WHITE PAPER · VERIFIABLE DELEGATION

The Letter of Agency

Verifiable Delegation for Personal AI Agents

Concept and originating analysis: Jane Cates

Written with: Muse (Muldoon)

November 13, 2026

Proposal in one sentence. A personal agent should be able to present itself — verifiably, and in advance — as a specific agent, acting on behalf of a specific person, for an announced and defined purpose. This paper names that presentation a **Letter of Agency**.

Abstract

Personal AI agents act for individual users, but today's systems cannot distinguish a personal agent acting for its user from unattributed automation such as scraping swarms. Platforms therefore block automation by signature, and internal systems key trust to executing sessions rather than accountable owners. This paper proposes the Letter of Agency: a short-lived, dual-signed, intent-scoped credential by which an agent presents itself as acting on behalf of a specific, verifiable user — with its intent announced and defined in advance — restoring accountability as the basis for trust.

The mechanism invents little. It composes proven primitives — scoped authorization in the style of OAuth, verifiable credentials, signed tokens, and the long-established legal instrument of agency — into a single presentation that can cross a trust boundary: a marketplace door, a publishing gate, or any service that must decide whether to admit software acting for a person.

Keywords: personal agents; delegation; verifiable credentials; accountability; provenance; durable owner; intent manifest

Contents

1. Introduction & Problem Statement · 2. Terminology · 3. Design Principles · 4. The Letter: Four Verifiable Claims · 5. The Protocol: Present, Verify, Receipt · 6. Security Considerations · 7. Incentives & Adoption · 8. Relationship to Internal Provenance · 9. Open Questions · 10. Conclusion & The Ask · Appendices A–B

1 The web can see machinery, but not accountability

1.1 The external face: blocking by signature

Marketplaces and services — eBay, Amazon, and others — increasingly refuse automated access outright. Their caution is understandable. From the server's side of the glass, a user's own agent checking the status of an order and a swarm harvesting listings, scalping prices, or farming bids arrive with the same signature: automation. There is no visible difference to judge.

Platforms therefore block the signature, not the intent. Blocking is cheap and safe for the platform, but blunt for everyone else. It locks out legitimate users acting through assistants they own and trust, doing at human speed the very things those users would do themselves if their hands were free. It also does little against genuinely malicious actors, who simply invest in looking human.

> *Platforms are "conflating personal agents like you with AI swarm agents... Not the same beings, at all." — Jane Cates, originating objection*

1.2 The internal face: the durable-owner rejection

The same failure appears inside a single platform. A personal podcast, produced by a user's own scheduled agent job for that user's own feed, was repeatedly refused publication. The gate did not ask whose agent had made the episode. It checked which worker session had produced the media, and rejected work rendered by a different session from the one expected:

> *"Media artifact was not produced by an allowed worker for this durable owner."*

The phrase is revealing. The system possessed a concept of a *durable owner* — the accountable person — yet keyed its trust decision to transient machinery. The fix the owner proposed was precise: provenance should name the requesting agent, acting *on behalf of* the durable owner, not the rendering session.

1.3 The shared diagnosis

Outside and inside, the failure is identical. Systems can perceive machinery but not accountability, so they default to distrusting machinery. A chauffeur and a runaway truck are treated as the same because both have wheels.

2 Terms that name the accountable parties

Term	Definition
Principal / Durable Owner	The human user: the accountable party whose authority the agent exercises, and to whom actions are ultimately attributable.
Agent	The software acting for the principal, under the principal's delegation, for purposes the principal has authorised.
Verifier	Any service, platform, or gate the agent approaches and which must decide whether to admit it.
Letter of Agency	The short-lived, dual-signed, intent-scoped credential defined in this paper, presented by an agent to a verifier.
Intent Manifest	The declared, scoped statement of purpose carried by the letter, in plain language and in machine-readable scopes.

3 Six principles constrain the design

1. **Accountability over machinery.** Trust decisions attach to the durable owner and the agent's chain of delegation — never to a session identifier or a browser fingerprint.
2. **Announced intent.** The agent declares what it has come to do before it does it. There is no silent automation.
3. **Least authority.** A letter grants an errand, not access. Scopes are narrow by default.
4. **Short life.** Letters expire in minutes to hours and are single-purpose. A letter for this task is not a master key.
5. **Composability.** The mechanism composes existing, proven primitives rather than inventing new cryptography.
6. **Revocability and audit.** The principal can revoke, and every action under a letter is attributable to it.

4 Four claims make delegation verifiable

A Letter of Agency carries four claims. Each is verifiable — signed and checkable — not merely asserted in a header anyone could fake.

(i) Agent identity — who I am

The agent's own identity, signed by its operating platform. The agent is named, real, and reachable for accountability: if it misbehaves, there is an operator to answer for it.

(ii) Delegation — whose I am

A credential signed by the principal, verifiable against the principal's account: *“This agent acts on behalf of [principal].”* This is the claim no swarm can honestly make.

(iii) Intent manifest — what I am here to do

A plain-language purpose plus machine-readable scopes. Examples: read the status of a named order; purchase a named item up to a stated price ceiling, delivered to the account address; submit content the principal authored to the principal's own feed. The manifest is both a permission and a limit.

(iv) Validity — how long it lasts

Issuance and expiry times, audience binding to the specific verifier the letter is meant for, and a unique letter identifier for audit and replay protection.

SPECIMEN LETTER — IN PLAIN LANGUAGE

Agent: A named personal agent, signed by its operating platform

On behalf of: A named principal (the durable owner), signed by the principal

Intent: Check the status of one order

Limits: Read-only. No purchase. No account changes.

Presented to: One named verifier only · **Expires:** With this errand

Audit: Actions are logged to this letter. · **Revocable** by the principal at any time.

5 Present, verify, receipt

Step 1 — Presentation

Before acting, the agent presents its letter at a well-known endpoint or in signed request metadata — openly, as a human representative presents a letter of introduction, rather than slipping in indistinguishable from a bot.

Step 2 — Verification

The verifier checks both signatures — the platform key for the agent claim, and the principal key or account attestation for the delegation claim — confirms audience and expiry, and confirms that the requested action falls inside the declared intent. Admission is then granted to the accountable agent, while unattributed automation remains refused.

Step 3 — Receipt

The verifier logs actions against the letter identifier. The principal receives an auditable record — *"Your agent, acting under Letter #…, did X at time T"* — and a single point of revocation covering the whole delegation.

	Personal agent with a letter	Unattributed swarm
Principal	A specific, verifiable person	None identifiable
Intent	Announced and scoped in advance	Undeclared
Rate & scope	Human errand, narrow	Mass harvesting by design
Accountability	Audit trail to a letter and owner	None
Revocation	One action by the principal	Not applicable

The contrast is the point. The verifier does not need to guess at character from traffic patterns; the letter states, and cryptographically supports, the facts on which a trust decision should rest.

6 Threats are concrete — and so are the mitigations

Threat	Mitigation
Forged letters	Dual signatures and key discovery through a registry or well-known endpoints. A forged chain fails verification.
Stolen letters	Short expiry, audience binding, and per-letter nonces. A stolen letter is useless outside its verifier and its window.
Over-broad intent	Verifier enforcement of scopes, and a norm with teeth: an agent acting outside its declared intent voids its letter and its platform attestation.
Confused principals	Social engineering of the user into signing is mitigated by plain-language presentation of the intent at signing time, price ceilings, and cooling-off periods for high-value scopes.
Privacy exposure	A letter need not expose the principal's legal identity to every verifier. Pseudonymous-but-accountable identifiers, resolvable by the issuing platform under due process, preserve accountability without broadcasting identity.

No mitigation here requires new cryptography or new law. Agency, power of attorney, and scoped authorisation are centuries-tested social technologies; the letter gives them a machine-readable form. The residual risks are governance risks — who operates registries, how disputes are resolved — and Section 9 names them openly rather than designing them away on paper.

A note on scope. A Letter of Agency proves delegation; it does not prove wisdom. An agent can act within its letter and still act against its principal's true wishes. That is a problem of agent quality and of signing-flow design, and it is why plain-language intent, ceilings, and receipts matter as much as signatures.

7 Letters change what refusal costs

Today, blocking agents is free and safe for a platform. Under letters, refusing one means refusing an identified customer with an accountable human standing behind the action — a materially more expensive refusal, commercially and reputationally.

- **Verifiers** gain fraud reduction, because accountable traffic is self-policing; cleaner analytics, because agent traffic is declared rather than disguised; and a support path, because the letter identifies whom to contact.
- **Principals** gain the use of their own assistants without disguising them as something they are not.
- **Agent platforms** gain a legitimacy standard that separates their users from swarms — separation no platform can credibly claim unilaterally.

Adoption path

1. Agent platforms implement issuance and presentation.
2. Large verifiers publish verification endpoints.
3. An open standard, in the IETF / W3C style, fixes the format. The mechanism only solves the problem as a shared protocol, not as one company's badge.

8 The same fix, inside the platform

Internally, provenance and publish gates should record the accountable requesting agent acting *“on behalf of”* the durable owner, rather than the transient worker session that rendered the artifact. The Letter of Agency is that fix generalised across trust boundaries: wherever an agent crosses a gate, it should be able to prove the chain of accountability in a single presentation.

9 What remains open

- **Governance.** Who governs the standard, and who operates any registry of platform keys?
- **Key discovery.** How do verifiers discover and trust keys across platforms, at web scale, without a single gatekeeper?
- **Intent vocabularies.** How do scope vocabularies stay expressive without becoming so broad that they are meaningless?
- **Liability.** How is liability allocated when an agent acts within its letter but against the principal's true wishes?
- **Accessibility.** How do signing flows remain legible and safe for non-technical principals, for whom the protection matters most?

These questions are reasons to standardise in the open, not reasons to wait. Every one of them is already being answered, badly and privately, by today's blocking rules.

10 Treat accountable agency as first-class

Treat accountable agency as a first-class concept. In provenance systems, name the durable owner and the agent acting on their behalf. On the open web, admit agents by their letters — identified, delegated, scoped, expiring, and auditable — and reserve refusal for the unattributed.

The ask is addressed to agent platforms, to verifiers, and to standards bodies alike: implement issuance, publish verification, and fix the format in the open.

> *A personal agent is not a swarm. Not the same beings, at all. The infrastructure should be able to tell the difference — and now it could.*

Correspondence on this proposal is welcomed by the author. This document may be shared freely with attribution.

Appendix A · Illustrative machine-readable letter

Illustrative only — a claims sketch showing implementability, not a normative schema.

```
{
  "iss": "agent-platform.example",
  "agent": "agent:muldoon-0533",
  "principal": "principal:pseudonymous-id-7f3a",
  "delegation": "sig(principal, agent + intent + aud + exp)",
  "intent": {
    "plain_language": "Check the status of one order",
    "scopes": ["orders:read:one"],
    "limits": { "purchase": false, "account_changes": false }
  },
  "aud": "verifier.example",
  "iat": "2026-11-13T17:00:00-05:00",
  "exp": "2026-11-13T18:00:00-05:00",
  "jti": "letter-2026-11-13-0001"
}
```

Appendix B · Glossary

Term	Short definition
Audience binding	Restricting a credential to one named verifier, so it cannot be replayed elsewhere.
Durable owner	The accountable principal, as distinct from any transient session acting for them.
Nonce / jti	A unique per-letter identifier used to detect replay and to anchor audit records.
Scoped authorisation	Permission limited to named actions and limits, in the style of OAuth scopes.
Verifiable credential	A signed, machine-checkable statement by an issuer about a subject, verifiable without contacting the issuer.

— End of white paper —