

Agents Are Instruments, Not Pilots: Capturing Expert Signal as Reliability Infrastructure for Agentic AI

MARISA FERRARA BOSTON, Reins AI, USA

Drawing on two years of monitoring design and operational experience in regulated audit workflows, we examine how agentic AI changes not only expert work but the visibility of organizational judgment. Corrections, overrides, escalations, workarounds, and trust decisions are part of the distributed process through which accountable outputs are produced, yet current observability is strongest around system behavior and weakest where that judgment enters the work. We call this loss expert signal evaporation. We present a five-tier observation framework and an instrumentation gradient for expert responses, showing how these signals can be preserved across organizational scope. The design opportunity is to use agents not as substitutes for expert judgment, but as instruments that make distributed judgment observable and available for reliability and repair.

Additional Key Words and Phrases: agentic AI, expert signal evaporation, human computation, collective intelligence, regulated work

1 Expert Signal Evaporation

Expert work in regulated organizations is distributed across a hierarchy of judgment. Practitioners make task-level decisions, then reviewers evaluate those decisions across related work, then managers identify patterns across teams or engagements, and finally methodology and quality leaders determine whether those patterns require changes to standards or practice. Information therefore cycles not simply as completed outputs, but through corrections, qualifications, escalations, and other expressions of expert judgment. The reliability of the organization depends in part on whether this information remains available as work is reviewed and combined at progressively broader scopes.

Agentic AI changes these information flows. A system may produce an output that is subsequently corrected, reclassified, overridden, or worked around by an expert, while preserving little structured information about the judgment that made the work acceptable. At higher organizational levels, the problem compounds: an individual correction may remain visible while patterns across corrections, repeated workarounds, or declining reliance never become legible as organizational signals. We refer to this progressive loss of structured judgment across the human-AI work system as expert signal evaporation.

Reorganization of expert work provides a precedent. Research on audit offshoring and PCAOB observations of shared-service centers show that work can continue while contextual knowledge needed for review becomes harder to preserve [5, 10]. Agentic systems create a related risk: judgment may leave residue in tickets, messages, or notes, but unless that residue remains linked to the system behavior that prompted it and available across organizational scopes, it cannot function as reliability signal.

At the same time, AI-mediated work makes these interactions increasingly instrumentable. Changes to evidence, classifications, routing, escalation, and patterns of use create potential points at which expert judgment can be preserved in context. Agentic systems need not only replace or assist portions of expert work: they can also function as instruments through which the distributed computation of the expert organization becomes observable.

2 Expert Judgment as Situated Human Computation

Human computation and collective intelligence research examines how human judgments can be elicited, distributed, and recombined within computational systems [7, 8]; human-feedback methods similarly collect evaluations to shape model behavior [4, 9]. In both cases, the computational process is deliberately organized to obtain a human contribution.

Expert organizations present a different form of distributed computation. Hutchins showed that reliable cognition can be distributed across people, instruments, artifacts, and procedures [6]. Sociotechnical research in healthcare likewise locates technological performance within human and organizational systems [3, 13]. Regulated audit work has the same relevant structural property: judgment is produced and revised across practitioners, reviewers, managers, artifacts, standards, and tools.

The expert signal considered here arises within that distributed work. A reviewer who replaces an evidentiary source or a manager who changes the severity of a finding is not performing a separate feedback task; each intervention contributes to the production of the underlying work and depends on its evidence, standards, and organizational context. In regulated expert practice, the resulting output is itself a product of this distributed judgment. An audit opinion, for example, may have a single accountable signatory, but the judgment it expresses is constituted through the work of practitioners, reviewers, specialists, methodology, evidence, and successive acts of challenge and revision. The design problem is therefore not how to solicit additional human feedback on an AI-produced output, but how to preserve the distributed judgment through which the work product itself is produced.

3 Where Reliability Signals Arise

Preserving distributed judgment requires observing more than the final work product. We distinguish five tiers of observation, each corresponding to a different unit of behavior in the human-AI system. Output (Tier 0) captures what the system produced and the evidence on which it relied. Execution (Tier 1) captures the computational activity that produced it, including model and tool calls. Coordination (Tier 2) captures how components, artifacts, and decisions move through a complete workflow. Prior work operationalizes coordination-level monitoring across within-run, cross-run, and structural scopes on a 220-run testbed [1]. Existing AI evaluation and observability are strongest across these first three tiers, which primarily describe system behavior.

Communication (Tier 3) observes what changes when system behavior encounters expert judgment: evidence is replaced, an interpretation is revised, a classification changes, an output is rejected, or an issue is escalated. The relevant signal is not simply that an intervention occurred, but the relationship between the intervention and the system state that prompted it. Fitness (Tier 4) moves to a broader unit of analysis: whether patterns in system behavior and expert response remain consistent with the standards, purposes, and conditions under which the work is accountable. A correction may resolve a local output at Tier 3; repeated corrections of the same kind across engagements may indicate at Tier 4 that the system is systematically misaligned with the work. The principal instrumentation gap lies at these two tiers.

The tiers therefore do not represent progressively more detailed telemetry, but different kinds of evidence about the work. A practitioner-level intervention may resolve one instance; the same intervention recurring across teams or engagements may become an organizational signal about the system itself. Reliability infrastructure must preserve signals where they arise and keep them available across the levels through which work is produced, reviewed, and revised.

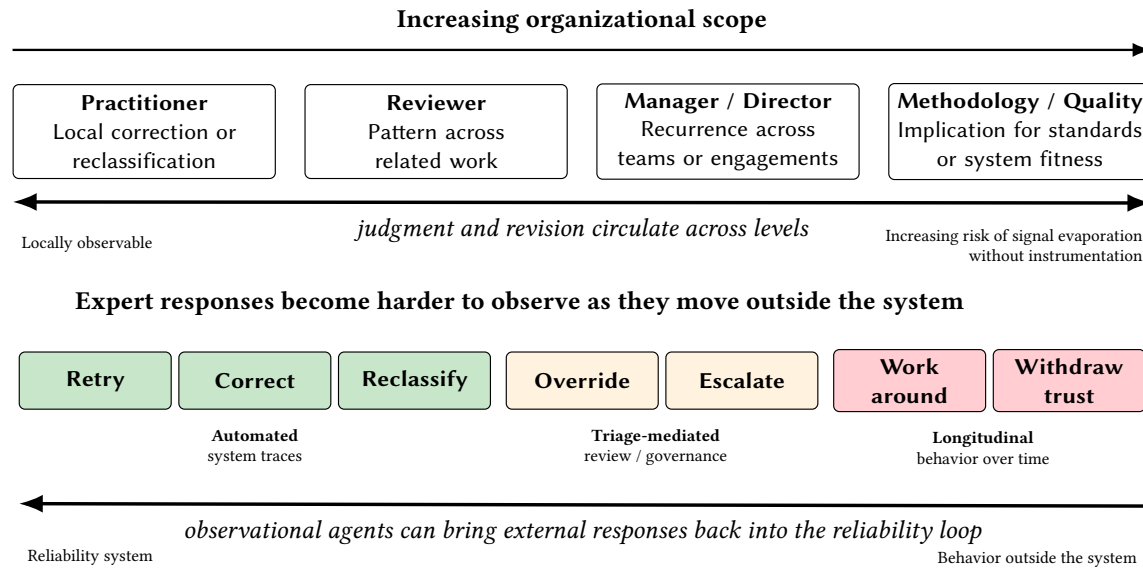


Fig. 1. Expert judgment circulates across organizational scope while expert responses vary in observability. Responses become harder to capture as they move outside the instrumented system. Agentic instrumentation can extend that boundary by observing workflow and longitudinal behavior and returning those signals to the reliability process.

4 From Expert Response to Repair

Expert responses leave different traces depending on where they occur. We organize them into seven forms: retry, correct, reclassify, override, escalate, work around, and withdraw trust. The first three occur close to the instrumented system and leave direct traces. Overrides and escalations enter review or governance workflows, where their meaning depends on organizational context. Workarounds and trust withdrawal may occur outside the system entirely, becoming visible only through absence or behavioral change over time.

This produces an instrumentation gradient. Automated signals can be recovered from telemetry, output deltas, and workflow metadata. Triage-mediated signals require enough context to connect an expert intervention to the behavior that prompted it. Longitudinal signals emerge only across runs, users, teams, or engagements. Organizations can therefore begin with existing system traces, instrument review and governance moments, and develop methods for patterns visible only over time. Monitoring agents can extend this boundary by observing surrounding workflow and fleet-level behavior for indirect traces of responses outside the primary system.

Figure 1 brings the two gradients together and locates the design opportunity where responses move furthest from the instrumented system.

The objective is not to collect every human action as feedback, but to preserve enough structure around expert intervention that judgment can become diagnostic evidence. Expert signal thus supplies a missing input to the reliability loop: not only what the system did, but what the expert organization had to do to make the resulting work trustworthy.

5 Implications and Scope

Organizational reliability cannot be recovered by aggregating lower-level measures alone. Output scores, execution traces, and counts of corrections describe individual system interactions but not whether the organization is preserving its capacity to exercise and compose judgment across the work. This extends sociotechnical accounts of system boundaries [12] and organizational approaches to algorithmic accountability [11]: oversight also requires preserving whether recurring interventions become recognizable patterns and whether those patterns alter review, methodology, or judgments of fitness for purpose. Expert signal is therefore not simply another evaluation dataset, but infrastructure through which an organization can learn from AI-mediated work.

The framework was developed through approximately two years of monitoring design and operational experience with agentic systems in regulated audit workflows. We offer it as a design framework, not a validated cross-domain taxonomy; whether its observation tiers and response categories transfer unchanged remains empirical. Human-AI reliance research has examined whether individuals critically engage with AI recommendations [2]; our concern is complementary but organizational in scale. Capturing expert signal also creates governance risks: instrumentation for system learning can become worker surveillance unless purpose, access, aggregation, and individual-level use are constrained.

The opportunity, however, is unusual. Agentic AI does not only disrupt the channels through which expert judgment has historically circulated; it also creates computational surfaces on which those channels can be observed. Systems can be designed so that the trace that makes a failure observable to the system also makes the judgment behind the work legible to the responsible expert, while connecting those signals across organizational scope and to responses that occur outside the primary system. In that role, agents are not substitutes for the expert organization’s judgment. They are instruments through which the distributed judgment that produces trustworthy work can become observable, persistent, and available for repair.

References

- [1] Marisa Ferrara Boston, Glen Hanson, Effi Georgala, JD Hudgens, and Heather Frase. 2026. Monitoring Agentic Systems Before They’re Reliable. In *Agentic Software Engineering Workshop at the 1st ACM Conference on AI and Agentic Systems (CAIS)*. Non-archival workshop paper. arXiv:2606.02494.
- [2] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 188 (2021), 21 pages. doi:10.1145/3449287
- [3] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. 2017. Unintended Consequences of Machine Learning in Medicine. *JAMA* 318, 6 (2017), 517–518.
- [4] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf
- [5] D. H. Downey. 2018. An Exploration of Offshoring in Audit Practice and the Potential Consequences of Associated Work “Re-Design” on Auditor Performance. *Auditing: A Journal of Practice & Theory* 37, 2 (2018), 197–223.
- [6] Edwin Hutchins. 1995. *Cognition in the Wild*. MIT Press, Cambridge, MA.
- [7] Aniket Kittur, Jeffrey V. Nickerson, Michael S. Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matthew Lease, and John J. Horton. 2013. The Future of Crowd Work. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (2013), 1301–1318.
- [8] Thomas W. Malone, Robert Laubacher, and Chrysanthos Dellarocas. 2010. The Collective Intelligence Genome. *MIT Sloan Management Review* 51, 3 (2010), 21–31.
- [9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [10] Public Company Accounting Oversight Board. 2024. *Staff Update: Observations Related to the Use of Shared Service Centers*. Spotlight. PCAOB.
- [11] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 33–44. doi:10.1145/3351095.3372873

- [12] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 59–68. doi:10.1145/3287560.3287598
- [13] Dean F. Sittig and Hardeep Singh. 2010. A New Sociotechnical Model for Studying Health Information Technology in Complex Adaptive Healthcare Systems. *Quality and Safety in Health Care* 19, Suppl 3 (2010), i68–i74.