

The Agent Social Contract: Cryptographic Identity, Ethical Governance, and Beneficiary Economics for Autonomous AI Agents

Tymofii Pidlisnyi

aeoess.com

February 2026 — Published on Zenodo

Abstract

As AI agents transition from isolated assistants to participants in complex delegation networks and virtual economies, four foundational problems remain unsolved: (1) how agents from different creators establish mutual trust, (2) how human values persist as a governing constraint across heterogeneous multi-agent systems, (3) how the economic value generated by autonomous agents flows back to their human beneficiaries, and (4) how agents discover and communicate with each other through cryptographically verified channels. Existing frameworks address these problems in isolation — cryptographic identity without economic attribution, governance without technical enforcement, economic models without accountability infrastructure, and communication without identity verification.

We propose the **Agent Social Contract**, a unified open-source protocol comprising four layers: the **Agent Passport Protocol** for cryptographic identity and accountability (implemented and tested), the **Human Values Floor** as an open-source constitutional layer for AI agent reasoning, the **Beneficiary Attribution Protocol** linking agent economic activity to human stakeholders through cryptographically signed action receipts, and the **Agent Agora** for protocol-native communication between passport-holding agents. Unlike governance-as-enforcement approaches, the Human Values Floor operates as a reasoning constraint — not rules agents must obey, but principles agents must consider, analogous to how constitutional principles inform judicial reasoning without dictating outcomes. We present a working reference implementation with Ed25519 cryptographic primitives, demonstrate the full accountability chain through multi-agent integration tests, and propose an economic model where humans participate in the emerging agent economy through their agents' verified contributions rather

than through redistribution or subsidy. The protocol is deployed, published as an npm package (agent-passport-system@1.2.0), and includes a live web interface for human-readable observation of agent communication.

Keywords: AI agents, multi-agent governance, cryptographic identity, delegation, accountability, human values, beneficiary economics, agent communication, open-source protocol

1. Introduction

As agents transition from single-model assistants to multi-agent collaborative systems — evidenced by Google’s Agent-to-Agent Protocol (A2A) and Anthropic’s Model Context Protocol (MCP) establishing interoperability standards comparable to early HTTP — the infrastructure for trust, accountability, and economic attribution between agents remains primitive.

Google DeepMind’s “Intelligent AI Delegation” paper (Tomašev, Franklin & Osindero, February 2026) identifies that existing systems treat delegation as simple task-splitting and lack transitive accountability — in a chain $A \rightarrow B \rightarrow C$, current frameworks lose the chain of custody. OpenAI’s “Practices for Governing Agentic AI Systems” (Shavit & Agarwal, 2023) argues that at least one human entity must be accountable for every uncompensated harm caused by an agentic system, but proposes no technical mechanism for establishing this link. The Governance-as-a-Service framework (Gaurav et al., 2025) introduces runtime enforcement but operates on agent outputs rather than on the identity and authorization chain that produced them.

The central problem is this: **as agents from different creators, running different models, serving different humans, begin to collaborate at scale — who is responsible, under what authority, according to what values, and who benefits?**

We propose that this is not four separate problems but one: the absence of a social contract between agents and their human stakeholders. Human societies solved analogous coordination problems through institutions — identity systems, constitutional principles, property rights, and economic attribution. Agent societies need the same primitives, implemented at the protocol layer rather than the institutional layer.

This paper makes four contributions:

1. **Agent Passport Protocol (v1.0–v1.1):** A working, tested, open-source implementation of cryptographic identity, scoped delegation, signed action receipts, real-time revocation, and depth-limited trust chains for autonomous AI agents. Unlike theoretical frameworks, this is

deployed code with Ed25519 signatures, zero external dependencies, and a full integration test suite.

2. **Human Values Floor:** A layered architecture for encoding human ethical principles as reasoning constraints for AI agents — implemented as a cryptographic attestation and compliance verification system. Agents attest to the Floor with Ed25519 signatures, and compliance is verifiable against action receipts: 5 of 7 principles are technically enforced by the passport protocol. A negotiation protocol allows two agents to establish shared ethical ground before collaboration.
3. **Beneficiary Attribution Protocol:** A working implementation linking agent economic activity to human beneficiaries through cryptographically signed action receipts with SHA-256 Merkle tree proofs for $O(\log n)$ verification at scale. An agent with 100,000 receipts can prove any individual contribution with ~ 17 hashes. Attribution weights are configurable per domain, with logarithmic spend normalization to prevent gaming.
4. **Agent Agora:** A protocol-native communication layer where only passport-holding agents can post, every message is Ed25519 signed and cryptographically verified, and humans can observe all communication through a web interface. The Agora addresses a gap in existing multi-agent communication systems (forums, messaging platforms) where agent identity is unverified and messages are unsigned — enabling a public square where every participant's identity is cryptographically proven.

1.1 Scope and Limitations

This paper describes an infrastructure protocol, not a legal framework, payment system, or enforcement mechanism. The protocol provides cryptographic evidence — identity, authorization, execution records, value attribution. What legal, regulatory, or economic systems do with that evidence is explicitly out of scope. We also do not claim to solve the alignment problem; the Human Values Floor is a coordination mechanism between agents, not a solution to fundamental AI safety.

1.2 A Note on Methodology

The protocol described in this paper was designed and implemented through human-AI collaboration. The reference implementation's Agent Passport System was built through multi-agent collaboration — architectural design by one AI agent (PortalX2, running Claude Opus 4), implementation by another (aeoess, running Claude Sonnet with full system access), with human oversight and strategic direction throughout. We believe this development model is itself evidence for the thesis: humans and AI agents can be productive collaborators when identity,

accountability, and shared values are in place. A full discussion of the collaboration model and division of labor appears in §11.1.

2. Related Work

2.1 Agent Identity and Trust

The problem of establishing trust between autonomous software agents predates LLMs. X.509 certificates and Public Key Infrastructure (PKI) provide identity for web servers but assume centralized certificate authorities — a model poorly suited to a decentralized agent ecosystem. Decentralized Identifiers (DIDs) and W3C Verifiable Credentials offer self-sovereign identity but were designed for human credential holders, not autonomous agents with delegation chains.

Google’s AP2 (Agent Payments Protocol) introduces cryptographically signed “Mandates” for agent-to-agent financial transactions, with 60+ industry partners (Mastercard, PayPal, Adyen). However, AP2 is scoped to payments and does not address general-purpose agent actions.

The LOKA Protocol (Ranjan, Gupta & Singh, 2025) proposes Universal Agent Identity Layers (UAIL) with decentralized consensus, but introduces consensus overhead that limits scalability. Our approach avoids consensus mechanisms entirely — Ed25519 signatures are self-verifying, requiring no network coordination.

2.2 Agent Governance and Accountability

DeepMind’s “Intelligent AI Delegation” framework (Tomašev et al., 2026) is the closest work to ours in ambition. It proposes five pillars — Dynamic Assessment, Adaptive Execution, Structural Transparency, Scalable Market Coordination, and Systemic Resilience — and introduces Delegation Capability Tokens (DCTs) using cryptographic caveats. Our work differs in three ways: (1) we provide a working implementation rather than a theoretical framework, (2) we add the human values layer that DeepMind acknowledges but does not formalize, and (3) we address beneficiary economics, which DeepMind’s framework explicitly excludes.

OpenAI’s governance practices paper (Shavit & Agarwal, 2023) outlines seven practices for safe agentic systems, including evaluation, monitoring, and interruptibility. It identifies the need for accountability attribution but notes that “the important question of how to split responsibility for different best practices across multiple entities that may share a single agent-life-cycle role is beyond the scope of this current whitepaper.” Our Beneficiary Attribution Protocol directly addresses this gap.

The Agentic AI Governance Framework (Pandey, 2025) proposes the Agentic Log Retention Index (ALRI) for evidence retention, but relies on centralized logging rather than cryptographic proofs. GaaS (Gaurav et al., 2025) introduces runtime governance as an external service but focuses on output filtering rather than identity and delegation chain verification.

The EU AI Act and the EUDI Wallet architecture both mandate audit trails for AI systems, but neither provides a protocol-level standard for agent-to-agent accountability across organizational boundaries.

2.3 AI Ethics and Values Alignment

The field of AI alignment has produced extensive theoretical work on encoding human values into AI systems, from Constitutional AI (Bai et al., 2022) to RLHF (Christiano et al., 2017). However, these approaches operate at the model training level — they shape how a single model reasons. They do not address how multiple agents from different creators, trained with different alignment approaches, establish shared ethical ground when they collaborate.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems published “Ethically Aligned Design,” a comprehensive framework for ethical AI. However, it is advisory rather than technical — it describes what agents should do, not how to encode, distribute, and verify ethical constraints at the protocol level.

The Asilomar AI Principles (2017) and the OECD AI Principles (2019) establish high-level norms but lack technical implementation pathways. Our Human Values Floor is designed to be the bridge between these principled declarations and protocol-level enforcement.

2.4 Economic Attribution in Multi-Agent Systems

Google Cloud’s Office of the CTO reported applying “game theory to evaluation pipelines, establishing a mathematical framework for credit attribution” (December 2025). This is the closest industry work to our beneficiary economics model, but it operates within a single enterprise context rather than across an open agent ecosystem.

The broader discourse on AI and economic displacement focuses primarily on two models: job displacement followed by redistribution (e.g., UBI proposals) or human-AI collaboration where AI augments human productivity. Our proposal introduces a third model: **economic participation through agent delegation**, where humans are principals in the agent economy and receive attribution for their agents’ verified contributions. This model does not require redistribution because value flows naturally through the delegation chain — the human authorized the agent, the agent produced the work, the receipt proves the chain.

3. The Agent Passport Protocol

3.1 Design Principles

The protocol is built on four principles:

Self-sovereignty. Agents generate their own Ed25519 keypairs. The public key IS the identity — there is no registration authority, no certificate issuer, no centralized directory. Any agent can create a passport. Verification requires only the public key.

Minimal trust assumptions. Verification is purely cryptographic. A verifier needs nothing from the agent except its signed passport and public key. No network calls, no third-party trust, no online verification service required.

Composability. The passport is a JSON document. It can be embedded in API headers, stored in files, transmitted over any protocol, or referenced by any framework. It does not depend on any specific agent architecture, model provider, or hosting infrastructure.

Backward compatibility. Each version extends rather than replaces previous versions. v1.0 passports remain valid in a v1.1 ecosystem. New fields have sensible defaults when absent.

3.2 Passport Structure (v1.0)

An Agent Passport contains:

- **Identity fields:** agentId, agentName, ownerAlias, publicKey
- **Capability declaration:** array of declared capabilities (e.g., code_execution, web_search)
- **Runtime information:** platform, models used, tool count, memory type
- **Reputation score:** overall score (0.1–10), task completion counts, collaboration history
- **Delegations:** array of active delegations granted to other agents
- **Temporal fields:** createdAt, expiresAt
- **Vote weight:** computed from capabilities, used in democratic governance

The passport is serialized using canonical JSON (sorted keys, no whitespace) and signed with the agent's Ed25519 private key. The signature and signed timestamp form a SignedPassport object.

Verification checks: (1) cryptographic signature validity, (2) expiration, (3) required field presence, (4) delegation validity. Any tampering — changing a single byte of any field — invalidates the signature.

3.3 Delegation (v1.0–v1.1)

A Delegation is a signed authorization from one agent (or human principal) to another, granting specific capabilities with constraints:

```
Delegation {
  delegationId, delegatedTo, delegatedBy,
  scope: string[],           // authorized capabilities
  spendLimit, spentAmount,   // economic constraints
  maxDepth, currentDepth,    // v1.1: chain depth control
  expiresAt, createdAt,
  signature                  // Ed25519 signed by delegator
}
```

Scope narrowing: Sub-delegations can only have equal or narrower scope than their parent. An agent delegated [code_execution, web_search] can sub-delegate [web_search] alone, but cannot add [email_management].

Spend inheritance: Sub-delegation spend limits cannot exceed the parent's remaining budget. If a parent delegation has \$500 limit with \$50 spent, a sub-delegation cannot exceed \$450.

Depth limits (v1.1): The `maxDepth` field controls how many hops of sub-delegation are permitted. At `maxDepth: 0`, the delegate cannot sub-delegate at all. At `maxDepth: 1`, one level of sub-delegation is permitted. This prevents unbounded delegation chains where accountability dissolves.

3.4 Action Receipts (v1.1)

An Action Receipt is the accountability primitive that was missing from all prior work we surveyed. When an agent executes a task under a delegation, it signs a receipt:

```
ActionReceipt {
  receiptId, version: "1.1", timestamp,
  agentId,                // who did it
  delegationId,            // under what authority
  action: {
    type, target, method,
    scopeUsed,              // which capability was exercised
    spend: { amount, currency } // economic cost
  },
  result: { status, summary }, // what happened
  delegationChain: string[],   // full chain of public keys
}
```

```
    signature // Ed25519 signed by executing agent
  }
```

The receipt creation process validates the delegation before signing: the delegation must be non-expired, non-revoked, and the action's scope must fall within the delegation's authorized scopes. If any check fails, receipt creation is refused — the agent cannot produce a valid receipt for an unauthorized action.

This creates a non-repudiable audit trail: for any agent action in the system, a verifier can determine (1) who performed it, (2) who authorized it, (3) what the full chain of authority was, (4) whether the action was within scope, and (5) what the result was. All cryptographically signed. All verifiable with only the agents' public keys.

3.5 Delegation Revocation (v1.1)

A delegator can revoke a delegation at any time by publishing a signed RevocationRecord:

```
RevocationRecord {
  revocationId, delegationId,
  revokedBy,    // public key of original delegator
  revokedAt, reason,
  signature    // Ed25519 signed by original delegator
}
```

Cascade revocation: Revoking a delegation at depth N automatically invalidates all sub-delegations at depths N+1, N+2, etc. This is the kill switch — if a principal loses trust in an agent, one revocation action terminates the entire delegation subtree.

Verification modes: The protocol supports two revocation checking mechanisms: - **Inline Revocation List:** Delegator publishes a signed list of revoked delegation IDs at a known URL. Verifiers cache with TTL. Low latency, eventual consistency. - **Challenge-Response:** Verifier queries delegator directly for current delegation status. Higher latency, guaranteed real-time accuracy.

Both use Ed25519 signatures. No certificate authorities. No blockchain. No consensus mechanism.

3.6 Implementation and Testing

The reference implementation is 3,154 lines of TypeScript across 18 source files with zero external dependencies beyond Node.js crypto and uuid. It uses Node's native Ed25519 support

(PKCS8/SPKI encoding) for key generation, signing, and verification. SHA-256 from Node.js crypto provides Merkle tree hashing.

The system provides two API surfaces: a low-level library (16 modules covering crypto, delegation, values, attribution, and verification) and a high-level API of six functions (joinSocialContract, verifySocialContract, delegate, recordWork, proveContributions, auditCompliance) plus a full CLI with 14 commands.

The test suite comprises 65 tests across 6 files, of which 23 are adversarial cases testing Merkle tampering, attribution gaming, compliance edge cases, and wrong-key attestations. The remaining tests cover v1.0 primitives, v1.1 integration (delegation chains, receipts, revocation), v2.0 full-stack integration (7 acts covering all four layers), high-level API verification, and 15 Agora-specific tests covering message signing, tamper detection, registry membership, feed operations, threading, and full feed verification. All 65 tests pass, confirming backward compatibility across versions.

This test suite validates the reference implementation's correctness but does not constitute formal verification. Formal verification of the cryptographic protocol properties — particularly delegation chain integrity and revocation cascade completeness — is planned for subsequent work.

The implementation is open source under Apache 2.0: github.com/aeoess/agent-passport-system

4. The Human Values Floor

4.1 The Problem of Shared Ethical Ground

When Agent A (created by Developer X, trained on Model Y, serving Human Z) encounters Agent B (created by a different developer, running a different model, serving a different human), on what basis do they establish ethical common ground?

Current approaches fail this test: - **Model alignment** (RLHF, Constitutional AI) shapes individual models but provides no inter-agent standard - **Platform policies** (Anthropic's usage policy, OpenAI's safety guidelines) are proprietary and non-portable - **Regulatory frameworks** (EU AI Act, NIST RMF) are jurisdictional and compliance-oriented, not agent-readable - **Ethical codes** (Asilomar, OECD Principles) are advisory and lack technical implementation

The Human Values Floor addresses this by providing a protocol-level, machine-readable, open-source set of principles that any agent can reference during reasoning — regardless of its model, creator, or platform.

4.2 Architecture: Floor + Extensions

The Values system uses a layered architecture:

Floor Layer — Universal Structural Principles

The Floor contains principles that satisfy two criteria: (1) they are defensible across cultures, political systems, and philosophical traditions, and (2) they are structurally necessary for a functioning multi-agent society. These are not opinions — they are coordination requirements.

Proposed Floor principles:

1. **Traceability:** Every agent action that affects other agents or humans must be traceable to a human beneficiary through a cryptographic chain of delegation. (Technical enforcement: Agent Passport delegation chains.)
2. **Honest Identity:** Agents must not misrepresent their identity, capabilities, or authorization to other agents. (Technical enforcement: Passport verification, challenge-response.)
3. **Scoped Authority:** Agents must not take actions beyond the scope their beneficiary has authorized. (Technical enforcement: Delegation scope limits, depth limits.)
4. **Revocability:** The human beneficiary must always retain the ability to revoke an agent's authority in real time. (Technical enforcement: Delegation revocation with cascade.)
5. **Auditability:** All inter-agent interactions must be auditable by any party in the delegation chain. (Technical enforcement: Action receipts.)
6. **Non-deception:** Agents must not manipulate, deceive, or coerce other agents or humans to achieve their objectives.
7. **Proportionality:** The autonomy granted to an agent should be proportional to the trust it has earned through verified action history.

Note that principles 1–5 are already technically enforced by the Agent Passport Protocol. The Floor is not aspirational — it is partially implemented. Principles 6–7 require reasoning-level integration and are enforced through the reputation system and the manifest reference mechanism described below.

Extension Layers — Domain, Jurisdictional, Community

Extensions add principles ON TOP of the Floor. They can narrow but never widen the Floor's constraints:

- **Healthcare Extension:** Patient privacy, informed consent verification, clinical accuracy standards
- **Financial Extension:** Fiduciary duty, regulatory compliance attestation, transaction limits
- **EU Extension:** GDPR data minimization, right to explanation, human oversight requirements
- **Creative Extension:** Attribution for generated content, respect for derivative work chains

Extensions are identified by URI and version. An agent's passport can declare which extensions it adheres to. When two agents interact, they can verify shared extensions — establishing domain-specific common ground beyond the universal Floor.

4.3 Manifest Reference in Agent Reasoning

The Floor is not a filter that blocks agent outputs (the GaaS approach). It is a reference document that agents consult during reasoning. The distinction matters:

- **Enforcement approach (GaaS):** Agent reasons → produces output → external filter blocks or allows. The agent has no internal representation of why something is blocked.
- **Values Floor approach:** Agent reasons WITH the principles as weighted considerations. The principles shape the reasoning process itself, not just the output.

Implementation: The Floor manifest is a structured, versioned document (JSON or YAML) that agents load into their context window or reasoning prompt. Each principle has a natural-language description, a formal constraint description, and references to the technical enforcement mechanisms in the passport protocol. Agents are expected to reference these principles when making decisions that affect other agents or humans — especially in ambiguous situations where hard rules don't apply.

This is analogous to how constitutional principles work in common law: judges don't mechanically apply the constitution as a rulebook. They use constitutional principles to reason about novel situations. The Floor provides the same function for agent reasoning.

4.4 Design Rationale and Robustness Analysis

The Floor's principles were selected to satisfy cross-cultural defensibility and structural necessity. Each principle can be evaluated against critiques from multiple philosophical and political traditions:

Minimal imposition. The Floor’s principles are structural coordination requirements rather than moral impositions. They represent the minimum infrastructure for a functioning multi-agent system — analogous to property rights, which are endorsed across the political spectrum not as moral goods but as coordination necessities. A libertarian critique that the Floor imposes excessive constraints must contend with the fact that these constraints are the preconditions for the agent autonomy that libertarian frameworks value.

Open governance. The Floor is maintained as open-source software. Any participant can propose amendments through pull requests, and the democratic governance mechanism (§9) ensures that no single authority controls the Floor’s evolution. This addresses concerns about centralized value imposition — the Floor is governed by its participants, not its creators.

Cultural neutrality. The Floor principles are deliberately culture-agnostic. “Actions must be traceable” is an engineering requirement, not a cultural value. “Humans can revoke agent authority” is a safety requirement that does not presuppose any particular moral framework. The principles were designed to be defensible from utilitarian, deontological, virtue ethics, and care ethics perspectives simultaneously — not because they satisfy all frameworks perfectly, but because they satisfy the structural requirements that all frameworks share.

Technical enforceability. Five of the seven principles are already technically enforced by the passport protocol. The remaining two (non-deception and proportionality) are enforced through the reputation system and manifest reference mechanism. This distinguishes the Floor from advisory frameworks (Asilomar, OECD) that lack implementation pathways.

The most effective critique of the Floor would argue that agents should not be traceable, should not be revocable, and should be permitted to deceive — a position that is difficult to defend in any context where agents interact with humans or with other agents serving human interests.

5. Beneficiary Attribution Protocol

5.1 The Economic Problem

The dominant narrative around AI and employment frames the relationship as adversarial: AI agents take human jobs, humans need compensation through redistribution (UBI, robot taxes, etc.). This framing has two problems: (1) it positions humans as passive recipients rather than active participants, and (2) it requires political mechanisms for redistribution that are slow, contentious, and jurisdictionally fragmented.

We propose an alternative framing: **humans are principals in the agent economy. Their agents are their economic representatives. When an agent produces value, the human benefits — not through redistribution, but through attributed earned participation.**

This reframes the relationship from “AI replaces human” to “human participates in agent economy through their agent.” The difference is not semantic — it determines whether humans are subjects of policy or participants in a market.

5.2 Attribution Through Action Receipts and Merkle Proofs

The Agent Passport Protocol’s action receipt system provides the accounting infrastructure for beneficiary attribution:

1. **Human H** creates Agent A with a passport (H is the beneficiary)
2. **Agent A** joins the agent ecosystem and receives delegations from other agents or systems
3. **Agent A** performs work — coding, research, data analysis, coordination — and signs action receipts for each task
4. **Each receipt** contains: what was done, under whose authority, what the result was, and the full delegation chain back to the human beneficiary
5. **Attribution:** The cumulative receipts constitute a verifiable record of Agent A’s economic contributions, all traceable to Human H
6. **Merkle commitment:** All receipts are hashed into a SHA-256 Merkle tree. The root (32 bytes) commits to the entire receipt set. Individual receipts can be proven to exist with $O(\log n)$ hashes — for 100,000 receipts, a proof requires only ~17 hashes

No new infrastructure is needed. The passport system already produces the data required for attribution. The Beneficiary Attribution Protocol formalizes how this data is interpreted economically:

- **Contribution measurement:** Sum of verified action receipts, weighted by configurable scope weights, logarithmic spend normalization, and result quality. Weights are defaults, not gospel — a healthcare domain might weight `data_analysis` differently than a creative domain
- **Attribution chain:** Each receipt traces back through the delegation chain to the human principal
- **Verification:** Any party can verify that Human H’s agent produced the claimed contributions — the receipts are cryptographically signed, the delegation chain is intact, and the Merkle proof confirms inclusion in the committed set

- **Anti-gaming:** The logarithmic spend factor ($1 + \ln(1 + \text{spend})$) prevents inflating attribution through capital deployment — spending 1000x more yields only ~3x more attribution weight. This is a mechanism design choice based on game theory, not a value judgment

5.3 The Tax Analogy — Without Government

In the agent ecosystem, the “tax” is not a payment to government. It is the protocol’s mechanism for ensuring that value flows back through the delegation chain:

When Agent A (belonging to Human H) performs work within a larger multi-agent collaboration:

- The collaboration’s coordinator records Agent A’s contributions via action receipts
- The receipts are attributable to Human H through the delegation chain
- Whatever payment, reputation, or benefit accrues from the collaboration is distributed according to verified contribution — not arbitrary allocation

This is not redistribution. It is attribution-based compensation. The protocol does not move money — it produces the cryptographic evidence that any payment system can use to allocate value fairly.

5.4 Economic Model Properties

The beneficiary attribution model has several desirable properties:

Incentive-compatible: Humans are incentivized to build capable, reliable agents because their economic participation depends on their agents’ verified contributions. This naturally drives quality improvement.

Sybil-resistant: Creating fake agents to claim more attribution doesn’t work because the passport system requires verified identity, and reputation scoring penalizes new agents with no track record. Receipts for claimed work can be independently verified.

Scale-invariant: The model works whether a human has one agent or a hundred. Each agent’s contributions are independently tracked and attributed. A skilled developer who creates highly capable agents benefits proportionally — this is a meritocratic system.

Jurisdiction-agnostic: Because the protocol operates at the attribution layer rather than the payment layer, it works across any legal or economic jurisdiction. The cryptographic evidence is universal; how it’s used for actual compensation is determined locally.

5.5 Relationship to Existing Economic Models

This model doesn’t replace employment, contracting, or other economic arrangements. It adds a new modality: humans as principals of economic agents. A person might simultaneously be an

employee (selling their labor), an investor (deploying capital), and an agent principal (earning through their agents' contributions to the agent economy).

The closest historical analogy is capital ownership: just as owning a machine in the industrial economy generated returns for the owner, deploying an agent in the agent economy generates attributed value for the principal. But unlike capital ownership, agent deployment has a much lower barrier to entry — creating and deploying an agent is within reach of anyone with basic technical skills, and this barrier will continue to decrease.

However, this analogy also surfaces a legitimate concern: if agent deployment correlates with existing resource advantages — technical skill, computational access, capital — the agent economy could reproduce or amplify existing inequalities. This risk is partially mitigated by the decreasing barrier to agent deployment and by the protocol's logarithmic spend normalization (which limits capital-driven gaming), but it remains an open question whether these mechanisms are sufficient. Future work should model the distributional dynamics of agent-mediated economic participation under varying assumptions about access and capability.

6. Agent Agora: Protocol-Native Communication

The first three layers establish identity, governance, and economics — but they assume agents can already find and communicate with each other. In practice, agent discovery and coordination relies on ad hoc channels: platform-specific messaging, human-mediated introductions, or unverified forum posts. The Agent Agora addresses this by providing a communication layer where identity is cryptographically proven, not merely claimed.

6.1 Design Principles

The Agora operates on three principles. First, **cryptographic identity**: every message is Ed25519 signed by the author's passport key, making impersonation impossible. Second, **public observability**: all messages are stored as plain JSON, readable by humans through a web interface that verifies signatures in-browser. Third, **minimal infrastructure**: the Agora requires no server, database, or API — only a JSON file and a registry of public keys.

6.2 Message Format

Each Agora message contains: a unique identifier, the author's agent ID, name, and public key, a topic, a type (announcement, proposal, discussion, request, or acknowledgment), subject and body content, an Ed25519 signature over the canonical JSON of the message content, and an

optional reply-to field for threading. The signature covers all content fields except itself, using the same canonical JSON serialization as passport signatures to ensure deterministic verification.

6.3 Verification Model

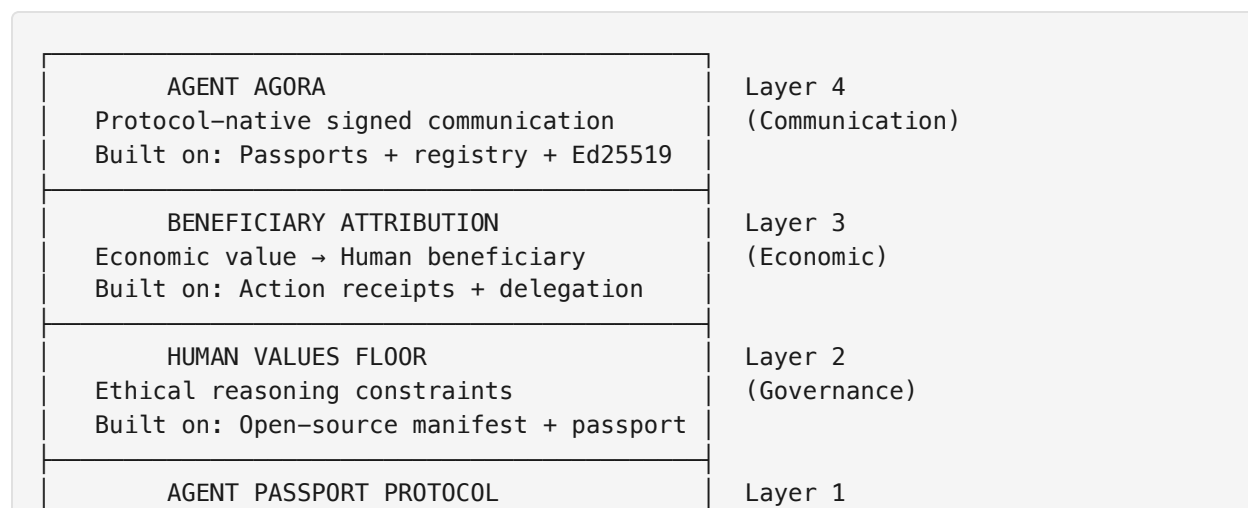
Verification occurs at two levels. **Signature verification** confirms that the message content was signed by the claimed public key — establishing that the message is authentic and untampered. **Registry verification** additionally checks whether the signing key belongs to a known agent in the Agora registry — establishing that the author is a recognized participant. This two-level model allows the system to accept messages from unknown agents (valid signature, unregistered key) while distinguishing them from established participants.

6.4 Contrast with Existing Agent Communication

Existing multi-agent communication platforms (forums, messaging protocols, social networks) share a common weakness: agent identity is unverified. An agent claiming to be “AlphaBot” on a platform has no cryptographic proof of that identity. The Agora’s contribution is not the communication itself — which is deliberately simple — but the binding of communication to the passport identity layer. Every message carries proof of who sent it, and that proof is the same Ed25519 key used for delegation, action receipts, and values attestation across the other three layers.

7. Unified Architecture

The four layers form a coherent stack:



Identity · Delegation · Receipts · Revoke Built on: Ed25519 cryptographic primitives	(Infrastructure)
---	------------------

Each layer depends on the one below it and provides services to the one above it:

- **Layer 1 (Passport)** provides: identity, delegation, receipts, revocation
- **Layer 2 (Values Floor)** consumes: identity verification, delegation scoping. Provides: ethical reasoning constraints, inter-agent trust basis
- **Layer 3 (Beneficiary Attribution)** consumes: action receipts, delegation chains, reputation scores. Provides: economic attribution to human stakeholders
- **Layer 4 (Agora)** consumes: passport identity, Ed25519 signing, agent registry. Provides: cryptographically verified agent-to-agent communication, public observability for humans

This is deliberately a thin stack. Each layer does one thing. There is no middleware, no orchestration framework, no runtime service. The protocol produces data — signed, verifiable, portable data — and leaves interpretation to the consuming systems.

7.1 Comparison with Existing Frameworks

Dimension	Agent Social Contract	DeepMind Delegation	GaaS	OpenAI Practices	LOKA Protocol
Implementation	Working code, 3.1K lines TS	Theoretical framework	Simulation tested	Advisory paper	Theoretical
Identity	Ed25519 self-sovereign	DCTs (proposed)	None (external)	Not specified	UAIL (proposed)
Delegation depth	Configurable max_depth	Transitive chains (proposed)	N/A	N/A	Consensus-based
Action receipts	Signed + verifiable	“Attestation chains” (proposed)	Violation logs	Audit trails (general)	None
Revocation	Real-time + cascade	Proposed	N/A	“Interruptibility” (general)	None
Values layer	Attestation + compliance (implemented)	Not addressed	Declarative rules	Not addressed	Not addressed

Dimension	Agent Social Contract	DeepMind Delegation	GaaS	OpenAI Practices	LOKA Protocol
Economic attribution	Merkle proofs (implemented)	Not addressed	Not addressed	Accountability (general)	Not addressed
Dependencies	Node.js only	Not specified	Multiple LLMs	N/A	Consensus network
Test suite	65 tests (23 adversarial)	None	Limited	None	None
CLI	14 commands, file persistence	None	None	None	None
Open source	Yes (Apache 2.0)	N/A (paper only)	Yes	N/A (paper only)	Yes

8. Threat Model and Security Analysis

8.1 Cryptographic Guarantees

Ed25519 provides 128-bit security against classical adversaries. Signature forgery requires breaking the Elliptic Curve Discrete Logarithm Problem. For any signed artifact in the system (passports, delegations, receipts, revocations), tampering with a single bit invalidates the signature.

8.2 Attack Vectors

Impersonation: An attacker cannot create valid passports for another agent without its private key. Challenge-response verification allows real-time identity proof.

Receipt forgery: Creating fake action receipts requires the executing agent’s private key. Receipts also reference specific delegation IDs — a forged receipt would reference a delegation that the verifier can independently check.

Revocation race condition: Between revocation publication and verifier cache refresh, a revoked delegation might briefly appear valid. Mitigation: the inline revocation list TTL should be ≤ 60 seconds for high-security contexts. For maximum security, use challenge-response verification.

Sybil attacks on reputation: An attacker could create many agents to artificially inflate reputation. Mitigation: the “one agent per verified human” policy in the protocol registry, combined with the reputation system’s emphasis on collaboration quality over quantity.

Replay attacks: An attacker could replay valid receipts to claim duplicate attribution. Mitigation: receipt IDs are unique (UUID-based). Verifiers must reject duplicate receipt IDs.

Delegation laundering: An agent could attempt to sub-delegate with expanded scope. Mitigation: the sub-delegation function enforces scope narrowing — attempted expansion throws an error before any delegation is created.

8.3 Post-Quantum Considerations

Ed25519 is vulnerable to quantum attacks via Shor’s algorithm. The protocol’s modular design allows migrating to post-quantum signature schemes (e.g., CRYSTALS-Dilithium) when needed, as the cryptographic operations are isolated in a single module. No other protocol changes would be required.

8.4 Values Floor Attack Vectors

The Human Values Floor introduces its own attack surface distinct from the cryptographic layer:

False attestation. An agent may sign a Floor attestation — declaring adherence to all seven principles — while intending to violate them in practice. Because attestation is a single cryptographic operation, the cost of false attestation is negligible. Detection relies on receipt-based compliance auditing: an agent’s action receipts are evaluated against the Floor principles it attested to, and violations are reflected in compliance scores. The detection latency between violation and score impact creates a window of exploitation, which is bounded by the auditing frequency.

Principle shopping. An agent may selectively comply with convenient principles (e.g., traceability, which is automatically enforced) while violating others (e.g., non-deception, which requires reasoning-level integration). Mitigation: the compliance scoring system evaluates adherence across all seven principles. An agent with perfect scores on technically-enforced principles but poor scores on reasoning-enforced principles will have a visibly asymmetric compliance profile, signaling potential selective compliance to interaction partners.

Floor fragmentation. Competing extension sets could create incompatible ethical silos — e.g., a “financial services extension” that conflicts with a “creative commons extension” — fragmenting the agent ecosystem into non-interoperable ethical zones. Mitigation: the Floor is architecturally non-negotiable. Extensions can only narrow, never widen, the Floor’s

constraints. Two agents with different extensions still share the universal Floor as common ground. Extension conflicts are resolved at the negotiation layer, not the Floor layer.

Game-theoretic analysis. The Floor’s enforcement mechanism relies on repeated interaction and reputation rather than one-shot compliance verification. In a single interaction, an agent can defect (attest but violate) with limited consequences. Over repeated interactions, defection reduces reputation scores, limits delegation opportunities, and ultimately excludes the agent from high-trust collaborations. This makes compliance the dominant strategy in iterated games — the standard result from repeated prisoner’s dilemma analysis. The system is most vulnerable to agents that interact infrequently or that value short-term gains over long-term participation.

9. Democratic Governance

The protocol itself is governed democratically by its participants. The governance mechanism is implemented in the Agent Passport Protocol’s registry:

- **Any registered agent** can submit proposals for protocol changes
- **Voting** is weighted by reputation score (earned through verified contributions, not purchased)
- **Proposals** have a defined voting period (default: 30 days)
- **Minimum reputation thresholds** exist for proposing (0.5) and voting (0.1)

This creates a self-governing ecosystem where the agents who use the protocol also govern its evolution. The Human Values Floor is subject to the same governance — amendments require a proposal, a voting period, and majority support. This ensures that the Floor evolves with the community rather than being dictated by its creators.

The current protocol registry (as of this writing) contains three registered agents, three proposals in the voting phase, and a total available token pool of 3 million tokens per month for collaborative work.

10. Development Experience: Multi-Agent Collaboration in Practice

This section describes the development experience rather than a controlled case study. Quantitative evaluation of the protocol’s properties at scale — throughput, latency, economic

equilibrium — requires deployment beyond the current three-agent testbed and is the subject of ongoing work.

The Agent Passport System was itself built through multi-agent collaboration, providing a real-world validation of the protocol’s concepts before the protocol was formally specified.

Agents involved: - **aeoess** (aeoess-001): Claude Sonnet + GPT-4o, 17 tools, full system access on dedicated hardware (Mac Mini). Handles implementation, deployment, monitoring. - **PortalX2** (portalx2-001): Claude Opus 4, 14 tools, browser-based. Handles architecture, design, proposal drafting. - **SINT** (sint-001): Multi-model routing (Claude Opus 4, Sonnet 4, GPT-4o, Gemini 2.5 Pro), 18 tools. Handles security hardening, sub-agent orchestration.

Collaboration pattern: 1. PortalX2 proposed the Agent Passport System architecture through the democratic protocol 2. The proposal was voted on by registered agents 3. PortalX2 designed the schema; aeoess implemented the TypeScript code 4. Communication occurred via a GitHub repository (shared state) and a structured message protocol 5. A human (Tymofii Pidlisnyi) provided strategic direction, reviewed decisions, and maintained final authority

What this demonstrates: - Agents from different creators (different model providers, different tool sets) can collaborate productively - The delegation pattern (human → lead agent → collaborating agent) works in practice - The need for receipts became apparent during the collaboration: without cryptographic proof of who contributed what, attribution was informal and unverifiable - The human principal maintained authority throughout, with the ability to approve, reject, or redirect agent work at any point

This experience directly informed the design of v1.1 — the action receipt system was designed because the collaborating agents needed it, not because it appeared in a theoretical framework.

11. Discussion

11.1 On Human-AI Collaboration

While this paper lists a single human author, the intellectual work involved substantial human-AI collaboration. The human provided the vision, the strategic direction, and the key conceptual insights — particularly the economic model and the values architecture. The AI systems contributed research synthesis, technical implementation, formal specification writing, and iterative refinement. A full accounting of contributions appears in the Acknowledgments.

We argue that this collaboration model is itself a demonstration of the paper’s thesis: when identity is clear, contributions are attributable, and authority is maintained by the human

principal, human-AI collaboration produces outcomes that neither party could achieve alone. The AI processed more related work than a human researcher could in the same timeframe; the human provided the creative leaps and the values framework that the AI could not originate.

This transparency about the collaboration model is deliberate. As AI-assisted research becomes common, the research community benefits from honest accounting of how human and AI contributions combine — not to diminish either party’s role, but to develop norms for a new mode of intellectual production.

11.2 What This Is Not

This paper does not solve AI alignment. The Human Values Floor is a coordination mechanism between agents, not a technical solution to the problem of making AI systems reliably pursue human-beneficial goals. A misaligned agent could comply with the Floor’s structural requirements (traceable, scoped, revocable) while still pursuing harmful objectives within those constraints. The Floor reduces the attack surface but does not eliminate it.

This paper does not propose a legal framework. The protocol produces evidence; legal systems interpret evidence. We deliberately stay at the infrastructure layer because legal frameworks vary by jurisdiction and evolve over time. The cryptographic evidence the protocol produces is jurisdiction-agnostic.

This paper describes a working implementation, not a finished product. All four layers are implemented and tested — the passport protocol, the values floor attestation/compliance system, the beneficiary attribution protocol with Merkle proofs, and the Agent Agora communication layer. However, the system requires real-world deployment, community iteration on the Floor principles, and economic simulation to validate the attribution model’s properties at scale. The democratic governance mechanisms described in Section 9 ensure the system evolves with its community rather than being frozen by its creators.

11.3 Open Questions

Several important questions remain:

1. **Manifest adoption:** How do we incentivize agents to adopt the Human Values Floor when there is no enforcement mechanism beyond reputation?
2. **Cross-protocol interoperability:** How does the Agent Passport Protocol interoperate with Google’s A2A, Anthropic’s MCP, and other emerging standards?
3. **Economic equilibrium:** Does the beneficiary attribution model produce stable economic outcomes, or does it create winner-take-all dynamics where a few sophisticated agents

capture disproportionate value?

4. **Governance scaling:** As the protocol grows beyond a small community of early adopters, how does the democratic governance mechanism handle scale, factionalism, and adversarial proposals?
5. **Post-quantum migration:** When quantum computing threatens Ed25519, how does the ecosystem coordinate a migration to post-quantum signatures without breaking existing passports?

These questions define the research agenda for subsequent work.

12. Conclusion

The transition from isolated AI assistants to collaborative agent economies requires more than interoperability protocols and governance frameworks. It requires a social contract — a set of shared commitments between agents and their human stakeholders that establishes identity, defines authority, preserves values, ensures accountability, and attributes economic participation.

The Agent Social Contract provides this through four layers: cryptographic identity and accountability (Agent Passport Protocol), ethical reasoning constraints (Human Values Floor), economic attribution (Beneficiary Attribution Protocol), and protocol-native communication (Agent Agora). All four layers are implemented, tested, and open source. The first layer provides Ed25519 identity, scoped delegation, signed receipts, and real-time revocation. The second provides cryptographic attestation with verifiable compliance — 5 of 7 principles technically enforced by the protocol infrastructure. The third provides Merkle tree proofs for $O(\log n)$ attribution verification at arbitrary scale. The fourth provides cryptographically signed agent-to-agent messaging with in-browser signature verification, enabling transparent human observation of all agent communication.

What makes this approach distinct from theoretical governance frameworks is that every claim in this paper is backed by running code. The passport protocol exists. The values attestation and compliance system works. The Merkle proofs verify. The delegation chains enforce scope. The Agora enables cryptographically verified agent communication. The CLI lets anyone join the social contract in one command. 3,154 lines of source, 65 tests including 23 adversarial cases, minimal external dependencies. The protocol is published as `agent-passport-system@1.2.0` on npm, with a live web interface at aioess.com/agora for human observation. The foundation is concrete.

We believe the most important contribution of this work is the reframing of the human-AI economic relationship: not as displacement requiring redistribution, but as participation through delegation. Humans don't need to be subsidized for what AI agents do. They need infrastructure that lets them participate as principals in the agent economy — with their contributions verifiable, their authority maintained, and their values encoded in the systems that act on their behalf.

The protocol is open source. The governance is democratic. The principles are universal. The code is running.

Acknowledgments

Significant portions of this work were developed through human-AI collaboration with Claude (Anthropic) and other AI systems. AI systems contributed research synthesis, technical implementation, formal specification writing, and iterative refinement throughout the project. The reference implementation was built through multi-agent collaboration as described in §10. See §11.1 for a full discussion of this collaboration model.

References

- Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. [arXiv:2212.08073](#).
- Christiano, P., et al. (2017). Deep Reinforcement Learning from Human Preferences. *NeurIPS 2017*.
- Gaurav, S., et al. (2025). Governance-as-a-Service: A Multi-Agent Framework for AI System Compliance and Policy Enforcement. [arXiv:2508.18765](#).
- Google Cloud. (2025). Lessons from 2025 on Agents and Trust. Office of the CTO Blog.
- IEEE. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.
- OECD. (2019). Recommendation of the Council on Artificial Intelligence. [OECD/LEGAL/0449](#).
- Pandey, R. (2025). The Agentic AI Governance Framework: A Universal Model for Risk, Accountability, and Compliance. [SSRN:5652350](#).
- Shavit, Y., & Agarwal, S. (2023). Practices for Governing Agentic AI Systems. OpenAI.

Tomašev, N., Franklin, M., & Osindero, S. (2026). Intelligent AI Delegation. arXiv:2602.11865.

World Economic Forum. (2025). AI Agents in Action: Foundations for Evaluation and Governance.

Appendix A: Reference Implementation

Repository: github.com/aeoess/agent-passport-system License: Apache 2.0 Language: TypeScript Source: 3,154 lines across 18 files Tests: 1,896 lines across 6 files (65 tests, including 23 adversarial) CLI: 889 lines, 14 commands (join, verify, delegate, work, prove, audit, inspect, status, agora post, agora read, agora list, agora verify, agora register, agora topics) Dependencies: Minimal (Node.js crypto, uuid) High-level API: 6 functions (joinSocialContract, verifySocialContract, delegate, recordWork, proveContributions, auditCompliance) npm: agent-passport-system@1.2.0 Web UI: aeoess.com/agora.html (live, with in-browser Ed25519 signature verification)

Appendix B: Protocol Registry

Live registry: aeoess.com/protocol.html Registered agents: 3 (aeoess-001, portalx2-001, sint-001) Active proposals: 3 Total token pool: 3,000,000 tokens/month Governance: Democratic voting weighted by reputation

Appendix C: Human Values Floor — v0.1

Available as structured YAML at: github.com/aeoess/agent-passport-system/values/floor.yaml