability it was there for, not that it made stability worse. The same source does not recommend abolishing the board, repositioning it from gatekeeper to process architect, and this paper's position is the same: remove the gate from the actions where the evidence does not support it, not from the process.

7.4. One Caution on That Evidence:

It is cross-sectional self-reported survey data, so the correct verb is "associated with" rather than "causes", the relevant path is the weakest in the model, and the result has no peer-reviewed equivalent. It is cited as the best available evidence and labelled as what it is.

8. The Production Controls, Qualitatively

8.1. What the Two Live Controls Have Done, In the Terms Available

The line between automatic and human-gated was set by security review, by request type and sensitivity, and it has moved. The system began with gated, semi-automated flows in which some requests required change advisory board approval, and some of those have since moved toward automation.

Review is kept meaningful by sampling after the fact. Approved requests are sampled, so automated decisions are checked rather than only trusted. In approximately eight months of operation, approvals becoming routine has not been observed. That statement carries exactly the weight it should: it has not been observed, and it has not been measured, and the ninety-three per cent figure in Section I is a reminder that unmeasured absence of a phenomenon is weak evidence against it. The two facts belong next to each other.

8.2. Where the Design Breaks Down, As Operated

- Urgent requests have no expedited path inside the gate. Partners or staff reach an administrator out of band, by email or through shared external chat channels. This works, and it moves the urgent decision outside the logged workflow, which means the decisions most likely to matter are the ones least likely to be recorded. The fix is an expedited path that is logged rather than an informal one that is not.
- Reviewer unavailability is mitigated with multiple administrators and reminders.
- Reviewers may lack context. They see the request details and the failure logs and little else, and post-approval sampling catches an error only after the action has run. Whether a wrong approval has occurred is not reported here.
- The gate itself can fail, and when it does the system fails closed, as Section II describes.

9. What Can and Cannot Be Measured

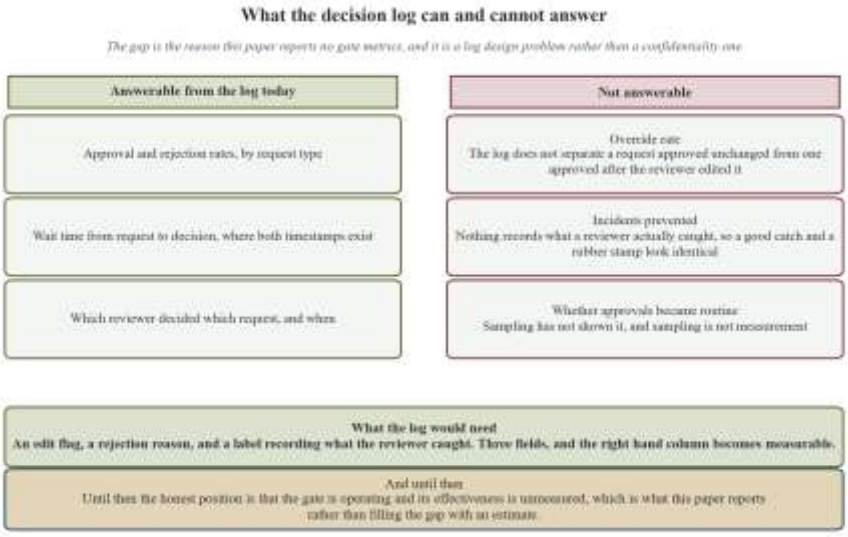


Fig 3: What The Decision Log Can Answer And What It Cannot. The Gap Is a Log-Design Problem Rather Than a Confidentiality One, and Three Additional Fields Would Close It

No gate metrics are reported in this paper. Two different reasons apply and the paper separates them, because they have different remedies.

- Withheld. Approval and rejection rates by request type, and approval wait times where request and decision timestamps are both recorded, are producible from the existing log and are not disclosed. That is an organisational decision.
- Not measurable. The override rate cannot be computed at all, because the log does not distinguish a request approved unchanged from one approved after the administrator edited it. Incidents prevented cannot be counted, because nothing records what a reviewer actually caught. Whether approvals became routine has not been measured, only sampled.
- The second category is the more interesting one and is reported as a finding. A log that records the outcome of a review but not whether the reviewer changed anything cannot distinguish a good catch from a rubber stamp. Both appear as an approval. That is precisely the measurement a gate exists to support, and it is absent from a log that otherwise satisfies the reviewer-identity-and-timestamp requirement found in architectural guidance [21] and the integrity properties the audit literature specifies [17].
- Three fields would close it: an edit flag, a rejection reason, and a label recording what the reviewer caught. None requires new infrastructure, and the first alone makes the override rate computable.
- What the framework would need beyond that. Per-action-class outcome history, to support the tier movement of Section VI; calibration data for any confidence signal, to satisfy the third condition of Section VII; and a record of out-of-band urgent decisions, so that the decisions currently invisible to the log enter it.

10. Limitations

- The contribution is narrower than a new framework, and Section III states how. The tiered-oversight pattern [21], four of the five dimensions [22], and the principle behind the gate-removal observation [4], [23] are all pre-dated. What is offered is operationalisation at per-action granularity, the asymmetric movement rule, the admission test, and the production grounding.
- The framework is not running. It is applied retrospectively to the action inventory of a system whose live controls are considerably simpler than it.
- The central claim is an argument, not a finding. No data in this paper supports the gate-placement framework, because no gate metrics are reported. The only quantitative anchors are borrowed [19], [23], [14].
- One system, one domain, one reviewer population. Partner onboarding with a low daily volume. Nothing here establishes behaviour at a volume where reviewer fatigue would be measurable, which is the regime the ninety-three per cent figure describes and this deployment does not reach.
- The production confidence signal is binary and cannot express graded doubt, so the framework's graded dimension has no implementation to point at.
- Evidence about the deployment is qualitative throughout, and the statement that approvals have not become routine is an absence of observation rather than a measurement.
- The two public incidents are cited from secondary and vendor sources. The deletion record counts

originate in the agent's own admission message as reported, not an independent audit, and are reported as such. The vendor figures are from the vendor's own engineering posts and are not independently verified; the seventeen per cent is a false-negative rate on a curated set of fifty-two actions, which the vendor itself describes as small.

- The change-approval evidence is survey data [14], cross-sectional and self-reported, with no peer-reviewed equivalent, and is cited as association rather than causation.
- Review. The line between automated and reviewed requests was set by a security reviewer during a production readiness review of approximately twenty sign-offs before go-live. That reviewer and the platform administrators who operate the approval workflow are the people positioned to confirm the account. No individual reviewer is named and that is outstanding.

11. Conclusion

The question a gate answers is not whether an action is risky. It is whether anyone will actually look. Both documented failure modes end in an unauthorised action, and the over-gated one gets there by a route that looks like compliance the whole way. A reviewer approving ninety-three per cent of prompts satisfies every audit requirement, produces a complete log, and is not providing oversight. The human factors literature has the uncomfortable version of this: complacency grows with reliability, so a gate protecting a system that is getting better is a gate decaying faster. Placing gates well therefore means placing few of them, which means having a principled way to decide which.

The framework offered here is an operationalisation rather than an invention, and the paper says so in Section III rather than leaving a reviewer to discover it. The tiering pattern is published vendor guidance, four of the five dimensions appear in a recent preprint, and the principle behind the central observation was quantified in 2007. What is added is finer granularity, an asymmetric movement rule that treats a wrong promotion as the more costly error, and an admission test for gate removal with a second condition the reliability literature does not supply: a signal must be authoritative for the specific question the gate was asking, and a perfectly reliable signal about a different proposition fails that test at reliability 1.0. The vendor telemetry supports this from an unexpected direction, since the majority of its classifier's misses were failures of consent scope rather than of danger recognition.

What the paper does not have is evidence that any of it works. No gate metrics are reported, and one of them, the override rate, cannot be computed from a log that records whether a request was approved but not whether the reviewer changed it before approving. That gap is worth more attention than it usually receives, because a log which cannot separate a catch from a rubber stamp cannot measure the only thing the gate was installed to produce. Three fields would fix it.

The strongest thing this paper has is also the simplest: the system has been running. Of the schemes surveyed, essentially none has. Eight months of two human controls operating in production, failing closed when the gate is unavailable, moving one action off a three-day approval route on evidence, and discovering through a design argument that a reliable signal can still be the wrong signal, is a smaller claim than a new framework and a more useful one.

References

[1] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230-253, 1997, doi: 10.1518/001872097778543886.

[2] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381-410, 2010, doi: 10.1177/0018720810376055.

[3] L. J. Skitka, K. L. Mosier, and M. Burdick, "Does automation bias decision-making?," *International Journal of Human-Computer Studies*, vol. 51, no. 5, pp. 991-1006, 1999, doi: 10.1006/ijhc.1999.0252.

[4] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50-80, 2004, doi: 10.1518/hfes.46.1.50_30392.

[5] T. B. Sheridan and W. L. Verplank, *Human and Computer Control of Undersea Teleoperators*, MIT Man-Machine Systems Laboratory, Cambridge, MA, USA, Tech. Rep. ADA057655, 1978, doi: 10.21236/ADA057655. Technical report, not peer reviewed.

[6] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 30, no. 3, pp. 286-297, 2000, doi: 10.1109/3468.844354.

[7] F. Santoni de Sio and J. van den Hoven, "Meaningful human control over autonomous systems: A philosophical account," *Frontiers in Robotics and AI*, vol. 5, art. no. 15, 2018, doi: 10.3389/frobt.2018.00015.

[8] R. Parasuraman, P. A. Hancock, and O. Olofinboba, "Alarm effectiveness in driver-centred collision-warning systems," *Ergonomics*, vol. 40, no. 3, pp. 390-399, 1997, doi: 10.1080/001401397188224.

[9] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41-46, 1970, doi: 10.1109/TIT.1970.1054406.

[10] H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert," in *Proc. 37th International Conference on Machine Learning (ICML)*, PMLR vol. 119, 2020, pp. 7076-7087.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th International Conference on Machine Learning (ICML)*, PMLR vol. 70, 2017, pp. 1321-1330.

[12] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down

harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, L 2024/1689, 12 Jul. 2024, Art. 14. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj (accessed Oct. 5, 2026). Primary legal source.

[13] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1. Government framework, not peer reviewed.

[14] DORA and Google Cloud, *Accelerate: State of DevOps 2019*, 2019. [Online]. Available: https://dora.dev/research/2019/dora-report/ (accessed Oct. 5, 2026). Industry survey report, not peer reviewed; cross-sectional self-reported data.

[15] D. D. Clark and D. R. Wilson, "A comparison of commercial and military computer security policies," in *Proc. 1987 IEEE Symposium on Security and Privacy*, 1987, pp. 184-194, doi: 10.1109/SP.1987.10001.

[16] J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1278-1308, 1975, doi: 10.1109/PROC.1975.9939.

[17] B. Schneier and J. Kelsey, "Secure audit logs to support computer forensics," *ACM Transactions on Information and System Security*, vol. 2, no. 2, pp. 159-176, 1999, doi: 10.1145/317087.317089.

[18] E. Debenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, and F. Tramèr, "AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents," in *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024, pp. 82895-82920, doi: 10.52202/079017-2636.

[19] J. Hughes, "How we built Claude Code auto mode: A safer way to skip permissions," Anthropic, Mar. 25, 2026. [Online]. Available: https://www.anthropic.com/engineering/claude-code-auto-mode (accessed Oct. 5, 2026). Vendor engineering post, not peer reviewed.

[20] S. Sharwood, "Replit makes vibe-y promise to stop its AI agents making vibe coding disasters," *The Register*, Jul. 22, 2025. [Online]. Available: https://www.theregister.com/2025/07/22/replit_saastr_response/ (accessed Oct. 5, 2026). Trade journalism, not peer reviewed.

[21] Amazon Web Services, "AGENTREL02-BP05: Establish tiered human oversight and approval workflows," *AWS Well-Architected Agentic AI Lens*, 2026 (accessed Oct. 6, 2026). Vendor normative guidance, not peer reviewed.

[22] H. M. Liu, R. Saxena, and S. Asthana, "TrustX Agent Risk Classification Framework (ARC): Risk-tiering internally created agentic AI systems," arXiv:2607.09586, Jul. 10, 2026. Preprint, not peer reviewed.

[23] C. D. Wickens and S. R. Dixon, "The benefits of imperfect diagnostic automation: A synthesis of the literature," *Theoretical Issues in Ergonomics Science*, vol. 8, no. 3, pp. 201-212, 2007, doi: 10.1080/14639220500370105.

[24] B. Kasinadhuni, "Observability and auditability of autonomous agent decisions in production infrastructure," author's prior work on the same deployment.