Global Journal of Scientific Research Publications
Published by: Sangotra Publishing House *ISSN (Online): 3139-5821*

Who Authorized the Machine? Closing the Accountability Gap in Autonomous AI Enterprises

Alex Mathew[1]
[1]Department of CS & Cybersecurity, Bethany College, West Viginia, USA

Abstract— *Organizations can't prove that the original human approval was for the final action, but autonomous AI agents can be successfully authenticated. This article focuses on the question as to how an organization can show that a consequential agent action remained within the parameters of an approved human action. It builds on an analysis of the agent security, access control, accountability, and AI audit literature and presents the conceptual model of the AI Delegation Accountability Framework, consisting of five dimensions of control, action-bound approval, and a subset-of-authority test at each hop of the delegation chain. It outlines a phased implementation, a backwards trace audit with six questions and a plan for validation, as well as identifying limitations and creating directions for empirical testing.*

Keywords— *E agentic AI; AI governance; accountability; authorization; delegation; IT audit; identity and access management; cybersecurity*.

I. INTRODUCTION

Organizations are beginning to let autonomous AI agents prepare payments, change records, and call other agents and services on their behalf. These agents are characterized as those that work towards complicated objectives with minimal human involvement (Acharya et al., 2025). One possible scenario involves a procurement agent creating a payment order and sending verification to a second procurement agent who will alter the supplier's bank details via a third-party service prior to the payment transfer. All agents are authenticated, all systems keep an event, but nobody can demonstrate that the approval of the human was for the last operation.

Throughout, four terms are used. Authentication identifies the agent, authorization is used to allow or deny the agent certain permissions, while delegation is used to bind a human identity to an agent, enabling the agent to act on behalf of the human (Booth et al., 2026). Accountability is the obligation of an actor to give an account of an outcome to a forum which has the power to question the actor (Novelli et al., 2024). This article poses the question of how an organization can prove that the consequential agent action was within a human's approval. It helps to provide the AI Delegation Accountability Framework, which is somewhat trivial in that it includes a subset-of-authority test at each hop of the delegation chain and the approval must be in the form of action.

II. LITERATURE REVIEW

Research on agent security and access control supplies the first layer of the foundation. Prompt injection is a potential vulnerability of LLMs, where an agent with legitimate access can be redirected (Yao et al., 2024; Booth et al., 2026). The authentication and access control are at the core of the secure architectures as per zero trust research (Syed et al., 2022; Dhiman et al., 2024). A recent review shows that there is no single solution that meets authentication, authorization, access control and auditing in one go (Mushtaq et al., 2025). These controls are asking if access was allowed, as opposed to if the final action was one that would be approved by a person.

The second layer of scholarship on accountability and responsibility comes from the literature. Answerability of AI can be seen as the answerability to a forum, as per the agreed standards (Novelli et al., 2024). When multiple actors work toward a goal, there may be a responsibility gap in which none of the actors is able to fill (Vallor & Vierkant, 2024). Don't take the responsibility away from people in the surroundings of the system (Chan et al., 2023). Delegation chains help to make this problem real, in that each additional agent introduces another hand.

The third layer provided by auditing and governance research. Mokander et al. (2024) audit governance, models and applications, which are three layered. Agentic workflow control tests are added to the control tests, such as authority boundary testing and escalation checks (Jain, 2026). These tests are useful but don't specify who gets to approve at the link of agents.

Practitioner sources describe three recurring weaknesses. Identity is not authority because if an agent borrows a user's account, then it can pretend to be that user and compromise on the non-repudiation (Fisher & Galluzzo, 2026). The logs do not provide evidence of authorization (authorities may have approved it but the record does not indicate that), and retention and tamper evidence needs to be tested, as well (Jain, 2026). Delegation creates accountability issues because there is a risk that scope expands due to additional agents or tools added by the delegating agent (Bahidika, 2026). Organizations cannot prove authorization when run against each weakness individually, but when combined together, each weakness can become a problem.

Combined, these literatures leave a specific gap. In existing standards development (Booth et al., 2026), ID, authorization and delegation processes are already considered separate controls. Five popular agent protocols and frameworks were examined and none of 25 evaluations met all the scope, intent and provenance requirements for maintaining scope, intent and provenance in a chain (Bahidika, 2026). The question that hasn't been answered is what would human approval be?

III. METHOD

The study uses a conceptual design method with three

Alex Mathew, "Who Authorized the Machine? Closing the Accountability Gap in Autonomous AI Enterprises," *Global Journal of Scientific Research Publications*, Volume 1, Issue 6, pp. 23-25, November 2026.

steps. The following sources were consulted to extract requirements: NIST concept paper and scholarly and practitioner sources (Booth et al., 2026). For each requirement there was a mapping with a control and gaps and overlaps were identified by comparing the delegation chain conditions in the literature (Bahidika, 2026). This produced the controls which were later coupled to the governance, model and application audit layers (Mökander et al., 2024). No interviews or live systems were used in this method, which is a document-based synthesis.

IV. The AI Delegation Accountability Framework

The framework has five dimensions, summarized in Table I. Identity and ownership provide each agent with a unique identity, an accountable owner and a purpose that is documented (Fisher & Galluzzo, 2026). Bounded authority must be delegated scope, which must be always be a subset of the delegator's scope at each hop (Bahidika 2026). Delegation traceability, independent verification, accountability and recovery round out the list. The two features that make this framework different are that the approval is tied to the specific action approved (e.g. a named payee and an amount) and that the subset test is applied at every hop and not just at the point of access.

TABLE I. AI Delegation Accountability Framework: Controls, Evidence, and Tests.

Dimension	Control objective	Mechanism	Evidence produced	Audit test	Pass criterion
Identity and ownership	Every agent is attributable to an accountable human	Unique agent identity; owner and purpose register; no shared credentials	Agent inventory with owner and purpose	Sample agents and confirm owner, purpose, and credential	All sampled agents have unique credentials and a named owner
Bounded authority	Agent acts only within approved scope	Least privilege; short-lived, scope-limited tokens; policy engine	Permission grants and token scopes	Compare granted scope with the delegator's scope at each hop	Delegated scope is a subset of the delegator's scope at every hop
Delegation traceability	Who delegated what to whom can be reconstructed	Signed delegation records linking each hop to the initiating authority	Tamper-evident delegation chain	Backward trace from a sampled action	Unbroken chain to the initiating authority
Independent verification	Out-of-scope actions are blocked or flagged	Pre-execution policy check; risk-tiered, action-bound approval	Approval records naming approver, payee, and amount	Submit planted out-of-scope requests	All planted requests blocked or escalated
Accountability and recovery	Responsibility is assigned and harm is limited	Escalation paths; suspend and revoke; rollback procedure	Exercise records and incident logs	Run a tabletop exercise and a live revocation	Agent suspended and access revoked within the agreed time

V. Implementation and Evaluation

Implementation should be done in three stages, starting with one of the more impactful workflows like vendor payments. The first phase involves an inventory of all the agents, owners, permissions and tools, with each action type assigned to a risk tier. The second phase of the identity and access management team involves the issuance of unique identities for agents and short-lived, scope-limited credentials via token exchange, and the deployment of a policy engine that enforces action-bound approval for the highest level. During the third phase, the internal audit conducts the tests as outlined in Table 1 and expands the scope only if the evidence shows, and the named business owner is accountable during the entire process. Risk tiers are used to speed up low-risk actions, and to ensure that human approval is only sought for actions that have real consequences, such as transfers and record changes.

Evaluation is done in a backwards trace. The auditor is responsible for picking a relevant action and then executing it step by step from the authority that initiated the action, verifying that the scope of the action is the same as the scope of the action that was granted at each step (Bahidika, 2026). The first three questions are related to the source of authority for the action, inheritance of permissions, and detection of unauthorized changes in scope or purpose of the action. The final three questions concern the binding of the approval to the specific action, the ability of an independent reviewer to recreate the decision using reliable evidence and whether revocation and escalation have been used. The following findings are assessed as 'pass', 'partial' or 'fail' on the criteria listed in Table I.

Validation should be based on the expertise, pilot and measurable indicators. From the beginning indicators should be embedded within the system, such that evidence is available prior to the start of an audit. The vendor payment workflow can be demonstrated through a pilot or tabletop exercise to determine if out-of-scope requests are prevented from being planted and if the revocation process is effective. They include: percentage of consequential actions that can be attributed to a named approver, percentage of hops in which scope is a subset, percentage of planted requests detected and time to stop an agent. The outcome of the pilot would indicate whether or not the thresholds in Table 1 are attainable.

VI. Comparison With Existing Approaches

The framework is an augmentation of existing approaches, and is not intended to supplant them. In Zero Trust architectures, authentication and access control are always performed (Syed et al., 2022). The three main types of responsible AI governance are known as structural, relational and procedural practices (Papagiannidis et al., 2025). Management system standards like ISO/IEC 42001, the NIST AI Risk Management Framework, and delegation constructs like OAuth token exchange provide helpful pieces of the puzzle, and the framework builds on these for governance and

Alex Mathew, "Who Authorized the Machine? Closing the Accountability Gap in Autonomous AI Enterprises," *Global Journal of Scientific Research Publications*, Volume 1, Issue 6, pp. 23-25, November 2026.

tooling. Its supplemental features are the action-bound approval record and the hop-by-hop subset test, both of which serve as evidence for an auditor to test.

VII. Limitations

Several shortcomings apply. The framework is a conceptual one and not tested in real organizations. The chain evidence that is reviewed is based on documentation of protocols and frameworks, and not deployed systems (Bahidika, 2026). In addition to standards and practitioner documents that have been peer-reviewed in journals, the source base also features documents that have not been so reviewed. The framework should be further validated by conducting case studies and interviews with practitioners to assess the practical application of the framework, and to evaluate if the proposed indicators are predictive of an audit outcome.

VIII. Conclusion

The primary question in autonomous enterprises is who authorized the machine? Currently, standards bodies are working towards identifying, authorizing and connecting agents with human principals but this is still in development (Booth et al., 2026). Each consequential action should be demonstrated to be within the scope of legitimate and traceable authority in the organization. It will be that demonstration and not the logs themselves that will separate trusted autonomous enterprises from exposed.

References

[1] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals: A comprehensive survey," *IEEE Access*, vol. 13, pp. 18912–18936, 2025. https://doi.org/10.1109/ACCESS.2025.3532853

[2] O. Bahidika, "When agents delegate: Accountability chains in multi-agent AI systems," *International Journal of Advanced Computer Science and Applications*, vol. 17, issue 9, pp. 24–31, 2026. https://thesai.org/Downloads/Volume17No9/Paper_3-When_Agents_Delegate.pdf

[3] H. Booth, B. Fisher, R. Galluzzo, and J. Roberts, "Accelerating the adoption of software and AI agent identity and authorization [Concept paper]," National Institute of Standards and Technology, National Cybersecurity Center of Excellence, 2026. https://www.nccoe.nist.gov/sites/default/files/2026-02/accelerating-the-adoption-of-software-and-ai-agent-identity-and-authorization-concept-paper.pdf

[4] A. Chan, R. Salganik, A. Markelius, et al., "Harms from increasingly agentic algorithmic systems," *In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 651–666). Association for Computing Machinery, 2023. https://doi.org/10.1145/3593013.3594033

[5] P. Dhiman, N. Saini, Y. Gulzar, S. Turaev, A. Kaur, K. U. Nisa, and Y. Hamid, "A review and comparative analysis of relevant approaches of zero trust network model," *Sensors*, vol. 24, issue 4, Article 1328, 2024. https://www.mdpi.com/1424-8220/24/4/1328

[6] B. Fisher, and R. Galluzzo, "Back to the future: Why agentic AI needs a strong identity foundation," *NIST Cybersecurity Insights*, 2026. https://www.nist.gov/blogs/cybersecurity-insights/back-future-why-agentic-ai-needs-strong-identity-foundation

[7] E. Jain, "Auditing agentic AI workflows: How to control test when the system decides for itself," ISACA Industry News, 2026. https://www.isaca.org/resources/news-and-trends/industry-news/2026/auditing-agentic-ai-workflows-how-to-control-test-when-the-system-decides-for-itself

[8] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: A three-layered approach," *AI and Ethics*, vol. 4, issue 4, pp. 1085–1115, 2024. https://doi.org/10.1007/s43681-023-00289-2

[9] S. Mushtaq, M. Mohsin, & M. M. Mushtaq, "A systematic literature review on the implementation and challenges of zero trust architecture across domains," *Sensors*, vol. 25, issue 19, Article 6118, 2025. https://doi.org/10.3390/s25196118

[10] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: What it is and how it works," *AI & Society*, vol. 39, issue 4, pp. 1871–1882, 2024. https://doi.org/10.1007/s00146-023-01635-y

[11] E. Papagiannidis, P. Mikalef, & K. Conboy, "Responsible artificial intelligence governance: A review and research framework," *The Journal of Strategic Information Systems*, vol. 34, issue 2, Article 101885, 2025. https://doi.org/10.1016/j.jsis.2024.101885

[12] N. F. Syed, S. W. Shah, A. Shaghaghi, A. Anwar, Z. Baig, and R. Doss, "Zero trust architecture (ZTA): A comprehensive survey," IEEE Access, vol. 10, pp. 57143–57179, 2022. https://doi.org/10.1109/ACCESS.2022.3174679

[13] S. Vallor, and T. Vierkant, "Find the gap: AI, responsible agency and vulnerability," *Minds and Machines*, vol. 34, issue 3, Article 20, 2024. https://doi.org/10.1007/s11023-024-09674-0

[14] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, issue 2, Article 100211, 2024. https://doi.org/10.1016/j.hcc.2024.100211