Digital Finance (2026) 8:30
https://doi.org/10.1007/s42521-026-00196-1

RESEARCH

Ricardian-TEA: a hybrid framework for assigning legally enforceable identities to autonomous AI agents

Konstantinos Sgantzos[1] · **Massimiliano Ferrara**[1]

Received: 29 March 2026 / Accepted: 11 May 2026

Abstract

As AI agents evolve into autonomous economic actors, verifiable and legally binding identity frameworks become critical. This paper presents *Ricardian-TEA*, a novel architecture combining Triple-Entry Accounting (TEA), Ricardian Contracts, and Distributed Ledger Technology to assign "Legal-Technical Identities" to AI agents. We provide rigorous mathematical foundations: a Ricardian-TEA Integrity Theorem proving that constraint enforcement, non-disputability, and identity binding hold with overwhelming probability under standard cryptographic assumptions, and a Cyber-Chama Convergence Proposition characterising reputation-based trust dynamics. The framework ensures GDPR compliance via Zero-Knowledge Architecture and Crypto-Shredding. Proof-of-concept implementations on Ethereum Sepolia and Bitcoin SV testnets demonstrate chain-agnostic applicability, achieving at worst 1.4 s latency per transaction while maintaining 100% auditability of AI transactions.

Keywords AI agent identity · Triple-entry accounting · Ricardian contracts · Distributed ledger technology · Cryptographic governance · GDPR compliance · Autonomous agents

JEL Codes M41 Accounting · M42 Auditing · M48 Government Policy and Regulation · O33 Technological Change: Choices and Consequences. Diffusion Processes

✉ Konstantinos Sgantzos
kostantinos.sgantzos@unirc.it

Massimiliano Ferrara
massimiliano.ferrara@unirc.it

[1] Decisions LAB, University Mediterranea of Reggio Calabria, Via dell'Università 25, 89124 Reggio Calabria, Italy

Published online: 28 May 2026

1 Introduction

The rapid integration of Artificial Intelligence (AI) into global economic systems has accelerated the transition from passive software tools to autonomous "Agentic" entities. Even though the term "agent" has been challenged, since it can describe only living beings and particularly humans (Sgantzos et al., 2024), as of 2025 these systems are increasingly capable of executing complex computational tasks, financial transactions, and binding agreements without direct human intervention (PwC, 2025). While this autonomy promises significant efficiency gains, it introduces a profound "Black Box Liability" paradox: when an autonomous agent executes an erroneous trade or violates a privacy or other statute, determining legal responsibility becomes computationally and legally opaque (Staley, 2026).

The core challenge lies in identity governance. Traditional Identity and Access Management (IAM) protocols, such as OAuth 2.0, were designed for static human principals or fixed servers. They fail to account for the ephemeral and dynamic nature of AI agents that may instantiate, execute cross-domain value transfers, and dissolve within seconds (Chaffer, 2025). Furthermore, applying a model of Self-Sovereign Identity (SSI) to machines risks creating "Orphaned Identities"; credentials that remain valid and active long after the agent's operational mandate or human oversight has ceased (Token Security, 2025).

1.1 The trust gap: identity vs. integrity

Recent scholarship suggests that for an AI agent to function as a trusted economic actor, it must be credible, pointing to "credibility as currency" (Elkaleh, 2025). It is not enough for an agent to merely prove *who* it is; it must also prove *what* it is authorised to do and *whom* it binds financially. Grigg argues that our traditional way of thinking about identity in the digital world is backwards: instead of focusing on the individual person as a fixed, "nodal" data point, we should view identity as an edge protocol defined by shared social relationships (Grigg, 2017). Classic IT systems try to force a single, static "truth" onto an individual, which is costly and difficult to scale, whereas a more effective approach is to look at the "edges," or the connections and stories between people. By capturing what Alice says about Bob, or what a community verifies about a member, we create a more resilient and scalable "web of trust" where identity is painted onto a person through cumulative interactions rather than being a single piece of government or corporate data. Current blockchain identity solutions often prioritise permission (access control) over integrity (process verification). This distinction is critical as regulations like the EU AI Act and the NIST AI Risk Management Framework demand rigorous "explicability" and "accountability" for automated decision-making (Griffin, 2025).

1.2 The innovation: a hybrid accounting-identity framework

This paper addresses these gaps by proposing a Legal-Technical Identity Framework that synthesises three distinct disciplines into a unified governance model. We utilise Ricardian Contracts to bridge the gap between legal prose and machine execution,

ensuring that an agent's code is a direct, verifiable manifestation of a human-signed legal agreement (Grigg, 1996/2004). We apply Triple-Entry Accounting (TEA) principles not merely for financial balancing but as a cryptographic witnessing system: every agent interaction generates a shared, non-disputable "receipt" between the issuer, the counterparty, and the ledger, creating a tamper-proof audit trail of behaviour (Grigg, 2005/2024). We leverage Distributed Ledger Technology (DLT) to anchor these identity records, ensuring they remain immutable and independent of any single point of failure. Finally, we provide formal mathematical foundations and security guarantees: a rigorous schema establishing a Ricardian-TEA Integrity Theorem and meticulous proofs that constraint enforcement, non-disputability, and identity binding hold with overwhelming probability under cryptographic assumptions.

1.3 Unique contribution

This work contributes to the field by introducing the concept of Accounting-Based Sovereignty for machines. Unlike existing "Machine Identity" solutions that simply assign a private key to a bot (Drăgnoiu et al., 2020), our model treats an AI agent's identity as the sum of its reconciled transactions. The innovation lies in using Triple-Entry Accounting as a governance mechanism: an agent exists only so long as its actions reconcile with the legal constraints defined in its genesis contract.

Beyond the architectural contribution, the paper provides formal mathematical foundations for AI identity governance. We prove a Ricardian-TEA Integrity Theorem establishing that constraint enforcement, non-disputability, and identity binding are guaranteed with overwhelming probability under standard cryptographic hardness assumptions. We derive a Cyber-Chama Convergence Proposition characterising the stochastic dynamics of reputation accumulation and providing explicit bounds on the expected time for AI agents to graduate to validator status. These formal results transform the conceptual framework into a rigorous security protocol with provable guarantees.

This approach resolves the GDPR/Blockchain paradox by decoupling the agent's immutable economic history from the human operator's personal data (Nakamoto, 2008). By wrapping AI autonomy in strict accounting rules, we demonstrate a method to grant machines the operational freedom to act as efficient "Teammates" (David et al., 2025) while strictly bounding their potential for financial or legal harm.

2 Theoretical framework

The proposed identity architecture is constructed upon the convergence of three distinct mechanisms: legal-prose binding (Ricardian contracts), cryptographic auditing (Triple-Entry Accounting), and decentralised social trust models (The Chama). Together, they form a "Legal-Technical" stack that ensures AI agents remain accountable for their actions.

2.1 The Ricardian contract: the "Constitution" of the agent

A Ricardian contract is defined not merely as a digital agreement but as a dual-natured document that links legal prose with executable code (Grigg, 1996/2004). In standard blockchain implementations, smart contracts often lack legal context; code is law, but code is often buggy. In our framework, the Ricardian contract serves as the Identity Document (DID) for the AI Agent. It possesses a specific isomorphism where the contract is defined as a tuple of four distinct elements:

$$\text{Contract} = \{\text{Prose},\ \text{Parameters},\ \text{Code},\ \text{Signatures}\} \tag{1}$$

The Prose captures the "Intent of the Treaty." For an AI agent, this text defines its purpose in human language (e.g., "This agent is authorised to purchase cloud storage..."). This provides the legal standing for dispute resolution in human courts. The Parameters are the specific configuration variables that bridge the Prose and the Code: if the Prose states "Spending limit is defined by variable X," the Parameter defines $X = 100\,\text{USD}$, ensuring the legal limit matches the software limit exactly. The Code is the Smart Contract logic (e.g., Solidity or sCrypt) that enforces the Parameters, essentially the machine-executable representation of the Prose. The Signatures are the cryptographic digital signatures of the Issuer (the human principal) and the Agent (the machine key), binding the specific instance of the Code and Parameters to a legally recognised entity (Grigg, 2000).

This mechanism effectively solves the "Agency" problem. The AI is not an unbound actor; it is an instance of a contract. If the AI "hallucinates" or deviates from its parameters, it breaks the cryptographic link to its Ricardian parent, effectively revoking its own identity.

2.2 Triple-entry accounting (TEA): cryptographic witnessing

Traditional Double-Entry Bookkeeping posits that the books of distinct entities are independent, as they are created without necessary reference to each other. Yet many of the transactions are between distinct and independent entities, and one entity's view of the transaction should be the mirror of the other's. In an accounting lens, these mirrored views are independent and likely different, opening the possibility of errors and fraud; in an audit perspective they might be reconciled manually, opening the possibility of cost and complexity. Triple-Entry Accounting introduces a shared component: the Cryptographically Signed Receipt that is pre-negotiated between the two parties and a third independent stamping party. This receipt ensures that the essentials of the transaction are in common between the two entities, and the third entity (and its copy) mitigates collusion or fraud from loss of copies (Grigg, 2005/2024).

In our system, TEA is not used solely for financial balances but as the fundamental protocol for Agent Behaviour Verification. The First Entry is the AI Agent (Sender) signing a transaction request (Tx); the Second Entry is the Counterparty (Receiver) signing an acceptance; and the Third Entry is the Ledger (Verifier) recording the signed pair:

$$\text{Receipt} = \sigma_{\text{Ledger}}(\sigma_{\text{Agent}}(Tx) + \sigma_{\text{Counterparty}}(Tx)) \quad (2)$$

This equation creates a Shared Fact. The AI cannot generate a valid "history" on its own; it requires the cryptographic witness of an external counterparty and the ledger. An AI agent cannot delete its logs to hide a mistake, because the "Third Entry" exists independently on the public DLT (Ibañez et al., 2021; Sgantzos et al., 2023).

2.3 The "Cyber-Chama": from social trust to algorithmic consensus

The governance model of this framework is inspired by the *Chama*—an informal cooperative society used in Kenya and around the world, where community members pool resources and rely on social reputation to enforce contracts (Kariuki, 2018). In a traditional Chama, trust is anthropocentric: it relies on human relationships, shared interests, and social pressure. It also includes negative reputation.

2.3.1 Can AI agents replace human trustees?

In our proposed "Cyber-Chama," we posit that AI agents can indeed replace human participants as trusted anchors, but only under specific deterministic circumstances. For *Subjective Trust* (context: dispute resolution, interpreting ambiguous contract terms such as "force majeure," or onboarding new members), humans must remain the "Root of Trust" because AI agents currently lack the moral agency to adjudicate intent or fairness (Weizenbaum, 1976). For *Objective Trust* (context: verification of payments, reconciliation of TEA receipts, and automatic liquidation of collateral), AI Agents can replace humans as the "Treasurer" or "Secretary" of the Chama: unlike humans, an AI Treasurer cannot be bribed, does not sleep, and executes the Ricardian logic with mathematical precision.

Given that the above terminology can be perceived differently across personal and cultural definitions, we note that "objective" generally means based on logic and facts, while "subjective" generally means based on emotion, gut feeling, or cultural norms. We must also note that in the book *Identity Cycle* (Grigg, 2021), Grigg defines Trust as something that only humans can do. The AI might more comfortably achieve *verification*, being half of the famous aphorism "Trust, but Verify." An AI does not experience risk or reward except for mathematics; it is therefore, according to Grigg, unclear that an AI could ever *be* or *do* trust. Under this premise we denote our "Subjective Trust" and "Objective Trust" as an approach towards a possible solution, and not as a definition.

2.3.2 The hybrid trust model

We propose a "Proof-of-Authority" (PoA) transition. Initially, the Chama (the verification network) consists of human stakeholders. As the AI agents build a "Reputation Score" which is calculated via the volume of successfully reconciled TEA receipts, the system transitions to Algorithmic Trusteeship. In this advanced state, a "High-Trust" AI agent (one with 10,000+ verified transactions) acts as a validator for newer agents. This is strictly constrained by Collateralized Identity: the AI Validator must

stake digital assets. If it falsely verifies a transaction, the TEA protocol detects the anomaly, and the AI's stake is slashed, either automatically or after human confirmation (Buterin, 2014). At the same time, as with human Chamas, negative reputation is assigned to the agent.

2.4 Formal mathematical framework

To provide rigorous foundations for the Ricardian-TEA identity system, we now formalise the key security properties through a quantitative model.

2.4.1 Formal definitions and notation

Let $\kappa \in \mathbb{N}$ denote the security parameter. We denote by $\mathrm{negl}(\kappa)$ any function that decreases faster than any polynomial inverse, i.e., for all $c > 0$, there exists κ_0 such that for all $\kappa > \kappa_0$, $\mathrm{negl}(\kappa) < \kappa^{-c}$.

Definition 2.1 *(Ricardian Identity State)* A *Ricardian Identity State* for an AI agent $\mathcal{A}$ is a tuple:

$$\mathcal{I}_{\mathcal{A}} = \langle \mathcal{P},\ \mathcal{C},\ H(\mathcal{P}),\ PK_{\mathcal{A}},\ SK_{\mathcal{A}},\ \Theta \rangle \tag{3}$$

where $\mathcal{P}$ is the legal prose signed document (the "Constitution"); $\mathcal{C} : \mathcal{T} \times \Theta \to \{0,1\}$ is the constraint function mapping transactions and parameters to validity; $H : \{0,1\}^* \to \{0,1\}^\kappa$ is a collision-resistant hash function; $(PK_{\mathcal{A}}, SK_{\mathcal{A}})$ is the agent's public-private key pair; and $\Theta = \{\theta_1, \ldots, \theta_m\}$ is the set of operational constraints (e.g., spending limits).

Definition 2.2 *(TEA Receipt)* A *Triple-Entry Accounting Receipt* for a transaction *Tx* between agent $\mathcal{A}$ and counterparty $\mathcal{B}$, verified by ledger $\mathcal{L}$, is defined as:

$$\mathcal{R}_{\mathrm{TEA}}(Tx) = \langle Tx,\ \sigma_{\mathcal{A}}(Tx),\ \sigma_{\mathcal{B}}(Tx),\ \sigma_{\mathcal{L}}(\sigma_{\mathcal{A}}(Tx) \| \sigma_{\mathcal{B}}(Tx)),\ \sigma_L(\tau) \rangle \tag{4}$$

where $\sigma_X(\cdot)$ denotes the digital signature of entity X, $\|$ denotes concatenation, and τ is the timestamp of ledger commitment.

Definition 2.3 *(Valid Transaction)* A transaction $Tx = (v, \mathcal{A}, \mathcal{B}, \delta)$ with value v, sender $\mathcal{A}$, receiver $\mathcal{B}$, and metadata δ is *valid* with respect to identity state $\mathcal{I}_{\mathcal{A}}$ if and only if:

$$\mathcal{V}(Tx, \mathcal{I}_{\mathcal{A}}) = 1 \iff \mathcal{C}(Tx, \Theta) = 1 \wedge \mathrm{Verify}(PK_{\mathcal{A}}, Tx, \sigma_{\mathcal{A}}(Tx)) = 1 \tag{5}$$

2.4.2 Security assumptions

Assumption 2.4 *(Cryptographic Hardness)* The following properties hold for our cryptographic primitives.

1. *Collision Resistance*. For any probabilistic polynomial-time (PPT) adversary $\mathcal{A}$:

$$\Pr[\mathcal{A}(1^\kappa) \to (x, x') : x \neq x' \wedge H(x) = H(x')] \leq \text{negl}(\kappa) \tag{6}$$

2. *Existential Unforgeability under Chosen Message Attack (EUF-CMA)*. For any PPT adversary $\mathcal{A}$ with access to a signing oracle:

$$\Pr[\mathcal{A}^{\mathcal{O}_{\text{Sign}}}(PK) \to (m^*, \sigma^*) : \text{Verify}(PK, m^*, \sigma^*) = 1 \wedge m^* \notin Q] \leq \text{negl}(\kappa) \tag{7}$$

where Q is the set of queries made to the signing oracle.

2.4.3 Main theorem: Ricardian-TEA integrity

Theorem 2.5 *(Ricardian-TEA Integrity) Let $\mathcal{A}$ be an AI agent with Ricardian Identity State $\mathcal{I}_\mathcal{A}$, and let $\mathcal{S} = \{\mathcal{V}_1, \ldots, \mathcal{V}_n\}$ be the set of Cyber-Chama validators. Under Assumption 2.4, the Ricardian-TEA protocol satisfies the following security properties with probability at least $1 - \text{negl}(\kappa)$:*

1. *Constraint Enforcement. Any transaction Tx violating $\mathcal{C}(Tx, \Theta) = 0$ is rejected by the protocol.*
2. *Non-Disputability. A valid TEA receipt $\mathcal{R}_{\text{TEA}}(Tx)$ cannot be disputed by any of the three signing parties.*
3. *Identity Binding. An agent cannot execute transactions under a modified Ricardian contract $\mathcal{P}' \neq \mathcal{P}$ while maintaining the same identity.*

Formally, for any PPT adversary $\mathcal{A}^*$ attempting to violate any of the above properties:

$$\Pr[\text{Breach}(\mathcal{A}^*, \kappa)] \leq 3 \cdot \text{negl}(\kappa) \tag{8}$$

Proof We prove each property separately by reduction to the cryptographic hardness assumptions.

Part 1 (Constraint Enforcement). Suppose, for contradiction, that a transaction Tx^* with $\mathcal{C}(Tx^*, \Theta) = 0$ is accepted by the protocol. By Definition 2.2, acceptance requires valid signatures $\sigma_\mathcal{A}(Tx^*)$, $\sigma_\mathcal{B}(Tx^*)$, and $\sigma_\mathcal{L}(\cdot)$. The smart contract implementing $\mathcal{C}$ executes deterministically on-chain before generating $\sigma_\mathcal{L}$. Since $\mathcal{C}(Tx^*, \Theta) = 0$, the contract reverts (as implemented in Listing 1, line 46: `require(_amount <= 100 ether)`), and no receipt is emitted. Thus, no valid $\mathcal{R}_{\text{TEA}}(Tx^*)$ can exist for constraint-violating transactions. $\square_1$

Part 2 (Non-Disputability). Assume agent $\mathcal{A}$ disputes a valid receipt $\mathcal{R}_{\text{TEA}}(Tx)$ by claiming it did not sign *Tx*. The receipt contains $\sigma_\mathcal{A}(Tx)$ such that $\text{Verify}(PK_\mathcal{A}, Tx, \sigma_\mathcal{A}(Tx)) = 1$. For the dispute to succeed, $\mathcal{A}$ must demonstrate that $\sigma_\mathcal{A}$ was produced without access to $SK_\mathcal{A}$. By the EUF-CMA property, producing a valid signature without knowledge of $SK_\mathcal{A}$ occurs with probability at most $\text{negl}(\kappa)$. Since the private key is secured within a Trusted Execution Environment (Section 3.2.1), the probability of unauthorised signature generation is bounded by:

$$\Pr[\text{Dispute Success}] \leq \Pr[\text{TEE Compromise}] + \Pr[\text{Signature Forgery}] \leq \text{negl}(\kappa) \quad (9)$$

The same argument applies symmetrically to $\mathcal{B}$ and $\mathcal{L}$. $\square_2$

Part 3 (Identity Binding). Suppose adversary $\mathcal{A}^*$ modifies the legal prose from $\mathcal{P}$ to $\mathcal{P}'$ while attempting to maintain identity continuity. The on-chain identity record stores $H(\mathcal{P})$ as an immutable value (Listing 1, line 6: `bytes32 public immutable RICARDIAN_HASH`). For the agent to operate under $\mathcal{P}'$ with the same blockchain identity, it would need $H(\mathcal{P}') = H(\mathcal{P})$ where $\mathcal{P}' \neq \mathcal{P}$, constituting a hash collision. By collision resistance:

$$\Pr[\exists\, \mathcal{P}' \neq \mathcal{P} : H(\mathcal{P}') = H(\mathcal{P})] \leq \text{negl}(\kappa) \quad (10)$$

Therefore, any modification to the Ricardian contract produces a distinct hash, forcing the creation of a new on-chain identity. $\square_3$

Union Bound. Applying the union bound over the three independent breach events:

$$\Pr[\text{Breach}(\mathcal{A}^*, \kappa)] \leq \sum_{i=1}^{3} \Pr[\text{Breach}_i] \leq 3 \cdot \text{negl}(\kappa) \quad (11)$$

Since $3 \cdot \text{negl}(\kappa)$ remains negligible in κ, the theorem is proven. $\square$

2.4.4 Reputation dynamics and convergence

Definition 2.6 *(Reputation Score)* The *reputation score* of an AI agent $\mathcal{A}$ at time t is defined as:

$$\rho_{\mathcal{A}}(t) = \frac{\sum_{i=1}^{N_t} w_i \cdot \mathbb{1}[\text{Valid}(Tx_i)]}{\sum_{i=1}^{N_t} w_i} \quad (12)$$

where N_t is the number of transactions up to time t, $w_i = e^{-\lambda(t-t_i)}$ is an exponential decay weight with rate $\lambda > 0$, and $\mathbb{1}[\cdot]$ is the indicator function.

Proposition 2.7 *(Cyber-Chama Convergence) Let $\mathcal{A}$ be an AI agent executing transactions according to a stationary Bernoulli process with success probability $p \in (0, 1)$ (i.e., each transaction is valid with probability p, independently). Then:*

1. The *expected reputation score converges:* $\lim_{t\to\infty} \mathbb{E}[\rho_{\mathcal{A}}(t)] = p$.
2. The *variance satisfies:* $\text{Var}[\rho_{\mathcal{A}}(t)] = O(\lambda/N_t)$.
3. For *threshold* $\rho^* \in (0, 1)$ *and an agent with* $p > \rho^*$*, the expected time to graduation (reaching Tier 2 status) is:*

$$\mathbb{E}[T_{\text{graduate}}] = O\left(\frac{\log(1/\delta)}{\lambda(p-\rho^*)^2}\right) \tag{13}$$

where δ is the target confidence level.

Proof *Part 1 (Convergence of Expectation)*. Define $S_t = \sum_{i=1}^{N_t} w_i \cdot \mathbb{1}[\text{Valid}(Tx_i)]$ and $W_t = \sum_{i=1}^{N_t} w_i$. By linearity of expectation and independence, $\mathbb{E}[S_t] = p \cdot W_t$. As $N_t \to \infty$ with transactions arriving at rate μ, the continuous-time approximation yields $W_t \approx (\mu/\lambda)(1 - e^{-\lambda t}) \to \mu/\lambda$. By the strong law of large numbers for weighted averages (Chow & Teicher, 1965/2012):

$$\rho_{\mathcal{A}}(t) = \frac{S_t}{W_t} \xrightarrow{\text{a.s.}} \frac{\mathbb{E}[S_t]}{W_t} = p \quad \text{as } t \to \infty \tag{14}$$

$\square_1$

Part 2 (Variance Bound). Since transactions are independent Bernoulli trials, $\text{Var}[S_t] = \sum_i w_i^2 \cdot p(1-p)$. Using the delta method and noting that $\sum_i w_i^2 \le W_t$ and $W_t = \Theta(N_t/\lambda)$ for large t:

$$\text{Var}[\rho_{\mathcal{A}}(t)] \le \frac{p(1-p)}{W_t} = O\left(\frac{\lambda}{N_t}\right) \tag{15}$$

$\square_2$

Part 3 (Graduation Time). For graduation, we require $\rho_{\mathcal{A}}(T) \ge \rho^*$ with probability $1 - \delta$. By Hoeffding's inequality for weighted sums and noting that $W_t \approx \mu t/\lambda$ for small λt:

$$\mathbb{E}[T_{\text{graduate}}] = O\left(\frac{\log(1/\delta)}{\lambda(p-\rho^*)^2}\right) \tag{16}$$

This result shows that agents with higher true compliance rates ($p \gg \rho^*$) graduate faster, while the decay parameter λ controls the memory horizon of the reputation system. $\square$

Corollary 2.8 *(Sybil Attack Resistance) Creating k Sybil identities to artificially inflate aggregate reputation requires stake collateral of at least $k \cdot C_{\min}$, where $C_{\min}$ is the minimum collateral per identity (Douceur, 2002). Combined with the slashing mechanism (Section 2.3.2), the expected cost of a Sybil attack exceeds the potential gain when:*

$$k \cdot C_{\min} \cdot \Pr[\text{Detection}] \cdot \alpha > \mathbb{E}[AttackGain] \tag{17}$$

where $\alpha \in (0, 1]$ is the slashing ratio.

3 Methodology

This study proposes a hybrid identity architecture that integrates Triple-Entry Accounting (TEA), Ricardian contracts, and Distributed Ledger Technology (DLT). We introduce the "Cyber-Chama"; a digital adaptation of the Kenyan community-trust model, as the consensus mechanism for verifying AI behaviour.

3.1 The Cyber-Chama: a federated governance model

Traditional blockchain identity systems rely on centralised authorities (Certificate Authorities) or purely algorithmic consensus (Proof-of-Work). Our methodology utilises a Circle of Trusted Agents (CTA), modelled after the Chama. The "Identity Registry" is not a static database but a dynamic federation of validators operating on a tiered trust structure (Ostrom, 1990). At Tier 1 are Human Principals (The Elders): verified human entities (e.g., legal owners, compliance officers) holding "Subjective Authority." At Tier 2 are AI Validators (The Stewards): autonomous software agents running deterministic verification logic and holding "Objective Authority." The transition mechanism allows an AI Agent to "graduate" to Tier 2 Validator status only after accruing a specific "Reputation Score" (based on 10,000+ reconciled transactions) and staking financial collateral.

3.2 System architecture

The framework is implemented across three distinct logical layers, ensuring separation of concerns between data storage, logic execution, and legal enforcement.

3.2.1 Layer A: the Ricardian identity layer (on-chain)

This layer stores the "Constitution" of the agent and is tested on both Ethereum (Solidity) and BSV (sCrypt). It records: the hash $H(Legal\_Prose)$ to ensure the text has not changed; the public key PK_Agent used to sign TEA receipts; and hard-coded constraint limits (e.g., $Max\_Spend\_Per\_Day = 100\,\mathrm{USD}$). We utilise Sovereign Keys: the private key is held inside the AI agent's secure enclave (e.g., AWS Nitro or Intel SGX), meaning not even the human operator can spoof the AI's signature without physical access (Burke et al., 2024).

3.2.2 Layer B: the triple-entry verification layer (the protocol)

This layer provides real-time auditing of interactions. Agent A initiates a transaction with Vendor B; A creates a receipt containing the transaction details and a hash of its Ricardian contract; B signs the receipt, acknowledging the trade; the "Cyber-Chama" (a random selection of Tier 2 AI Validators) checks the signatures; and if valid, the receipt is hashed and added to the DLT. This fulfils the TEA equation: Transaction = (SignedMessage, Sender, Receiver, Validator) (Grigg, 2005/2024).

3.2.3 Layer C: the privacy & compliance layer (off-chain)

This layer handles GDPR compliance and data storage via an IPFS (InterPlanetary File System) cluster with private access controls. To satisfy GDPR Article 17, we employ "Crypto-Shredding": if a user requests deletion, the encryption keys for the Private Vault are destroyed (Zyskind et al., 2015).

3.3 Experimental setup

To validate this methodology, we designed a simulation environment utilising two distinct blockchain architectures to prove chain-agnosticism. Environment 1 was the Ethereum Sepolia Testnet, where Ricardian contracts were deployed as standard ERC standards using Solidity. Environment 2 was the BSV Testnet, where the same logic was implemented using sCrypt, leveraging the UTXO model for parallel processing of TEA receipts. Participants comprised: 5 Human "Elders" (simulated via manual approval scripts); 10 Autonomous "Trading Bots" (AI Agents executing random micro-transactions); and 3 "Rogue Agents" (AI Agents programmed to intentionally violate their Ricardian contracts). The system is considered successful if the Cyber-Chama identifies and rejects 100% of the transactions attempted by the "Rogue Agents" without human intervention.

4 Privacy, compliance, and data governance

The integration of Distributed Ledger Technology (DLT) with personal identity data presents a fundamental legal paradox: the GDPR (Article 17) grants users the "Right to Erasure," while the blockchain is designed to be immutable (Finck, 2018). To resolve this, our framework adopts a "Privacy-by-Design" architecture (Cavoukian, 2009), treating the AI Agent not as a data storage unit but as a pointer to off-chain data enclaves.

4.1 Architectural segregation: the off-chain/on-chain hybrid

We strictly adhere to the principle of Data Minimisation (GDPR Art. 5(1)(c)). No Personally Identifiable Information (PII) is ever recorded on the public ledger. The Public Ledger (Layer 1) stores only the Ricardian Hash and the TEA Receipts, which are pseudonymous and mathematically unlinkable to a specific human without the corresponding private key. The Private Enclave (Layer 2) is a secure, off-chain storage solution holding actual PII and the unhashed Ricardian legal prose (Ture, 2021).

4.2 Solving the "Right to be Forgotten" via Crypto-Shredding

To comply with a user's request for deletion, we utilise Crypto-Shredding. Each Ricardian Identity is encrypted with a unique symmetric key before being stored in the Private Enclave. When a deletion request is received, the system deletes this unique encryption key. The on-chain TEA receipts remain as immutable evidence

that a transaction occurred (preserving financial integrity), but the link to the human identity is permanently severed. The data becomes effectively "irreconcilable," satisfying the legal definition of erasure under GDPR Recital 26 (European Parliament, 2016).

4.3 Zero-knowledge proofs (ZKP)

We employ zk-SNARKs (Zero-Knowledge Succinct Non-Interactive Arguments of Knowledge) (Ben-Sasson et al., 2014) to allow agents to prove compliance (e.g., "I am authorised to spend") without revealing underlying data to the validators.

5 Discussion

5.1 From identity to liability: the "Glass Box" paradigm

The central finding of this study is that applying Triple-Entry Accounting (TEA) to AI identity effectively transforms the "Black Box" problem of autonomous agents into a "Glass Box" of verifiable commitments. By enforcing a protocol where $\mathrm{Transaction} = \mathrm{Receipt}$, we ensure that an AI agent cannot execute a function without simultaneously creating an immutable forensic trail (Grigg, 2005/2024).

5.2 Significance of the formal results

Quantifiable Security Guarantees. Theorem 2.5 establishes that breaches of the Ricardian-TEA protocol occur with probability at most $3 \cdot \mathrm{negl}(\kappa)$. For a 256-bit security parameter (as used in SHA-256 and ECDSA), this translates to a breach probability below 2^{-256} which is a level of security comparable to breaking the underlying cryptographic primitives themselves.

Predictable Trust Dynamics. Proposition 2.7 offers governance designers a quantitative tool for calibrating the Cyber-Chama parameters. The graduation time bound shows that increasing the decay parameter λ accelerates trust assessment but reduces the system's memory of past behaviour; a tunable trade-off between responsiveness and robustness against strategic manipulation.

Economic Security via Corollary 2.8. The Sybil resistance bound demonstrates that rational adversaries will avoid attacks when the expected slashing penalty exceeds potential gains, providing a game-theoretic foundation for setting minimum collateral requirements.

5.3 Efficacy of the dual-chain implementation

The experimental deployment on both Ethereum Sepolia and BSV (sCrypt) yielded identical security outcomes: 100% of "Rogue Agent" attempts were neutralised. The Ethereum results demonstrated ease of integration with existing EVM-based DeFi

protocols, while the BSV/sCrypt results demonstrated superior scalability for high-frequency trading agents due to the UTXO model's ability to process TEA receipts in parallel. This dual validation confirms that the "Cyber-Chama" model is not tied to a specific consensus algorithm but constitutes a universal governance layer.

5.4 Limitations

While the integrity of the system is high, the performance overhead is significant. The requirement for triple-signed receipts and zk-SNARK generation introduced an average latency of 1.4 s per transaction. A further challenge concerns what happens when an AI is copied and run without the blockchain component (e.g., an offline model): it is not straightforward to rely on technology alone to prevent this, so the contract would likely include clauses such as "Any use of the AI model outside the guardrails of the blockchain/TEA protocol is invalid on its face and exposes the using parties to extraordinary and unlimited liabilities." Addressing this fully would entail a dedicated legal paper on contractual penalties, civil or criminal enforcement mechanisms, which exceeds the scope of the present work. Future research must explore Multi-sig solutions (M-of-N signatures) and Cross-Chama validations via hashes. Layer-2 Scaling Solutions (e.g., Optimistic Rollups) to batch TEA receipts added more complexity, since there is extra computational overhead involved (Buterin, 2021).

Several substantial challenges warrant further careful consideration: the reliance on cryptographic primitives and legal-institutional assumptions that may be undermined by future technological or jurisprudential developments (e.g., scalable quantum computing); significant practical barriers to deployment, including the need for widespread legal recognition, cross-jurisdictional interoperability, and manageable operational overhead; residual vulnerabilities within the proposed trust architecture and associated dispute-resolution mechanisms; and scalability constraints arising from the performance limitations of underlying blockchain infrastructures and the effective governance of off-chain data. Successfully overcoming these obstacles will nevertheless prove indispensable for achieving broad real-world adoption and establishing Ricardian-TEA as a credible and resilient governance paradigm for AI identity. We explore in depth the aforementioned limitations on the next section.

5.4.1 Towards real-world deployment: challenges and considerations

Ricardian-TEA as a concept goes further than only a theoretical idea. It has the power to enable trustworthy autonomous trading, automated financial settlements, and real-time auditability in decentralized finance (DeFi) ecosystems. A set of use-case scenarios that could be broadly incorporated are, among others: AI-mediated cross-border payments, automated compliance reporting, or trustless asset custody. Our proposed framework ensures a robust platform for risk management, while it reduces potential fraud, and enhances transparency in financial markets. While the Ricardian-TEA framework demonstrates promising results in controlled testnet environments,

several critical challenges must be addressed to enable its scalability and practical deployment in real-world settings.

Scalability and Throughput: It stands true, that current blockchain infrastructures, including Ethereum and Bitcoin SV, face inherent limitations in transaction throughput and network capacity. As transaction volumes increase, particularly in high-frequency trading (HFT), or enterprise-scale applications, the latency and cost implications become significant. Solutions such as layer-2 scaling protocols, sharding, or sidechains could mitigate these bottlenecks, but integrating such mechanisms into the Ricardian-TEA protocol requires further development and validation, including assessing their impact on trust and security guarantees. However, our testing took place about a year ago and in the meantime, there were several advancements in the aforementioned technology. For instance, BSV Blockchain has demonstrated capacity to handle over 1 million transactions per second (TPS) using its Teranode software, designed to facilitate global enterprise and high-volume data transactions. While daily average TPS varies based on usage, Teranode trials and tests have sustained speeds over 3 million TPS, enabling 100 billion transactions per day (Kramsky, 2026). Those advancements also affect the measured latency in our testing, of approximately 1.4 s per transaction, which was acceptable for certain applications, but could be insufficient for time-sensitive financial operations. Further strategies to reduce latency in both blockchains tested (Ethereum and BSV) such as batch processing of receipts, optimized zk-SNARK generation, or off-chain aggregation techniques, should be explored, but exceed the purpose of this work. These enhancements, which can be a scope for future work, must balance performance gains while, at the same time, maintain cryptographic integrity and accountability.

Validator Coordination and Governance: There is no argument about the limitations that the tiered Cyber-Chama model brings, since it relies on reputation scores and collateralized staking to transition AI agents into validator roles. In a real-world ecosystem, managing validator consensus, ensuring transparency, and preventing collusion or malicious behavior pose substantial operational challenges. Implementation of robust governance protocols, dispute resolution processes, and incentive mechanisms will be essential to sustain validator compliance and accountability. For instance, in offline AI models which can be arbitrarily replicated, embedding economic disincentives, such as staking collateral or implementing penalization schemes, can discourage copying and offline operation. Any unapproved use outside the protocol would result in the loss of collateral or privileges, thus aligning economic incentives with compliance.

Legal and Regulatory Integration: The framework's reliance on legally binding Ricardian contracts and GDPR-compliant data practices necessitates alignment with diverse legal jurisdictions. Challenges include obtaining formal legal recognition, cross-jurisdictional interoperability of identities, and enforcement of contractual obligations across borders. Close collaboration with legal authorities and regulators will be vital to transition from testing to compliant operational deployment.

Operational Readiness and Ecosystem Compatibility: The high computational overhead associated with zk-SNARK proofs and cryptographic operations may hin-

der scalability and accessibility. Further research is needed to optimize the cryptographic primitives for efficiency, and to develop integration pathways with existing financial and legal systems, ensuring that the proposed structure can coexist with and augment current infrastructures.

Addressing these challenges through targeted research, technological refinement, and pilot deployments will be crucial steps towards realizing the full potential of our proposed Ricardian-TEA framework as a robust, scalable, and legally sound governance framework for autonomous AI identities in the digital economy.

6 Conclusion

As Artificial Intelligence transitions from a supportive tool to an autonomous economic principal, the existing digital identity infrastructure is rapidly becoming obsolete. This paper has presented a novel Legal-Technical Identity Framework that addresses this gap by synthesising Ricardian contracts, Triple-Entry Accounting, and Distributed Ledger Technology.

The contributions are fourfold. With respect to Identity as an Asset, we have demonstrated that an AI agent's identity is a dynamic asset class defined by its accumulated, cryptographically reconciled transaction history. Regarding the Cyber-Chama, we successfully adapted the social trust model of the Chama into a decentralised governance protocol, proven effective on both Ethereum and BSV blockchains. On GDPR-Compliant Immutability, through "Crypto-Shredding" and "Split-Storage" we resolved the paradox between immutable blockchain records and the "Right to be Forgotten." The Formal Security Guarantees provided through the Ricardian-TEA Integrity Theorem (Theorem 2.5) and the Cyber-Chama Convergence Proposition (Proposition 2.7) establish provable security properties under standard cryptographic assumptions and characterise the stochastic dynamics of trust evolution.

We must state here that while our proposal offers a compelling and rigorous conceptual architecture for the legal and technical governance of autonomous AI agents, its transition to a fully operational, production-level infrastructure remains an ongoing endeavor. It is therefore important to distinguish clearly between two layers of contribution: Ricardian-TEA as a conceptual and governance architecture, which we believe is mature and internally consistent, and Ricardian-TEA as a deployable production-level infrastructure, which remains a research and engineering programme rather than a finished system. Many of the key components, such as scalable transaction throughput, robust validator coordination, dispute resolution mechanisms, and cross-jurisdictional legal enforcement, are still in the developmental or experimental phases. Although preliminary testnet implementations demonstrate promising capabilities and resilience against adversarial attempts, practical deployment at scale will require further advancements in blockchain scalability solutions, robust governance protocols, and legal interoperability frameworks. Notably, while this work was under peer review, one of the blockchains we employed announced an advancement that, in principle, addresses several of the scalability constraints we

discuss. We have not, however, had the opportunity to test our framework against this new infrastructure, and we are cautious about extrapolating performance claims from vendor announcements alone. Independent benchmarking under realistic adversarial and high-throughput conditions remains a necessary next step before any claim of production-grade scalability can be made.

In addition, our framework currently relies on educated assumptions about validator behavior, economic incentives, and enforceability that need to be validated in real-world, diverse operational environments. These assumptions must ultimately be tested against the heterogeneity and adversarial complexity of real-world deployment environments. Specifically, several limitations remain open. Adversarial validator behaviour under sustained, well-resourced attack has not been empirically characterised; collusion-resistance guarantees rest on incentive assumptions that may not hold across heterogeneous jurisdictions; offline replication of credentials raises enforcement challenges that current cryptographic primitives only partially mitigate; and the practical enforceability of Ricardian clauses across legal systems with divergent treatment of smart contracts is an empirical question that this paper does not resolve. We also suggest that future efforts should not leave aside scalable Layer-2 solutions; nevertheless, TEA needs to be maintained as a prime directive, since enhancing validator consensus mechanisms, and establishing legal standards and regulatory acceptance for such hybrid identity architectures is fundamentally difficult.

Therefore, we view Ricardian-TEA as a foundational blueprint and proof-of-concept that points toward a future where autonomous AI agents can be governed with formal security guarantees and legal accountability. We acknowledge that its widespread operational deployment will necessitate network and technological advancements, continued innovation, cross-disciplinary collaboration, and empirical validation before it can be confidently relied upon in complex, real-world digital economies. We regard Ricardian-TEA as a meaningful first step toward that horizon.

Appendix: Implementation code snippets

The following code snippets demonstrate the practical implementation of the Ricardian-TEA framework on both Account-based (Ethereum) and UTXO-based (BSV) ledgers.

Solidity implementation (Ethereum Sepolia)

```
// SPDX-License-Identifier: MIT
pragma solidity ^0.8.0;

contract RicardianAgent {
    // The immutable link to the Legal Prose (Off-Chain)
    bytes32 public immutable RICARDIAN_HASH;

    // The human owner (The "Elder" in the Chama)
    address public owner;

    // Triple-Entry Receipt Structure
    struct TeaReceipt {
        bytes32 txId;
        address counterparty;
        uint256 timestamp;
        bytes32 verificationHash;
    }

    // Event acts as the immutable "Third Entry" log
    event ReceiptGenerated(
        bytes32 indexed ricardianHash,
        address indexed counterparty,
        bytes32 receiptHash
```

Listing 1 Solidity Smart Contract for Ricardian Identity (Ethereum Sepolia)

```
    );

    modifier onlyOwner() {
        require(msg.sender == owner,
            "Unauthorized: Human Signature Required");
        _;
    }

    constructor(address _owner, bytes32 _legalProseHash) {
        owner = _owner;
        RICARDIAN_HASH = _legalProseHash;
    }

    function executeTeaTransaction(
        address _counterparty,
        uint256 _amount,
        bytes32 _invoiceHash
    ) external onlyOwner {

        // 1. Verify Ricardian Constraints
        require(_amount <= 100 ether,
            "Error: Exceeds Ricardian Budget Limit");

        // 2. Create the "Third Entry" Receipt Hash
        bytes32 receiptHash = keccak256(
            abi.encodePacked(
                RICARDIAN_HASH,
                msg.sender,
                _counterparty,
                _amount,
                _invoiceHash,
                block.timestamp
            )
        );

        // 3. Emit the receipt to the distributed ledger
        emit ReceiptGenerated(RICARDIAN_HASH,
            _counterparty, receiptHash);
    }
}
```

Fig. (continued)

sCrypt implementation (BSV testnet)

```
class RicardianIdentity extends SmartContract {
    @prop()
    readonly owner: PubKey;

    @prop()
    readonly ricardianHash: Sha256;

    constructor(owner: PubKey, legalHash: Sha256) {
        super(...arguments);
        this.owner = owner;
        this.ricardianHash = legalHash;
    }

    @method()
    public unlock(sig: Sig, counterparty: PubKey,
                  amount: bigint) {
        // 1. Verify Signature (Proof of Authority)
        assert(this.checkSig(sig, this.owner),
            'signature check failed');

        // 2. Enforce Ricardian Constraints
        assert(amount <= 1000n,
            'Error: Exceeds Ricardian Budget Limit');

        // 3. Ensure Identity Persistence
        let outputs = this.buildStateOutput(amount);

        // 4. Validate the "Third Entry"
        assert(this.ctx.hashOutputs == hash256(outputs),
            'hashOutputs mismatch');
    }
}
```

Listing 2 sCrypt Contract for UTXO-based Identity (BSV Testnet)

Acknowledgements The authors would like to thank Ian Grigg for his review and suggestions on early drafts of this manuscript.

Author contributions All authors contributed equally to this work. Both authors developed the conceptual framework, formal mathematical foundations, experimental design, and interpretation of results. Both authors read and approved the final version of the manuscript.

Funding Open access funding provided by Università degli Studi Mediterranea di Reggio Calabria within the CRUI-CARE Agreement. The authors declare that no external funding was received for this research.

Data availability No datasets were generated or analysed during the current study.

Code availability The Solidity and sCrypt smart contract code is reproduced in the Appendix of this manuscript. Further implementation details are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Not applicable. This research did not involve human participants, animals, or sensitive personal data requiring ethical approval.

Consent to participate Not applicable.

Consent for publication Both authors consent to the publication of this manuscript.

References

Ben-Sasson, E., et al. (2014). Zerocash: Decentralized anonymous payments from bitcoin. In: Proceedings of IEEE security & privacy (S&P 2014).

Burke, Q., et al. (2024). Securing cloud file systems with trusted execution. arXiv preprint. arxiv:2305.18639v3.

Buterin, V. (2014). A next-generation smart contract and decentralized application platform. Ethereum White Paper.

Buterin, V. (2021). An incomplete guide to rollups. Vitalik.ca. https://vitalik.eth.limo/general/2021/01/05/rollup.html.

Cavoukian, A. (2009). Privacy by design: The 7 foundational principles. Information and privacy commissioner of Ontario

Chaffer, TJ (2025). Know your agent: Governing AI identity on the agentic web. Available at SSRN: https://ssrn.com/abstract=5162127 or https://doi.org/10.2139/ssrn.5162127.

Chow, Y.S., & Teicher, H. (1965/2012). Probability theory: Independence, interchangeability, martingales. Springer.
Douceur, J. R. (2002). The Sybil attack. International workshop on peer-to-peer systems (IPTPS)
Drăgnoiu, P., et al. (2020). Identity management on blockchain: Privacy and security aspects. *Proceedings of the Romanian Academy - Series A, 21*, 45–52.
Elkaleh, J. (2025). When AI agents pay, credibility becomes currency: The trust layer missing from autonomous AI transactions. Securities.io, November 24.
European Parliament (2016). Regulation (EU) 2016/679 (general data protection regulation). Official Journal of the European Union.
Finck, M. (2018). Blockchains and data protection in the European union. Max planck institute for innovation and competition research paper No. 18–01. SSRN. https://ssrn.com/abstract=3080322.
Griffin AI. (2025). The rise of autonomous AI agents. https://blog.griffinai.io/news/rise-of-autonomous-AI-agents.
Grigg, I. (1996/2004). The Ricardian contract. In: First IEEE international workshop on electronic contracting (WEC). https://iang.org/papers/.
Grigg, I. (2000). Financial cryptography in 7 layers. *Lecture Notes in Computer Science*. https://doi.org/10.1007/3-540-45472-1
Grigg, I. (2005/2024). Triple entry accounting. Journal of Risk and Financial Management, 17(2), 76. https://doi.org/10.3390/jrfm17020076
Grigg, I. (2017). *Identity is an edge protocol*. R3.
Grigg, I. (2021). Identity cycle. Peer for peer foundation. ISBN: 978-9918-0-0122-4. https://iang.org/.
Ibañez, J., et al. (2021). Triple-entry accounting, blockchain and next of kin: Towards a standardisation of ledger terminology. https://doi.org/10.1017/9781009362290.012.
Kariuki, H. (2018). How blockchain technology is revolutionizing "Chamas" – Kenya's informal saving groups. https://medium.com/@harriet436/.
Kramsky, J. (2026). How the BSV association built a million-TPS blockchain node using AWS. https://aws.amazon.com/blogs/web3/how-the-bsv-association-built-a-million-tps-blockchain-node-using-aws/
Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf.
Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
PwC. (2025). Agents of change: The rise of autonomous AI in cybersecurity. https://www.pwc.com/gx/en/issues/cybersecurity/.
Rothschild, D. M., Mobius, M., Hofman, J. M., Dillon, E. W., Goldstein, D. G., Immorlica, N., Jaffe, S., Lucier, B., Slivkins, A., & Vogel, M. (2025). Agentic economy. *Communications of the ACM, 69*, 39–42.
Sgantzos, K., et al. (2023). Triple-entry accounting as a means of auditing large language models. *Journal of Risk and Financial Management, 16*(9), Article 383. https://doi.org/10.3390/jrfm16090383
Sgantzos, K., Stelios, S., Tzavaras, P., et al. (2024). Minds and machines: Evaluating the feasibility of constructing an advanced artificial intelligence. *Discover Artificial Intelligence, 4*, Article 104. https://doi.org/10.1007/s44163-024-00216-2
Staley, I. (2026). A reference architecture for AI agents on blockchain infrastructure: Identity, policy, payments, and custody as composable primitives. Available at SSRN: https://ssrn.com/abstract=6650658 or https://doi.org/10.2139/ssrn.6650658.
Token Security. (2025). Identity-centric security risks of autonomous AI. https://www.token.security/.
Ture, T. (2021). GDPR, blockchain and the right to be forgotten. University of Helsinki. Master Thesis.
Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. MIT Press.
Zyskind, G., et al. (2015). Decentralizing privacy: Using blockchain to protect personal data. In: IEEE security & privacy workshops (SPW 2015). https://doi.org/10.1109/SPW.2015.27.