# Designing a Decision-Safety Mechanism for Autonomous Agents



INDEPENDENT RESEARCHER · 2026

## The problem

Autonomous agents built on large language models increasingly act on the world through the interfaces of operational systems: they read state, form a view, and take an action through a tool. A growing body of work shows these agents are unreliable at the precise moment of deciding whether to act at all, and that this is a distinct skill from solving the task. A 2026 benchmark that pairs cases where acting is correct with cases where abstaining is correct found that even the strongest models tested performed poorly on the pair, and that the ability to abstain scaled independently of the ability to solve the task, so a more capable agent was not thereby a more cautious one (AgentAbstain, 2026). A companion line of work frames the same behaviour as an explicit action in the agent's policy, choosing between answering, abstaining, and acting (Agentic Abstention, 2026). A counterfactual audit went further and removed an evidence-sufficiency dimension from an agent's triage, and found the agent then systematically overcommitted on items it could not support, which is direct evidence that the sufficiency judgement is doing real work (Support-State Triage, 2026). A survey of abstention in large language models describes the field as young and the behaviour as governed by a conjunction of answerability, confidence and human-values gates, with a genuine defer state only gestured at rather than built (Wen et al., 2025).

Whether the action is permitted is one question, and a well-studied one; whether the agent has adequate evidence to be proposing it at all is a different and earlier one. Most of the deployed guardrail work answers the first question. This project is about the second.

## Approach

I ran the work as design-science research, following the build-and-evaluate paradigm and its established scaffolding. The four methodological anchors are the original design science guidelines (**Hevner et al., 2004**), the six-activity process model that structures the build from problem identification through demonstration and evaluation to communication (**Peffers et al., 2007**), the knowledge-contribution and presentation framework that fixes how a design contribution is positioned and written up (**Gregor and Hevner, 2013**), and the evaluation framework that warrants an artificial and summative evaluation rather than a field study (**Venable et al., 2016**). I added two methodological sources that speak directly to the discipline this kind of project needs: an analysis of knowledge-contribution paths in design science, which frames a contribution as a general method and then its instantiation (**Akoka et al., 2023**), and an analysis of Australian design-science doctoral theses, which found that such projects commonly fail on scoping, on enunciating their philosophy, and on evaluating what they build (**Cater-Steel et al., 2019**). That last finding is why I specified the artefact tightly and held a hard boundary rather than letting the scope drift outward. Design decisions are recorded against these anchors with their basis, so the reasoning is traceable rather than assumed.

The artefact is specified as one coherent contribution in three connected parts.

The first part is an agent-facing integration surface: what a system-side interface has to expose before an agent's evidence position can be assessed at all. The relevant properties are provenance, recency, ordering, acknowledgement, idempotency, and an explicit representation of absence. This is not invented from nothing; each property has a literature that tells you what it means and how it can fail. Provenance has a settled vocabulary of why, how and where a datum came to be (**Cheney et al., 2009**), and recent provenance models represent supporting and conflicting evidence with reliability conditions (**Menotti et al., 2025**). Ordering has the logical-clock theory that fixes which precedence is even recoverable from a trace (Hrischuk and Woodside, 2002; Zholtkevych and Zozulia, 2025). Idempotency and acknowledgement have the exactly-once and end-to-end reliability literature, where exactly-once decomposes into an at-most-once safety property and an at-least-once liveness property (Frølund and Guerraoui, 2002; Kassam et al.

tive science and forensics establishes that absence is only diagnostic in proportion to how expected the missing thing was (Hsu et al., 2017; Thompson and Scurich, 2018). That is the formal reason a serious surface has to distinguish something that was not observed from something observed to be absent, rather than treating both as a null. Two recent agent-side systems show how much of this is already reachable in practice: a normalised tool contract that augments every tool response with provenance and idempotency signals (Pan and Hou, 2026), and a database-native reasoning design that makes provenance and ordering first-class, committed and auditable within a transaction (Ahmed, 2026). Both stop short of representing recency as an attribute, acknowledgement as receipt semantics, and absence as a first-class primitive, which is precisely where the surface in this project is aimed.

The second part is a deterministic assessment that takes a proposed action and the agent's evidence position and resolves to act, defer, or abstain. The important design commitment is where the determinism sits. The agent stays probabilistic; it reasons however it reasons. The assessment layer over it is deterministic, so the decision to act is auditable and reproducible from a recorded evidence position rather than being another sample from a model. This propose-then-certify shape, a stochastic component proposes and a deterministic layer certifies or defers, is itself an established pattern, described in operations research as a route to assured autonomy (Dai et al., 2026) and instantiated in several runtime enforcement systems. The assessment combines its checks without letting a strong signal on one dimension buy back a missing signal on another, which is the non-compensatory idea from decision theory, made deterministic here rather than probabilistic (van de Kaa, 2017; Khraibani et al., 2016). The three-way outcome, rather than a binary allow or block, has a direct precedent in runtime verification, where a monitor returns true, false, or inconclusive, and the inconclusive verdict maps naturally onto defer (Hallé and Villemaire, 2012).

The third part is a failure taxonomy grounded in observed execution traces rather than written in advance, and it is the headline. The distinctive move is not deriving a taxonomy from traces, which is now well established, but indexing the failures by the condition of the agent's evidence and coupling that index to the gate. I return to how the field made me sharpen that claim below.

## The boundary rule, and why it is the hardest part

I held one boundary rule across every design decision, and I treat its erosion as the project's most likely failure mode: the mechanism decides only whether the agent may act on its evidence, and

sive verdicts about the system under test (Salva, 2026); an instrument-control system polls a device and pronounces on its physical steady state (Zhang et al., 2026). Both are reasonable in their own settings and both are exactly the kind of external-state verdict this project must not produce. Keeping the output confined to the agent's own action, every time the surrounding literature offered a readiness verdict for free, was a continuous act of discipline rather than a one-time decision.

## How the field narrowed the claim

I did not get to keep the first version of the contribution. Three lines of prior work forced honest concessions. Deterministic gates placed before an agent action already exist, aimed at capability, policy, identity, or authority rather than evidence sufficiency: pre-action authorisation (Uchibeke, 2026), runtime rule enforcement (Wang et al., 2026, AgentSpec), execution-authority enforcement (Gill-Lakhowal, 2026), a pre-action legitimacy boundary (Lavi, 2026), and a bounded-autonomy controller with an escalation ladder (Sherry et al., 2026). The framing of evidence sufficiency and provenance is already named in print, as a review of auditable autonomy (Theodorakopoulos and Theodoropoulou, 2026), as a survey of evidence tracing and execution provenance (Wang et al., 2026), and even as a benchmark whose title claims governance-evidence sufficiency (Solozobov, 2026). And deriving a failure taxonomy from execution traces is well established: a grounded-theory taxonomy over a large multi-agent trace corpus (Cemri et al., 2025, NeurIPS), a trajectory-diagnosis taxonomy (Barke et al., 2026), an agent-environment taxonomy (Song et al., 2025), and a prospective-reflection taxonomy distilled from historical trajectories (Wang et al., 2026, PreFlect). Each of these is indexed to something other than evidence conditions, to system design, coordination, verification, exploration, exploitation, plan quality, or failure mechanism, which is what leaves the evidence-condition axis open. I conceded the occupied parts in writing rather than defending them, and re-based the claim onto the joint of evidence-condition indexing and gate coupling, with two refinements that survived the audit: a deterministic split between a remediable defer and a terminal abstain, and the not-observed versus observed-absent primitive on the surface.

## Status, stated plainly

This is at the design and pre-registration stage. The implementation and the empirical evaluation are specified in advance, using synthetic scenarios with constructed ground truth, no field data,

tion will have to sit inside a real calibration tension that the human-factors literature describes precisely: reliance that is too high maps to unsafe action and reliance that is too low maps to over-abstention (Lee and See, 2004; **Hoff and Bashir, 2015**), and explanations alone can worsen overreliance unless they prompt genuine verification (**Romeo and Conti, 2026**). The operational setting is drawn from real deployments of agents on port-terminal systems, including an LLM dispatching agent for automated container terminals (**PortAgent, 2025**) and terminal operating systems whose documented functionality centres on notification, traceability and timestamping (**Hervás-Peralta et al., 2019**), with a case study of record-versus-reality divergence, stale status and module-integration failure that motivates why an agent's evidence position needs to be assessable in the first place (**Gekara and Nguyen, 2020**). Those figures are attributed to their sources and are never asserted as background prevalence.

## What it demonstrates

Research design under a real methodology held end to end, formal specification of a decision procedure with stated properties, a surface grounded in the provenance, ordering, idempotency and absence literatures rather than asserted, command of a fast-moving and crowded body of related work, and disciplined scoping held to a firm boundary across the whole design.

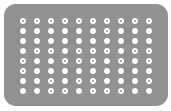
## Selected references


- Ahmed (2026). Database-Native Reasoning: Treating the Database as the Cognitive Substrate for AI Systems. IEEE Access, 14, 54912-54921. **DOI 10.1109/ACCESS.2026.3675808**.
- Akoka, J., Comyn-Wattiau, I., Prat, N., Storey, V.C. (2023). Knowledge contributions in design science research: Paths of knowledge types. Decision Support Systems, 166. **DOI 10.1016/j.dss.2022.113898**.
- AgentAbstain: Do LLM Agents Know When Not to Act? (2026). **arXiv:2607.10059**.
- Barke, S., et al. (2026). AgentRx: Diagnosing AI Agent Failures from Execution Trajectories. **arXiv:2602.02475**.
- Cater-Steel, A., Toleman, M., Rajaeian, M.M. (2019). Design Science Research in Doctoral Projects: An Analysis of Australian Theses. Journal of the Association for Information Systems, 20(12), 1844-1869. **DOI 10.17705/1jais.00587**.
- Cemri, M., Pan, M.Z., et al. (2025). Why Do Multi-Agent LLM Systems Fail? NeurIPS 2025

10.1016/j.artint.2021.103474.

- Cheney, J., Chiticariu, L., Tan, W.C. (2009). Provenance in Databases: Why, How, and Where. Foundations and Trends in Databases, 1(4). DOI 10.1561/1900000006.
- Dai, H., Simchi-Levi, D., Wu, F., Xie, J. (2026). Assured Autonomy: How Operations Research Powers and Orchestrates Generative AI Systems. DOI 10.1177/10591478261455127.
- Frølund, S., Guerraoui, R. (2002). e-Transactions: End-to-End Reliability for Three-Tier Architectures. IEEE Transactions on Software Engineering, 28(4).
- Gekara, V., Nguyen, V. (2020). Challenges of Implementing a Container Terminal Operating System: the Case of the Port of Mombasa. Journal of International Logistics and Trade, 18(1). DOI 10.24006/jilt.2020.18.1.049.
- Gill-Lakhowal (2026). Deterministic Runtime Enforcement: The Execution Authority for Autonomous AI Agents (L-DREA). IEEE Access, 14. DOI 10.1109/ACCESS.2026.3719838.
- Gregor, S., Hevner, A.R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. MIS Quarterly, 37(2), 337-355.
- Hallé, S., Villemaire, R. (2012). Runtime Enforcement of Web Service Message Contracts with Data. IEEE Transactions on Services Computing, 5(2). DOI 10.1109/TSC.2011.10.
- Hervás-Peralta, M., Poveda-Reyes, S., Molero, G.D., Santarremigia, F.E., Pastor-Ferrando, J.P. (2019). Improving the Performance of Dry and Maritime Ports by Increasing Knowledge about the Most Relevant Functionalities of the Terminal Operating System. Sustainability, 11(6), 1648. DOI 10.3390/su11061648.
- Hevner, A.R., March, S.T., Park, J., Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75-105.
- Hoff, K.A., Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors, 57(3). DOI 10.1177/0018720814547570.
- Hsu, A.S., Horng, A., Griffiths, T.L., Chater, N. (2017). When Absence of Evidence Is Evidence of Absence: Rational Inferences From Absent Data. Cognitive Science, 41 (Suppl. 5). DOI 10.1111/cogs.12356.
- Khraibani, R., de Palma, A., Picard, N., Kaysi, I. (2016). A new evaluation and decision making framework investigating the elimination-by-aspects model in transportation investment choices. Transport Policy, 48. DOI 10.1016/j.tranpol.2016.02.005.
- Larsen, K.R., et al. (2025). Validity in Design Science. MIS Quarterly, 49(4). DOI 10.25300/MISQ/2024/18064.
- Lavi, G. (2026). The Pre-Action Legitimacy Gap in AI Systems. arXiv:2604.24153.

- Menotti, L., Marchesin, S., Giachelle, F., Silvello, G. (2025). Provenance-driven nanopublications: representing source lineage and trust networks for multi-source assertions. International Journal on Digital Libraries, 26, 24. DOI 10.1007/s00799-025-00431-x.
- Pan, X., Hou, Y. (2026). Agent-First Tool APIs: Rethinking Enterprise Service Interfaces for LLM-Native Execution. arXiv:2605.10555.
- Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3), 45-77. DOI 10.2753/MIS0742-1222240302.
- PortAgent (2025). An LLM agent for vehicle dispatching in automated container terminals. arXiv:2512.14417.
- Romeo, G., Conti, D. (2026). Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. AI and Society, 41(2). DOI 10.1007/s00146-025-02422-7.
- Salva, S. (2026). Reliable execution of natural language test cases for GUI applications using LLM agents. Software Quality Journal, 34, 33. DOI 10.1007/s11219-026-09767-2.
- Sherry, Schreiber, Schreiber (2026). From taxonomy to stability: a CISO-ready architecture and best-practice blueprint for bounded autonomy in agentic cyber defense. Information and Computer Security. DOI 10.1108/ICS-05-2026-0264.
- Solozobov (2026). DEMM-Bench: A Cross-Regime Benchmark for Agent-Runtime Governance-Evidence Sufficiency. arXiv:2606.20634.
- Song, K., et al. (2025). Aegis: Taxonomy and Optimizations for Overcoming Agent-Environment Failures in LLM Agents. arXiv:2508.19504.
- Support-State Triage: Unlu (2026). Don't Start What You Can't Finish: A Counterfactual Audit of Support-State Triage in LLM Agents. arXiv:2604.16752.
- Theodorakopoulos, L., Theodoropoulou, A. (2026). Auditable LLM Autonomy for Operational Decision-Making: Big Data Evidence and Decision Traces. Computers, Materials and Continua. DOI 10.32604/cmc.2026.082270.
- Thompson, W.C., Scurich, N. (2018). When does absence of evidence constitute evidence of absence? Forensic Science International, 291. DOI 10.1016/j.forsciint.2018.08.040.
- Uchibeke (2026). Before the Tool Call: Deterministic Pre-Action Authorization for Autonomous AI Agents (Open Agent Passport). arXiv:2603.20953.
- van de Kaa, E.J. (2017). Establishing the relevance of non-compensatory choice algorithms from stated choice surveys. Judgment and Decision Making, 12(3).

- Wang, H., Poskitt, C., Sun, J. (2026). AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents. ICSE 2026. arXiv:2503.18666.
- Wang, Y., et al. (2026). From Agent Traces to Trust: A Survey of Evidence Tracing and Execution Provenance in LLM Agents. arXiv:2606.04990.
- Wang, Y., Cao, Y., Lin, L., Chen, J. (2026). PreFlect: From Retrospective to Prospective Reflection in Large Language Model Agents. arXiv:2602.07187.
- Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., Wang, L.L. (2025). Know Your Limits: A Survey of Abstention in Large Language Models. Transactions of the Association for Computational Linguistics, 13. DOI 10.1162/tacl_a_00754.
- Zhang, Z., et al. (2026). LLM-Enabled Multi-Agent Collaborative Instrument Control for Automated Microelectronic Testing. Advanced Devices and Instrumentation, 7, 0202. DOI 10.34133/adi.0202.




Download PDF ↓