Machine Intelligence & Responsible Systems
PERSPECTIVE | COMMISSIONED SYNTHESIS

Human Oversight in Agentic AI: Authority, Interaction, and Failure Containment

Brian Baker - Corresponding author

Abstract

This perspective examines human oversight in agentic artificial-intelligence systems. The organizing question is where authority should reside when planning, delegation, tool use, and recovery are distributed across agents. Ten related scholarly sources are synthesized through a decision-centered framework spanning problem definition, mechanism, measurement, evaluation, implementation, and governance. The review does not invent experiments, pooled estimates, or unreported quantitative results. It instead evaluates the strength and transferability of the available evidence, with particular attention to assuming that adding a human approval step guarantees meaningful control. The resulting framework links technical or empirical performance to explicit use conditions and identifies tests that should precede wider adoption in agentic systems acting on organizational resources.

Keywords: agentic AI, human oversight, autonomous agents, safety, auditability

Synthesis lens	Principal question
Mechanism	the interaction among task decomposition, shared state, tool permissions, monitoring, and shutdown
Evaluation	intervention latency, coordinated failure, authorization drift, and recovery from partial execution
Primary risk	assuming that adding a human approval step guarantees meaningful control

1. Introduction

Research on human oversight in agentic artificial-intelligence systems sits at the boundary between a promising component and a consequential decision. The core question is where authority should reside when planning, delegation, tool use, and recovery are distributed across agents. Answering it requires more than assembling favorable results. It requires a chain of evidence that makes the problem boundary, mechanism, measurement, comparator, uncertainty, and accountable decision visible to the reader.

This article presents a structured narrative review of ten related publications. Sources were selected for topical relevance, traceable publication metadata, and complementary coverage of technical, empirical, and governance questions. No meta-analytic pooling is attempted because the included studies differ in designs, populations, system boundaries, and outcomes. The purpose is decision-oriented synthesis: to identify what each source can support, where transfer remains uncertain, and which evidence would justify the next stage of use.

The central risk is assuming that adding a human approval step guarantees meaningful control. A top-line metric or broad policy label can appear decisive while concealing the conditions that produced it. Accordingly, the synthesis separates observation from interpretation and implementation from validation. This separation is especially important when the same system creates benefits for one user or institution while transferring cost, risk, or administrative burden to another.

2. Evidence Base and Problem Boundary

A defensible review begins by defining the unit of analysis. For the present topic, the relevant boundary includes inputs, institutional or technical context, the mechanism producing an output, the person or system acting on that output, and the downstream outcome. Evidence falls outside the boundary when it measures only an intermediate proxy yet is used to claim an end-to-end benefit.

Calisto et al. (2022) addresses "BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions." In this synthesis, that work is used as a source on problem definition within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Borghoff et al. (2025) addresses "Human-artificial interaction in the age of agentic AI: a system-theoretical approach." In this synthesis, that work is used as a source on conceptual boundary within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Seaborn et al. (2021) addresses "Voice in Human–Agent Interaction." In this synthesis, that work is used as a source on measurement within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Together, these works provide complementary entry points, but their conclusions should not be flattened into a single ranking. Differences in data generation, setting, temporal horizon, and outcome definition may be substantively meaningful. A review should therefore preserve those differences and state where analogy, rather than direct replication, connects one domain to another.

3. Mechanism and System Design

The operative mechanism is the interaction among task decomposition, shared state, tool permissions, monitoring, and shutdown. This formulation discourages component-only reasoning. A model, platform, policy, assay, or infrastructure service acquires practical meaning only when its interfaces and use conditions are specified. The evidence chain should describe what information is available, what transformation occurs, which assumptions make that transformation credible, and who is authorized to act.

Chandra et al. (2022) addresses "To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents." In this synthesis, that work is used as a source on system mechanism within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Samson and Zaleskiewicz (2025) addresses "Warmth and Competence in Human-AI Agent Interactions." In this synthesis, that work is used as a source on representation choice within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Wang et al. (2022) addresses "Learners' perceived AI presences in AI-supported language learning: a study of AI as a humanized agent from community of inquiry." In this synthesis, that work is used as a source on failure analysis within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

These studies also show why modular claims require interface tests. A component may perform well while the complete pathway fails because inputs drift, users interpret outputs differently, recovery semantics are ambiguous, or institutional incentives change. Design documentation should therefore include not only the intended pathway but also foreseeable deviations, degraded modes, and conditions under which the system should stop or defer.

4. Evaluation and Uncertainty

Evaluation should center on intervention latency, coordinated failure, authorization drift, and recovery from partial execution. Average performance remains useful, but it should be accompanied by distributions, error categories, relevant subgroups or operating regimes, and uncertainty at the point where a decision changes. Comparators must isolate the value of the proposed addition rather than compare a mature intervention with an intentionally weak baseline.

Gnewuch et al. (2023) addresses "More Than a Bot? The Impact of Disclosing Human Involvement on Customer Interactions with Hybrid Service Agents." In this synthesis, that work is used as a source on

comparative evaluation within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Lim et al. (2025) addresses "Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations." In this synthesis, that work is used as a source on uncertainty within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

External validation should introduce meaningful shifts in setting, time, population, workload, market structure, or environmental conditions. A nominally independent sample can still be too similar to test transfer. Conversely, a failed transfer is scientifically informative when the changed conditions are measured and the mechanism of failure is examined rather than hidden in an aggregate score.

5. Translation and Governance

Translation to agentic systems acting on organizational resources depends on more than technical readiness. Decision rights, documentation, maintenance, monitoring, appeal, and recovery are part of the intervention. The governance requirement is least-privilege tools, immutable logs, explicit handoffs, and rehearsed containment procedures. Each control should be connected to a named failure mode and should identify an owner, evidence source, review interval, and escalation path.

Sedlakova and Trachsel (2022) addresses "Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?." In this synthesis, that work is used as a source on implementation within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Li and Zhang (2023) addresses "Chatbots or me? Consumers' switching between human agents and conversational agents." In this synthesis, that work is used as a source on governance within human oversight in agentic artificial-intelligence systems. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for agentic systems acting on organizational resources. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

A credible implementation plan also defines sunset and revalidation conditions. New data, software updates, institutional restructuring, population change, or shifts in incentives can invalidate previously reasonable assumptions. Monitoring should therefore test the validity of the evidence chain, not merely whether a service remains online or a headline metric remains stable.

6. Research Agenda

Future studies should preregister or otherwise predefine the intended decision, principal outcomes, exclusions, and analysis plan when feasible. They should report missingness, sensitivity analyses, negative or null findings, and the cost of errors to differently situated stakeholders. Reproducible artifacts should include sufficient metadata to reconstruct data provenance, model or analytical versions, parameter choices, and the sequence from evidence to conclusion.

Cross-disciplinary work needs a shared object of explanation rather than a collection of adjacent vocabularies. For human oversight in agentic artificial-intelligence systems, that shared object is the complete decision pathway. Technical, clinical, social, environmental, or economic expertise should each change the design or interpretation of the study. When evidence remains incomplete, the useful output is a bounded hypothesis, a transparent uncertainty statement, or a request for additional measurement.

7. Conclusion

The literature supports a disciplined approach to human oversight in agentic artificial-intelligence systems: define the decision, preserve system boundaries, test consequential failure modes, connect uncertainty to action, and assign accountable control. The strongest conclusion is not that one method is universally superior. It is that credible use in agentic systems acting on organizational resources requires an auditable link from source evidence to design choice, evaluation result, and governed decision.

References

Borghoff, U. M., Bottoni, P., & Pareschi, R. (2025). Human-artificial interaction in the age of agentic AI: a system-theoretical approach. Frontiers in Human Dynamics, 7. https://doi.org/10.3389/fhumd.2025.1579166

Calisto, F. M., Santiago, C., Nunes, N., & Nascimento, J. C. (2022). BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. Artificial Intelligence in Medicine, 127, 102285. https://doi.org/10.1016/j.artmed.2022.102285

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. Journal of Management Information Systems, 39(4), 969-1005. https://doi.org/10.1080/07421222.2022.2127441

Gnewuch, U., Morana, S., Hinz, O., Kellner, R., & Maedche, A. (2023). More Than a Bot? The Impact of Disclosing Human Involvement on Customer Interactions with Hybrid Service Agents. Information Systems Research, 35(3), 936-955. https://doi.org/10.1287/isre.2022.0152

Li, C. Y., & Zhang, J. T. (2023). Chatbots or me? Consumers' switching between human agents and conversational agents. Journal of Retailing and Consumer Services, 72, 103264. https://doi.org/10.1016/j.jretconser.2023.103264

Lim, S., Schmälzle, R., & Bente, G. (2025). Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations. Computers in Human Behavior Artificial Humans, 5, 100172. https://doi.org/10.1016/j.chbah.2025.100172

Samson, K., & Zaleskiewicz, T. (2025). Warmth and Competence in Human-AI Agent Interactions. Polish Psychological Bulletin, 148-160. https://doi.org/10.24425/ppb.2025.153987

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). Voice in Human–Agent Interaction. ACM Computing Surveys, 54(4), 1-43. https://doi.org/10.1145/3386867

Sedlakova, J., & Trachsel, M. (2022). Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?. The American Journal of Bioethics, 23(5), 4-13. https://doi.org/10.1080/15265161.2022.2048739

Wang, X., Pang, H., Wallace, M. P., Wang, Q., & Chen, W. (2022). Learners' perceived AI presences in AI-supported language learning: a study of AI as a humanized agent from community of inquiry. Computer Assisted

Language Learning, 37(4), 814-840.
https://doi.org/10.1080/09588221.2022.2056203