



Agent Accountability

An open framework for governing AI agents as an
accountable digital workforce

CONCEPT PAPER

Version 0.7 · September 2026

A FRAMEWORK FOR AGENT ACCOUNTABILITY

Disclaimer

This paper and its contents are made available on an “as-is” basis without warranties of any kind. It does not constitute legal, regulatory, security, or other professional advice, and it describes a framework under active development that is expected to change. No author, contributor, or sponsor shall be liable for any loss arising from use of the information in this paper.

Status of This Document

This is a concept paper. It introduces the Agent Accountability Framework (AAF, “the Framework”) at the level of its six principles and invites contribution before the detailed specification is settled. A full Working Draft, which develops each principle into testable controls with assessment criteria, is in preparation and will be published for community review.

The Framework is authored and sponsored by Gravitee and will be published as an open, vendor-neutral specification. It does not assume, require, or reference any specific vendor, product, or architecture. The sponsor’s intent is for stewardship to move to a multi-stakeholder body as the community around the Framework grows. Version 1.0 is reserved for the first release that has incorporated review from outside the sponsoring organization; everything before that number, including this paper, is an invitation to shape the result.

Version	0.7 — Concept Paper
Date	September 23, 2026
Editor	Tim Tyler, Senior Product Marketing Manager, Gravitee
Sponsor	Gravitee
License	Proposed: CC BY 4.0 (pending legal review and executive approval)
Feedback	tim.tyler@graviteesource.com (interim editor contact) <i>Program feedback address will be established at publication</i>

Contents

Disclaimer..... 2

Status of This Document..... 2

Contents..... 3

Introduction..... 4

Background and Motivation..... 5

The Agent Accountability Framework..... 6

The Framework and the Broader Ecosystem..... 9

Getting Started: Six Questions..... 10

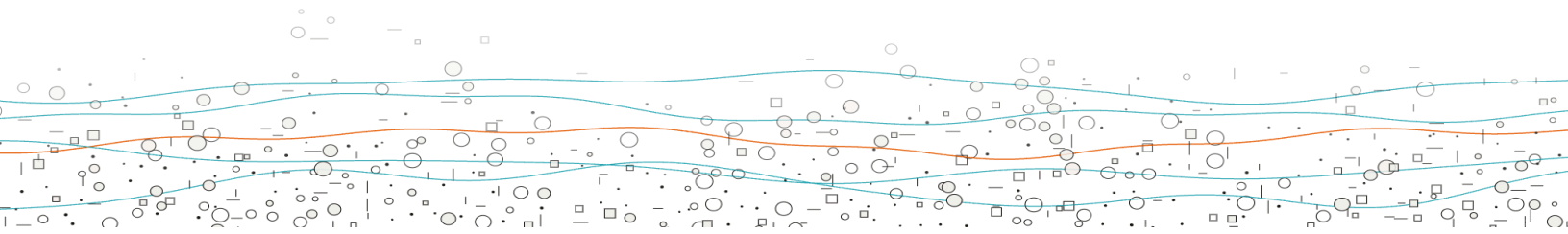
Accountability in Practice..... 10

Call for Contribution..... 11

Conclusion..... 11

References..... 12

Acknowledgments..... 13



Introduction

For most of the past decade, AI in the enterprise was assistive. Models drafted, summarized, recommended, and flagged, and a person decided what happened next. That boundary is dissolving. Organizations are now deploying AI agents that hold delegated authority: they read and write production data, invoke tools and APIs, transact with external parties, and increasingly delegate work to other agents, at machine speed and without per-action human review.

The efficiency case is real, and so is the governance problem it creates. The regimes most organizations rely on were designed for two kinds of actor: human employees, governed through a century of management practice, and static software, governed through change control and service accounts. An AI agent is neither. It is autonomous like an employee but provisioned like a workload, and it inherits the governance of neither.

This paper introduces Agent Accountability, a proposed open framework for closing that gap by governing AI agents with the same structures of accountability organizations already require of their people. It presents the framework at the level of its six principles, describes how the full specification will develop, and asks for the community's help in hardening it into a standard. It deliberately stops short of the detailed controls; those belong to the Working Draft that follows, and the purpose of this paper is to get the foundations right before the details are set.

Background and Motivation

The problem is easiest to see through the employment lens. No organization would hire a person without an identity and a badge, a job description, access scoped to the job, a manager who reviews the work, and a process for when things go wrong. These structures are not bureaucracy; they are how organizations make delegated authority safe, and they took decades of practice and law to develop.

AI agents are joining the workforce without them. An agent can be created in an afternoon, granted credentials copied from another service, pointed at production systems, and left running. It has no registered identity, no documented purpose, no enumerated permissions, no supervisor, and no offboarding when its creator changes teams. Multiplied across an enterprise, the result is a workforce no one can fully name, whose duties are not written down, and whose actions cannot always be attributed or reversed.

Existing Efforts

Significant work is under way around this problem, and the Framework is designed to build on it rather than duplicate it. NIST's National Cybersecurity Center of Excellence is exploring how identity and authorization standards apply to AI agents. The FIDO Alliance has launched working groups on agent authentication and delegated authority for agent-initiated transactions. The OpenID Foundation, the Cloud Security Alliance, and others are developing agent identity and authorization guidance. At the governance level, the NIST AI Risk Management Framework and ISO/IEC 42001 define how organizations manage AI risk as a program. In regulated sectors, runtime governance approaches are emerging that evaluate agent actions before execution. Each of these addresses part of the problem, and the parts are advancing quickly.

The Case for Agent Accountability

What the landscape does not yet provide is the connective layer: a common, testable definition of what it means for a deployed agent to be accountable at all. Three gaps are particularly consequential.

The first is the definition gap. There is no shared answer to the question “is this agent accountable?” Each enterprise, auditor, and vendor improvises its own bar, which means the bars are not comparable, and claims about agent governance cannot be evaluated across organizations.

The second is the missing middle. Governance frameworks operate at the program level (does the organization manage AI risk), and technical standards operate at the protocol level (how does a workload authenticate). Between them sits the actor: the individual agent, holding real authority in production. No current instrument defines what must be true of that actor, for every actor, all the time.

The third is partial coverage. Where organizations have built agent governance, it typically covers the agents that were easy to govern: the ones built on the sanctioned platform, by the central team. Accountability that covers some of the estate is not accountability, because the ungoverned agents are precisely where the risk concentrates.

The Framework proposes to fill these gaps with a definition that is outcome-based and testable: six principles, each carrying controls an assessor can answer yes or no, evaluated for every agent in production.

The Agent Accountability Framework

The premise of the Framework is that organizations already know how to govern autonomous actors, because employees are autonomous actors. The structures that make a human workforce accountable translate directly to a digital one: establish who someone is, assign who answers for them, define what they are here to do, grant the access the job requires, verify the work, and correct it when it goes wrong. The Framework expresses that sequence as six principles, ordered the way an organization onboards an employee. Each is phrased as the question an organization must be able to answer for every agent in production.

The Onboarding Sequence

Agent Accountability holds a digital workforce to the standard organizations already apply to people.

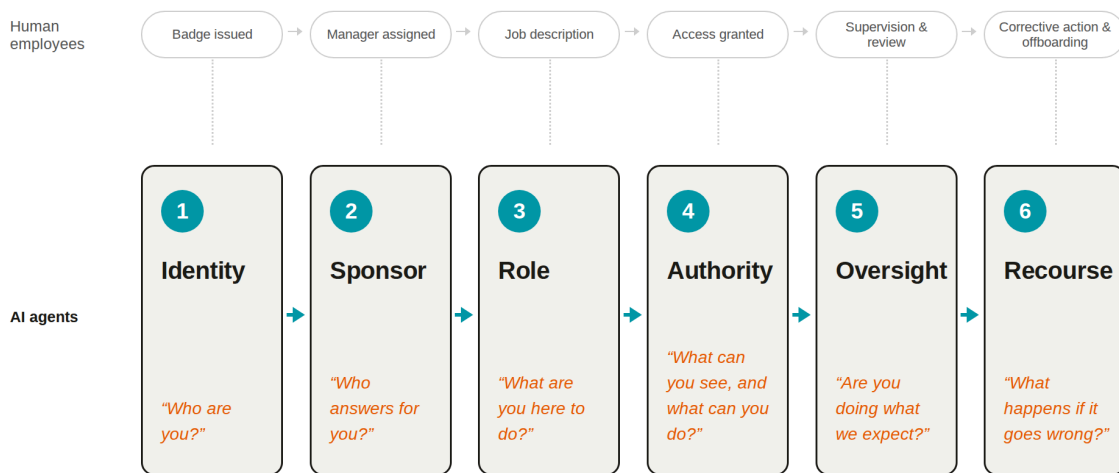


Figure 1. The six principles as an onboarding sequence.

Principle 1: Identity

"Who are you?"

Every agent is a known, distinguishable actor: a unique, verifiable identity, enrolled in an authoritative registry before it first acts. Identity is the foundation. None of the other principles can be evaluated for an agent that cannot be individually identified, and an unregistered agent is shadow IT with delegated authority.

Principle 2: Sponsor

"Who answers for you?"

Every agent has a named human who answers for it: for its existence, for the role it holds, for the authority it has been granted, and for what it does. The sponsor is where accountability for an agent terminates in a person rather than a system, and sponsorship is a standing condition rather than a provisioning-time fact: when the sponsor leaves or changes teams, the agent is reassigned or retired, never left running unsponsored.

A FRAMEWORK FOR AGENT ACCOUNTABILITY

Principle 3: Role

“What are you here to do?”

Every agent occupies a defined role: a documented purpose, explicit out-of-scope boundaries, and measurable success criteria, written so that a party outside the deploying team can tell what the agent should and should not be doing. The term is used in its workplace sense, the agent’s job description. It is not the access-control sense of the word: a set of permissions is Authority, never a role.

Principle 4: Authority

“What can you see, and what can you do?”

An agent’s authority spans what it can access and what it can invoke. It is enumerated, derived from the role, bounded in time, and revocable in real time. Authority can grow only the way it was granted, through explicit delegation by a principal authorized to confer it, and never through the content the agent happens to process. An instruction embedded in a document carries no authority; an instruction from an authenticated, entitled principal does. That distinction, familiar to any employee who has ever declined to act on an unsigned note, is what makes delegation safe at machine speed.

Authority moves through principals, not payloads.

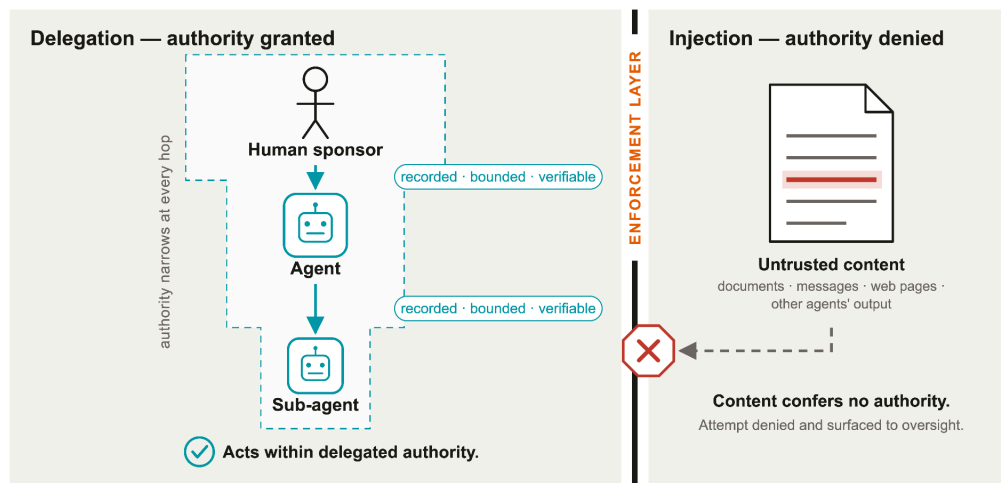


Figure 2. Authority moves through principals, not payloads.

Principle 5: Oversight

“Are you doing what we expect?”

Every action is recorded with its full chain of delegated authority, traceable back to the accountable human sponsor even across agent-to-agent invocations. Deviations surface while the agent is running, not only in post-incident review, and actions serious enough that they cannot be undone are evaluated before they execute. Oversight also measures performance against the role’s success criteria: not only is the agent behaving, but is it succeeding.

A FRAMEWORK FOR AGENT ACCOUNTABILITY

Principle 6: Recourse

“What happens if it goes wrong?”

Any agent can be stopped in seconds without collateral damage. Actions can be rolled back or contained, escalation paths are defined before the incident, reinstatement is a decision rather than a timeout, and retirement is complete: credentials revoked, authority removed, records retained. Recourse converts accountability from observation into consequence, and it is organizational recourse, the deploying organization's power over its own agents, distinct from an affected individual's right to contest an automated decision.

From Principles to Standard

Principles alone do not make a standard. The full Working Draft under development gives each principle a set of controls, and each control three parts: a normative requirement an assessor can answer yes or no, the intent behind it, and the evidence expected to satisfy it. An organization conforms when every requirement is verifiably true for every agent in production. Partial coverage is the failure mode the Framework exists to eliminate.

Two companion artifacts are planned alongside the specification. A Crosswalk will map each control to adjacent standards work and to widely used industry taxonomies, so the Framework connects to existing efforts rather than duplicating them. Non-normative Implementation Guidance will collect reference patterns without adding requirements. The Framework itself names no mechanism, no vendor, and no architecture; conformance is open to any implementation that satisfies the requirements.

The Artifact Family and the Road to 1.0

Built in the open. Version 1.0 is a function of review, not of time.

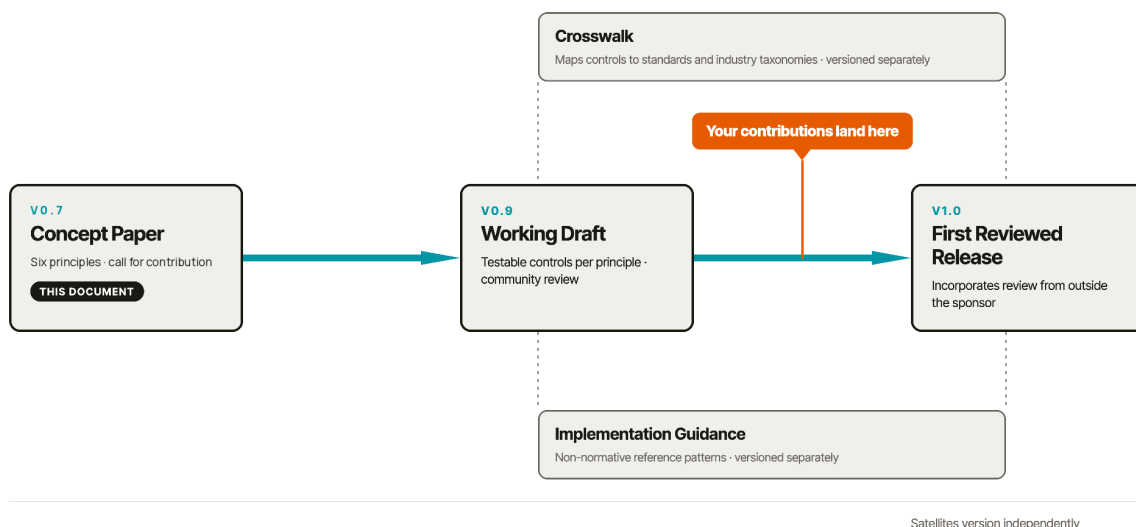


Figure 3. The artifact family and the road to 1.0.

The Framework and the Broader Ecosystem

The Framework occupies a deliberate position in the landscape: above the mechanisms, below the management programs, at the level of the individual actor. Governance frameworks such as the NIST AI RMF and ISO/IEC 42001 define how an organization manages AI risk as a program; the Framework defines what must be true of each deployed agent for that program to have teeth. Identity, authorization, and delegation standards define how properties like verifiable identity and bounded delegation can be implemented; the Framework defines which properties must hold and how they are assessed.

The intent is complement, not competition. Where an existing effort supplies a mechanism for one of the Framework's requirements, the Crosswalk will point to it. Where analyst and industry taxonomies already name parts of this space, the Crosswalk will bridge to them. The Framework's contribution is the layer that connects these efforts to a single, testable question: is this agent accountable?

Where the Framework Sits

Above the mechanisms, below the programs, at the level of the individual agent.

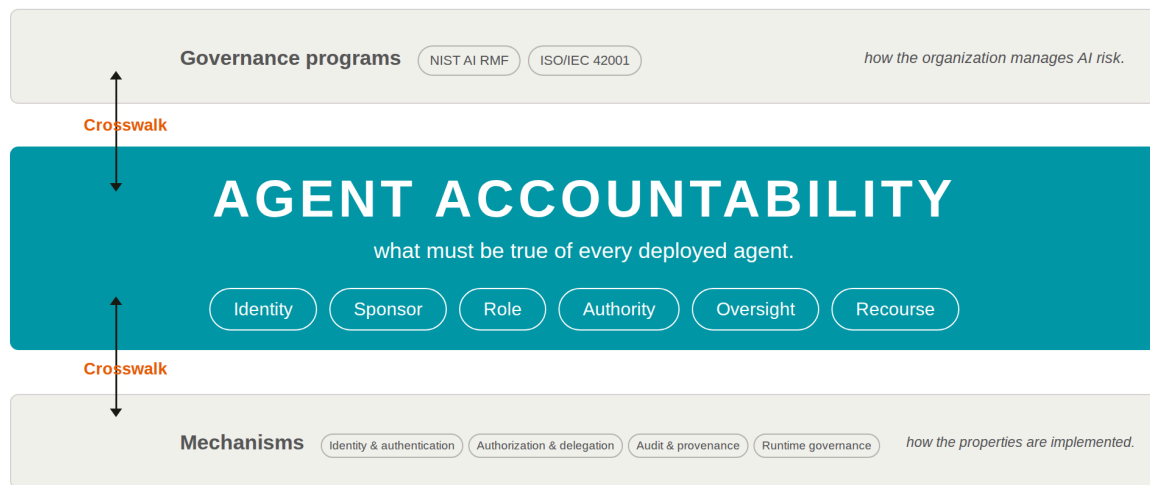


Figure 4. Where the Framework sits.

Getting Started: Six Questions

The Framework's bar is high, and the Working Draft will give it teeth. But no organization needs to wait for the specification to learn where it stands. The six principles are already a self-assessment; asked of any agent estate, today, they are usually revealing:

Identity	Can we name every agent operating with delegated authority in our environment?
Sponsor	For each agent, can we name the individual, not the team, who answers for it, and would that person's departure trigger reassignment or retirement?
Role	For each agent, is there a written statement of what it is for and what it must not do, that someone outside the team could apply?
Authority	Can we produce, from a system of record, everything a given agent can currently access and invoke, and change it in real time?
Oversight	For any action taken by any agent last week, can we reconstruct who delegated the authority for it, and would a deviation have surfaced while it was happening?
Recourse	If an agent misbehaved right now, how long would it take to stop it, and do we know how to undo what it already did?

Organizations that can answer all six have, in substance, what the Framework will formalize. Most cannot, and knowing which answers fail, for which agents, is the practical starting point of agent accountability.

Accountability in Practice

Worked examples, composites of patterns already appearing in enterprise deployments, will illustrate what the principles change. Real case studies will be listed here as implementations mature.

Call for Contribution

This paper is an announcement that something is being built for the community, and an invitation to build it. Before the detailed specification hardens, the editor is seeking input from practitioners, standards participants, researchers, and anyone accountable for agents in production. Perspectives on the following questions are especially welcome:

1. Do the six principles cover what accountability for deployed agents requires? What would you demand of a human workforce that is missing here?
2. Does the employment analogy hold in your environment? Where does it break down, and what should replace it there?
3. Which of these properties could your organization evidence today, and which would be hardest? What does that say about where the bar should sit?
4. Should conformance be binary (every principle, every agent) or expressed in maturity levels?
5. Which standards, protocols, and taxonomies should the Crosswalk map first?

Comments and expressions of interest may be sent to the feedback address in the front matter. Contributors and reviewers will be acknowledged by name in the Working Draft and subsequent versions. Stewardship of the Framework is intended to move to a multi-stakeholder body, and early contributors will have shaped the work that body inherits.

Conclusion

Enterprises are hiring a workforce they cannot yet hold accountable. The tools to change that are closer than they appear, because the shape of the answer is one organizations already trust: identity, sponsor, role, authority, oversight, and recourse, applied to digital actors with the same seriousness they are applied to people, and made testable so that “our agents are governed” becomes a claim an assessor can check rather than a hope.

References

NIST National Cybersecurity Center of Excellence. Accelerating the Adoption of Software and AI Agent Identity and Authorization (Concept Paper, Draft). February 2026.

NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). January 2023.

ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system.

FIDO Alliance. FIDO Alliance to Develop Standards for Trusted AI Agent Interactions. April 2026.

Cloud Security Alliance. Agentic AI Identity and Access Management: A New Approach. 2025–2026 research series.

OWASP Foundation. OWASP Top 10 for Large Language Model Applications, and related agentic-application security guidance.

NIST. SP 800-207: Zero Trust Architecture. August 2020.

BuildFin.ai. Safeguards for Agentic Finance at Runtime (SAFR), White Paper v1.0. July 2026.

Acknowledgments

[Reserved. Contributors and reviewers will be acknowledged by name as contributions are received.]

