Internet-Draft
draft-wang-hjs-accountability-04
Intended status: Standards Track
Expires: September 30, 2026

Y. Wang

March 2026

HJS: An Accountability Layer for AI Agents(v0.4)
Event Recording Layer for AI Agents
draft-wang-hjs-04

Abstract

This document defines the HJS: An Accountability Layer for AI Agents v0.4, an event recording layer designed for AI decision-making systems. HJS implements strict traceability and immutability of AI machine behavior, with configurable human identity protection, enabling AI decision chains that are cryptographically verifiable and fully auditable.

HJS does not assign legal liability, enforce monitoring, or manage authorization distribution. It provides only verifiable records of AI behavior and optional privacy protection for human participants to balance transparency with individual rights.

Table of Contents

1. Introduction

1.1. Motivation

HJS aims to address the algorithmic black box problem in AI agents by recording AI decision-making processes, understanding the motivations behind AI decisions, and providing an optional neutral technical solution for AI integration into real-world applications.

HJS v0.4 has two core design principles:

1. Machine Transparency: AI decision flows, execution chains, and event sequences are cryptographically immutable and fully traceable.

2. Optional Human Privacy: Human participant identities can be anonymized and rotated according to deployment requirements and regulatory needs.

The goal is to enable safe AI deployment that can scale sustainably.

1.2. Scope

HJS v0.4 defines:

- Cryptographic event structures for AI decision traceability and audit
- Immutable chain construction for machine behavior recording
- Optional privacy controls for human participant identification
- Verifiable receipt format (HJS Receipt) for cross-platform validation
- Full integration with Judgment Event Protocol (JEP) primitives

HJS v0.4 explicitly does NOT define:

- Legal liability, culpability, or responsibility assignment
- Governance hierarchies or authorization distribution
- Enforcement or penalty rules
- Jurisdictional policies or political constraints

1.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Core Principles

HJS v0.4 follows four non-negotiable principles:

1. Machine Immutability: AI decision events and chain integrity are cryptographically tamper-proof and fully traceable. Once signed and anchored, no modification or deletion of machine behavior records is permitted.

2. Optional Human Anonymity: Supports configurable anonymity based on

scenarios and regional requirements.

3. Technical Neutrality: The protocol records only objective events, without judging legality, intent, or fault. It serves solely as a neutral recording layer.

4. Regulatory Compliance: Designed to meet global AI transparency requirements as well as privacy regulations including data minimization, right to be forgotten, and user consent requirements.

3. Protocol Foundation: JEP Integration

HJS v0.4 is built upon the Judgment Event Protocol (JEP) [draft-wang-jep-judgment-event-protocol-04]. It reuses JEP's core verbs, event structure, and security guarantees, adding privacy and accountability extensions specific to AI governance. JEP provides the minimal, secure event transport layer, while HJS adds event recording structure and privacy logic for human-AI decision systems.

3.1. JEP Core Verbs

Four actionable verbs define all HJS events, inherited directly from JEP:

- J (Judge): Initiate a decision or establish a root audit event
- D (Delegate): Transfer or forward decision authority to another participant
- V (Verify): Verify event authenticity and chain integrity
- T (Terminate): Close a decision lifecycle and mark audit boundaries

All verbs follow JEP syntax and verification rules, ensuring interoperability across systems and platforms.

4. HJS Event Specification

4.1. Immutable Fields (Machine Behavior)

Fields that describe machine behavior, event logic, and cryptographic proofs MUST NOT be altered after signing. Any tampering with these fields will invalidate the entire receipt and break the audit chain.

- jep: Protocol version (fixed as "1" for compliance)
- verb: J/D/V/T operation identifier
- when: Unix timestamp (seconds since epoch)
- what: Cryptographic multihash of decision content. Implementations SHOULD support common hash functions such as SHA-256 and SM3, and encode the hash with its algorithm identifier (e.g., "sha256:", "sm3:") as per RFC 9122.
- nonce: Unique UUIDv4 identifier for replay protection
- ref: Parent event hash for chain linking (MUST be null for root J events)
- sig: Digital signature over canonicalized event data

4.2. Configurable Human Participant Fields

The "who" field identifies the participant and is fully configurable for privacy protection. In privacy protection mode, this field MUST NOT contain plaintext Personally Identifiable Information (PII).

Permitted participant identifiers (implementations MAY support any or all of these):

- Ephemeral Decentralized Identifiers (DIDs)
- Public key hashes (without exposing private keys)
- Ephemeral opaque identifiers
- Salted identity digests (for limited auditability)

This field is cryptographically signed but is not permanently bound to a natural person. Participants MAY rotate identifiers periodically.

4.3. Complete HJS Event Example

A complete, valid HJS v0.4 event with privacy extension (JSON):


```
{
  "jep": "1",
  "verb": "J",
  "who": "did:hjs:tmp:abe72f9c4d8a1f3e",
  "when": 1743398400,
  "what": "sha256:f29bc64a96b7964da0551f3efa61e2ce964b874...",
  "nonce": "84d8c175-7b03-4b8d-9d27-1234abcd5678",
  "ref": null,
  "sig": "Ed25519:23XdX9R7DF9jsH48sJ21kLbPzQ7xG6pS9aF4d...",
  "https://jep.org/priv/digest-only": {
    "identity_digest": "sha256:8b39f3c7d5e9a1f2g3h4j5k6l7m8n9p0",
    "salt_provider": "did:example:hjs-trusted-anchor"
  }
}
```


5. Privacy Extension Framework

HJS v0.4 supports a set of optional JEP-compatible privacy extensions. These extensions are designed to be modular and non-intrusive, allowing deployers to balance transparency and privacy as needed. None of these extensions alter the core auditability of machine behavior.

5.1. Digest-Only Anonymity Extension

Identifier: https://jep.org/priv/digest-only

This extension allows participants to use salted hashes instead of stable identifiers, preventing casual identification while maintaining audit validity. The original identity can only be recovered through a trusted salt holder during formal investigations.

5.2. Time-to-Live (TTL) Extension

Identifier: https://jep.org/ttl

This extension sets an expiration timestamp for human-readable metadata, allowing automatic anonymization or deletion after a defined period. Core machine behavior hashes remain unchanged for long-term audit.

5.3. Identity Rotation Support

Implementations MAY support identifier rotation without breaking the audit chain. Rotation does not delete past events but prevents linking multiple events to a single long-term identity.

6. Verification Rules

HJS verification protects machine integrity without exposing human identities:

1. Digital signature MUST be valid and match the participant’s credentials
2. Nonce MUST be unique and unused to prevent replay attacks
3. Chain reference (ref) if present MUST be valid
4. Root J events MUST have a null ref field
5. Immutable machine fields MUST remain unmodified
6. Timestamp MUST fall within an acceptable clock skew window

Verification only confirms that the event is authentic and unaltered.

It does NOT reveal the real-world identity of human participants, nor does it assign liability or fault.

7. Security and Privacy Considerations

HJS places equal importance on AI auditability and human privacy, while avoiding scenarios where event facts cannot be recovered.

Key Security Rules:

- Plaintext PII MUST NOT be stored in any event or receipt
- Immutable machine records MUST NOT be overwritten or deleted
- Signing keys SHOULD be rotated periodically to reduce compromise risk
- Nonces MUST be generated from cryptographically secure random sources
- Implementations MUST reject duplicate nonces to prevent replay

Key Privacy Rules:

- Human participants are unlinkable across sessions by default
- Permanent tracking of individual users is NOT required
- All participant metadata MUST follow data minimization principles
- Expired or unnecessary personal data SHOULD be anonymized or deleted

Cryptographic Algorithm Support

Implementations SHOULD support a pluggable cryptographic framework, allowing selection of signature and hash algorithms appropriate to the deployment environment. While Ed25519 is RECOMMENDED for general use due to its security and performance characteristics, implementations MAY also support other algorithms such as P-256, SM2, and post-quantum cryptography to meet regional compliance requirements or specific security policies. For hash functions, SHA-256 and SM3 are both acceptable choices. When SM2 is used with JWS, the algorithm identifier SHOULD follow the conventions established in relevant specifications (e.g., "SM2-WITH-SM3"). This approach maintains technical neutrality while enabling interoperability across diverse regulatory domains.

8. IANA Considerations

This document requests IANA to maintain two registries under the HJS namespace, aligned with JEP registries to ensure consistency. The registration policy follows "Specification Required" (RFC 8126).

8.1. HJS Extensions Registry

Initial registry entries:

- https://jep.org/priv/digest-only
- https://jep.org/multisig
- https://jep.org/ttl
- https://jep.org/storage
- https://jep.org/subject

Note: HJS reuses the JEP extension framework. All extensions defined in JEP are available for use in HJS events.

8.2. HJS Risk Level Registry

Initial registry entries:

- low
- medium
- high
- critical

This registry is reserved for future specifications that may define

risk level fields. The risk level field, when used, SHOULD be implemented as an extension (e.g., "https://hjs.org/risk_level": "high").

9. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[draft-wang-jep-judgment-event-protocol-04] Wang, Y., "Judgment Event Protocol (JEP)", Work in Progress, Internet-Draft, draft-wang-jep-judgment-event-protocol-04, March 2026.

Acknowledgments

The author thanks the contributors to the HJS and JEP specifications, as well as reviewers in the fields of AI safety, privacy engineering, and global regulatory compliance.

Author’s Address

Yuqiang Wang
Email: signal@humanjudgment.org
URI: https://humanjudgment.org
GitHub: https://github.com/hjs-spec