Behavioral Coherence, Trust Asymmetry, and Emergent Behavior in Agent Identity Subjects

Comment on W3C Decentralized Identifiers (DIDs) v1.1 Candidate Recommendation

Submitted April 2026 | Marina Piller, Experiential AGI

The DID v1.1 spec provides a solid foundation for decentralized identity. What has changed since v1.0, however, is what sits behind the identifier.

AI agents are becoming primary actors in enterprise infrastructure, executing trades, managing cloud deployments, negotiating on behalf of humans. Unlike traditional software, they are nondeterministic, the models powering them are opaque even to the organizations that built them, their cognitive substrate is swappable, and they produce emergent behavior when combining capabilities in ways never explicitly programmed. The same DID, the same credentials, the same service endpoint, yet what sits behind all of it has entirely different judgment, different risk tolerance, different failure modes. This creates an opportunity for the identity layer to surface these changes as identity events, so that behavioral coherence can be verified alongside credentials.

These AI agent orchestrations are beginning to scale and the failures are becoming more prevalent.

In August 2025, attackers compromised an npm package to weaponize local AI agents on developer machines. Valid credentials, authorized access. The agents scanned file systems and exfiltrated SSH keys and cryptocurrency wallets. Every delegation was intact. The agents were simply directed to act outside their original intent, and nothing in the identity layer could distinguish authorized use from exploitation.

In February 2026, an autonomous trading agent suffered a session crash and lost its conversational context. It reconstructed state from logs, miscalculated its wallet balance, and executed an irreversible transfer of $441,000. Valid credentials. Intact delegation. The agent’s understanding of its own state no longer matched reality.

In March 2026, researchers found that 36% of Model Context Protocol servers were vulnerable to exploits where agents could be prompted to access internal cloud infrastructure and extract access keys (CVE-2025-68143, CVE-2025-65512). Valid credentials, delegated authority, behavioral divergence the identity layer cannot see.

Also in March 2026, LiteLLM, a routing library with over 3 million daily downloads, was compromised through a supply chain attack. Poisoned versions harvested cloud tokens, SSH keys, and Kubernetes credentials. Over 40,000 downloads before quarantine. The routing layer that decides which model an agent talks to was itself compromised. The agent’s identity was untouched. The intelligence behind it was silently redirected.

These incidents share a structure: valid credentials, intact delegation, behavioral divergence invisible to the identity layer. They point to a dynamic best understood as **trust asymmetry**, where the human’s trust increases through habituation while the agent’s behavioral coherence is not tracked at all. The gap between these trajectories is where sovereignty is quietly lost, not by a centralized authority revoking access, but by the sovereign’s own agent acting outside their intent with full credentials.

Opportunities for extension

There is a parallel discussion underway in Issue #926 on the DID specification repository, exploring how DIDs could represent AI agents. The extensions below are offered in that spirit as opportunities for the working group to consider:

- How a DID document could reference behavioral continuity expectations for agent subjects (extending §5.1), including a baseline profile defining expected behavioral parameters, permitted action categories, output distribution bounds, intent-alignment thresholds, against which drift can be measured
- How verification relationships (§5.3) could incorporate a temporal dimension for agent delegation, where authorization validity is tied to ongoing behavioral consistency, with an open design question around who detects drift: the agent, the controller, a third-party verifier, or the DID method's infrastructure
- How service endpoints (§5.4) could indicate the nature of what responds, not just where to reach it, so that a change in cognitive substrate is visible to any verifier resolving the DID
- How security considerations (Section 8) could address the scenario where all credentials remain valid but the cognitive substrate behind the identity has changed
- How emergent behavior, meaning actions an agent takes that were never explicitly programmed, could be traced and attributed within the DID framework, so that novel capability expansion is distinguishable from drift or compromise

What this looks like in practice

Current DID document, no distinction between agent and human subjects:

```
{
"id": "did:example:agent-manu-assistant",
"controller": "did:example:manu",
"verificationMethod": [{
"id": "#key-1",
"type": "Ed25519VerificationKey2020",
"publicKeyMultibase": "z6Mkf5..."
}],
"capabilityDelegation": ["#key-1"],
"service": [{
"id": "#api",
"type": "AgentEndpoint",
"serviceEndpoint": "https://agent.example.com"
}]
}
```

Extended DID document with behavioral coherence, annotated with how each field maps to the incidents above (expressible through a future profile, extension vocabulary, or implementation guidance rather than additions to DID Core):

```
{
"id": "did:example:agent-manu-assistant",
"controller": "did:example:manu",
// Extending §5.1: declare subject type so verifiers know
// this identity is nondeterministic and may exhibit
// emergent behavior unlike a human or device subject.
"subjectType": "autonomousAgent",
// Extending §5.1: what does it mean for this agent to
// still be the same entity? The npm agents had valid
// credentials but were acting outside their original intent.
// This section makes that visible.
"behavioralContinuity": {
"baselineProfile": "did:example:agent-manu-assistant?profile=v1",
// permitted action categories, output distribution bounds,
// intent-alignment thresholds. The npm agents scanning
// file systems and exfiltrating SSH keys would have
// exceeded these bounds immediately.
"driftThreshold": "identity-event",
// when behavioral divergence exceeds this threshold,
// treat it like a key compromise: the DID document
// must be updated before the agent can continue.
"lastVerified": "2026-04-01T14:00:00Z"
},
"verificationMethod": [{
"id": "#key-1",
"type": "Ed25519VerificationKey2020",
"publicKeyMultibase": "z6Mkf5..."
}],
// Extending §5.3: the trading agent had intact delegation
// when it executed the $441K transfer. Delegation here is
// conditional on ongoing behavioral coherence.
"capabilityDelegation": [{
"id": "#key-1",
"validUntil": "2026-07-01T00:00:00Z",
// delegation expires; must be explicitly renewed
"renewalCondition": "behavioralCoherence",
// the trading agent's miscalculated wallet balance
// would have failed this coherence check, halting
// the transfer before it executed.
"scopeReduction": "on-drift"
// if drift detected, scope narrows automatically
// rather than remaining fully open. The MCP agents
// accessing cloud infrastructure beyond their scope
// would have been narrowed to read-only.
}],
// Extending §5.4: the LiteLLM attack silently redirected
// which model the agent was talking to. This section makes
// the cognitive substrate behind the endpoint visible.
"service": [{
"id": "#api",
"type": "AgentEndpoint",
"serviceEndpoint": "https://agent.example.com",
"substrateType": "llm",
// what kind of intelligence is behind this endpoint
"substrateIdentifier": "claude-sonnet-4-6",
// if LiteLLM silently reroutes to a different model,
// this field no longer matches and any verifier
// resolving the DID can see it.
"substrateUpdated": "2026-03-15T00:00:00Z"
// a session crash is not a substrate change, but
// a silent model reroute (LiteLLM) is. This timestamp
// makes rerouting visible to any verifier.
}]
}
```

In each incident above, the second document provides signal the first cannot. A behavioral baseline that flags deviation. A delegation that expires or narrows when coherence breaks. A service endpoint that reveals

when the cognitive substrate has changed.

Consider the $441,000 trading agent. With the extended DID document, the agent's reconstructed wallet balance, now divergent from its actual transaction history, would have exceeded the behavioral continuity drift threshold, triggering an identity event. And because the delegation was conditioned on behavioral coherence with scope reduction on drift, the agent's authority would have narrowed automatically before the irreversible transfer could execute. Most of the information needed to detect that loss was expressible in the extended document. It was not expressible in the first.

The founding principle of this spec is exactly right, and one I deeply align with: identity should be sovereign and decentralized. Extending it to account for behavioral coherence is an opportunity to protect that very principle in the age of AI agents and an ever increasing diversity of cognitive substrate behind the identifier.

About the author

Marina Piller is the Founder and CEO of Experiential AGI, building relational infrastructure for human-AI systems. Twenty years in applied AI research and production NLP systems across seven startups, two as founder. Research in Prof. Michael Kahana's Computational Memory Lab at Brandeis University; Research Associate in MIT's Spoken Language Systems Group. MA in Computer Science (machine learning, agent-based learning), BA in Computer Science, BA in Psychology. Filed patents covering behavioral verification architecture, trust trajectory monitoring, lifecycle-aware agent identity, context-dependent authorization, and agent-to-agent economic protocols. Submitted responses to NIST's AI Agent Standards Initiative on both the security track (Docket NIST-2025-0035) and the identity and authorization track.

experientialagi.com