

Article

Can We Trust Autonomous Agents? A Formal Framework for Verifiable Agent Economies

Md Monjurul Karim  and Qiang Qu ^{*} 

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China;
karim@siat.ac.cn

^{*} Correspondence: qiang@siat.ac.cn

Abstract

Autonomous AI agents discover services, negotiate, and settle payments. However, open agent markets lack reliable means to verify delivery, constrain delegated authority, and make identity renewal economically consequential. To address these failures, we formalize the *verifiable agent economy*, where evidence and programmable constraints govern identity, behavior, and payment. We introduce the agentic economy stack and formulate trust-asymmetric coordination through four requirements: delivery incentive compatibility, verifiable settlement, budget and authority safety, and Sybil resistance. To operationalize these requirements, we construct a blockchain-based mechanism combining verifiable credentials, stake, reputation, constrained wallets, and receipt-conditioned settlement. Within the class of receipt-conditioned payment rules, its geometric-mean payment rule minimizes worst-case multiplicative exposure to residual value uncertainty. The mechanism satisfies delivery incentive compatibility and Sybil resistance under explicit deterrence and identity-cost conditions, while its budget and authority guarantee is limited to actions routed through the enforcement path. Across five detection laws, numerical analysis identifies an interior verification-budget optimum. At their respective optima, our mechanism achieves 46% higher welfare than Verification-Only and raises honest delivery from 77% to 99%. A measured sampling-audit verifier shifts but preserves the optimum, while false positives and thin markets tighten participation and staking requirements.

Keywords: verifiable agent economy; autonomous AI agents; autonomous payments; trust-asymmetric coordination; decentralized identity



Academic Editor: Anthony Jnr. Bokolo

Received: 31 July 2026

Revised: 23 September 2026

Accepted: 28 September 2026

Published: 1 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Autonomous AI agents are moving beyond task automation toward delegated economic action [1]. They can discover services, negotiate offers, and initiate payments on behalf of human or organizational principals [2,3]. Consider a flight-booking workflow in which an assistant agent selects and purchases a service without human intervention. A valid confirmation can still conceal three failures: the service agent may deliver a lower-quality option than promised, the assistant may overpay for an unverifiable claim, or a previously reliable service agent may underperform once the immediate gain exceeds the continuation value of its reputation. These failures arise not simply from errors in agent reasoning but from delegating economic authority before identity, conduct, and delivery can be reliably verified.

Although these failures share a trust problem, they require different safeguards. Misdelivery is a delivery-discipline problem because settlement can confirm an output without

establishing that it satisfies the contracted quality [4–6]. Overpayment is a delegated-authority and pricing problem: negotiation benchmarks report financial losses by purchasing agents [2], while constrained-wallet protocols limit the spending and actions an agent may authorize [7]. Strategic underperformance after a compliant history is a continuation-value problem, combining unobservable effort [8] with a reputation that can be abandoned through identity renewal [9]. Together, these failures give rise to a condition we term *trust-asymmetric coordination*, in which principals delegate economically consequential authority to agents whose identity, capability, effort, and execution cannot be adequately observed or verified by the delegating party.

Advances in agent architecture, reasoning, and tool use make this coordination problem increasingly consequential [10,11]. Existing research addresses its components at different layers: conceptual architectures describe agent-economy institutions [1,12,13]; empirical studies investigate emerging agent-mediated markets and protocols [14,15]; and payment, identity, and wallet standards provide primitives for settlement, reputation, and delegated control [3,16]. These streams do not yet jointly formalize four interdependent requirements: disciplining delivery, pricing residual uncertainty, enforcing delegated authority, and making identity renewal economically consequential. The gap is therefore not the absence of primitives, but the absence of a formal account of how they must be composed, under which conditions they support trustworthy coordination, and what exposure remains when verification is incomplete.

To address this gap, we ask: *What requirements define a verifiable agent economy, and under what conditions does a composed coordination mechanism satisfy them?* We define the *verifiable agent economy* as an institutional regime in which evidence and programmable constraints govern economic exchange among autonomous agents. We introduce the agentic economy stack to organize the relevant primitives, model trust-asymmetric coordination as a game, and use this framework to construct a mechanism that satisfies four coordination requirements. Each guarantee holds under a specific scope. Within the class of receipt-conditioned payment rules, geometric-mean settlement minimizes residual exposure. Delivery incentive compatibility and Sybil resistance follow when the stated deterrence and identity-cost conditions hold, while budget and authority safety follows when transactions use the prescribed enforcement path. These guarantees apply to digital transactions in which delivered artifacts provide verifiable settlement evidence. Embodied agents therefore remain outside the theorem scope. We evaluate the mechanism through theorem checks, model-based simulations, and an implementation-based sampling-audit verifier. The results identify an interior optimum for the verification budget and show how verifier quality, false-positive rates, and market thickness shape coordination outcomes.

This paper makes four contributions:

- We operationalize the *verifiable agent economy* through cumulative criteria for autonomous economic actors and the agentic economy stack, which organizes coordination, trust, settlement, human integration, and governance.
- We formulate trust-asymmetric coordination as a game subject to delivery incentive compatibility, verifiable settlement, budget and authority safety, and Sybil resistance. We also identify offer-stage misrepresentation, assistant-side incentives, collusion, and specification gaming within the verified evidence boundary as explicit non-claims.
- We construct a mechanism combining credentials, stake, reputation, constrained wallets, and receipt-conditioned settlement. We establish exposure optimality within the stated rule class and derive the deterrence, identity-cost, participation, and enforcement conditions governing the remaining guarantees.
- We evaluate the formal and implementation-level behavior of the mechanism using theorem checks, simulations, verifier measurements, and ablations. The results dis-

tinguish structural properties from outcomes that depend on verifier quality, market thickness, and parameter calibration.

The remainder of the paper is organized as follows. Section 2 reviews prior work. Section 3 defines the verifiable agent economy and presents the agentic economy stack. Sections 4 and 5 formulate the coordination problem and analyze the mechanism. Section 6 reports the numerical analysis, and Sections 7 and 8 discuss the implications and conclude.

2. Related Work

2.1. Prior Framings of the Agent Economy

Research on agent economies spans economic principles [17–19], market implications [1,20], and architectural proposals [13,21–23]. Tomasev et al. [12] classify sandbox economies and propose safety principles without specifying a coordination mechanism. Xu [13] organizes identity, cognition, settlement, and governance in a layered blockchain architecture, whereas Baxi [24] bounds exposure by conditioning permissions on robustness assessment. Empirical studies investigate emerging economic conditions and payment and identity protocols [14,15]. These works provide concepts, architectures, bounds, and observations relevant to our problem but address different combinations of assumptions and outcomes.

The term *verifiable agent economy* also appears in discussions of ERC-8004 [16,25]. There, it denotes a registry-stack objective rather than an operational definition of an economic regime. Building on those identity, reputation, and validation primitives, we specify cumulative criteria for the regime and four constraints for exchange under hidden action, residual value uncertainty, delegated authority, and renewable identity. The novelty claim is limited to formalizing this composition, proving the receipt-conditioned exposure result within its rule class, and deriving explicit conditions for the remaining guarantees; it does not extend to the individual primitives.

2.2. Agent Payments, Verification, and Trust Infrastructure

Technical research provides complementary primitives for payments, verification, identity, and delegated control. Programmable payment protocols can condition transfers on attested computation [3]. Payment systematizations and x402 security analyses identify the boundary of that guarantee: finality can confirm a transfer against the available delivery evidence without establishing that the output meets the contracted quality [5,6,26]. Verification-native clearing escalates high-exposure cases to arbitration [27], while reliable delivery checking requires costly auditing [4]. ERC-8004-style registries provide identity, reputation, and validation [16,25]; constrained wallets bound an agent's spending authority [7]; and surveys map these components across Web3 and AI-agent systems [28]. Together, these mechanisms address distinct failure surfaces but do not jointly determine the verification budget, the evidence-conditioned settlement price, or the stake and reputation needed to deter delivery shortfalls. Our model links these design choices rather than replacing the underlying protocols.

2.3. Coordination, Negotiation, and Trust in Multi-Agent Markets

Negotiation benchmarks [2], agent-market environments [29], platform theory [30], and controlled studies [31–36] provide complementary accounts of agents as market participants and of the markets in which they interact. Pricing agents can learn collusive strategies [37–41], although the robustness of such behavior is contested [42,43]. Blockchain-based coordination mechanisms add incentives, reputation, and zero-knowledge verification [44–46], while surveys identify their security boundaries [9,47,48]. This literature shows why capability and prior compliance alone cannot establish reliable conduct. We

narrow that insight to post-execution delivery discipline by linking observable evidence to immediate and continuation consequences; the resulting guarantees hold only under the stated conditions and do not establish truthfulness or collusion resistance.

2.4. Economic Foundations Revisited for Autonomous Agents

The framework draws on, rather than replaces, established economic theories. Agency theory motivates the hidden-action constraint and the use of continuation value to discipline effort [8,49]. Transaction-cost economics explains why verification may remain costly even when agents reduce search and bargaining costs [50–52]. Institutional economics motivates explicit rules for identity, reputation, and enforcement, while unresolved legal status limits what technical enforcement can establish [53]. Sections 3.2 and 4 translate these foundations into design constraints; Section 7.2 develops the questions that remain outside the formal mechanism.

2.5. Comparative Coverage and Remaining Gap

Table 1 makes the preceding synthesis explicit across seven comparison dimensions. The table records specialization rather than deficiency: conceptual studies emphasize definitions and architectures, formal studies isolate selected mechanisms and bounds, and empirical studies provide observed evidence. The remaining gap is their joint formalization, not the absence of any individual component. Section 3 next operationalizes that integration through the definition and layered reference model.

Table 1. Coverage of seven comparison dimensions across representative prior work. ✓ fully addressed; ~ partially addressed; × not addressed. The map records scope and evidence rather than overall quality, so × means only that a work does not target that dimension. For this work, ~ under empirical and implementation evidence denotes model simulations and an implementation-based sampling audit, without field validation of a deployed system.

Work	Formal Definition	Layered Model	Formal Problem	Formal Mechanism Analysis	Verification-Based Trust	Empirical/Implementation Evidence	Economic Grounding
Rothschild et al. [1]	~	×	×	×	×	×	~
Tomasev et al. [12]	~	~	×	×	~	×	~
Xu [13]	✓	✓	×	×	✓	×	~
Gondauri & Batiashvili [14]	✓	×	×	×	×	✓	✓
Jin et al. [15]	×	×	×	×	×	✓	×
Baxi [24]	~	~	✓	✓	~	~	~
ERC-8004 discourse [25]	×	×	×	×	✓	~	×
This work	✓	✓	✓	✓	✓	~	✓

3. The Verifiable Agent Economy

This section defines the institutional conditions under which an autonomous software agent functions as an economic actor. It distinguishes an agentic economy from task automation, identifies the verifiable subclass studied in this paper, and connects its actor criteria to the formal requirements developed in Section 4.

3.1. From Economies to Agent Economies

Economic exchange requires more than transacting actors. Authority to commit resources, mechanisms for value transfer, and institutions for identity, contracting, reputation, and enforcement make commitments attributable under incomplete information. Automation changes how a task is executed; a distinct economic regime arises when software receives bounded authority to control delegated resources, create attributable obligations, and accumulate a persistent market history.

Definition 1 (Agentic economy). *An agentic economy is an economic system in which autonomous AI agents exercise bounded authority delegated by human or organizational principals to participate in discovery, negotiation, exchange, and settlement. Their commitments and use of delegated resources are governed by institutional rules that create attributable obligations, exposure, and reputational consequences for their principals.*

An agent's economic role therefore depends on a persistent institutional position connecting identity, authority, settlement, evidence, and future market access.

Definition 2 (Economic actor). *An autonomous agent is an economic actor in an agentic economy if it possesses: (C1) a persistent identity, distinguishable from other agents and cryptographically or institutionally bound to a principal; (C2) delegated authority, expressed as an enforceable mandate over spending, permissible actions, counterparties, disclosure, and operating context; (C3) payment capability, through which it sends or receives value in settlement of obligations within that mandate; (C4) verifiable behavior, under which consequential actions produce attributable evidence evaluated against a specified claim and verification procedure; and (C5) a persistent, updatable reputation state, initialized at admission and revised through attributable conduct in ways that affect future interaction. The five criteria are cumulative; identity alone is insufficient.*

The criteria can operate under different institutional arrangements. In an operator-mediated regime, a central entity attests identity, records behavior, executes payment, and maintains reputation. In an open regime, participants cannot rely exclusively on one operator's integrity and continuity, so evidence, programmable constraints, and records resistant to unilateral revision must perform those functions.

Definition 3 (Verifiable agent economy). *A verifiable agent economy is the subclass of agentic economies in which criteria (C1) to (C5) are enforced through verifiable evidence and programmable constraints without sole reliance on a central operator. Identity rests on verifiable credentials; consequential conduct produces attributable evidence; authority and payment follow machine-enforceable limits and settlement rules; and reputation persists in records resistant to unilateral revision and inexpensive reset. Trust is asymmetric because principals and counterparties bear unequal exposure while lacking equivalent access to agent type, policy, effort, and execution context. Verification does not remove that asymmetry; it makes specified conduct and exposure limits enforceable despite incomplete observability.*

Definition 3 is operational rather than nominal. Building on Xu's architectural treatment [13], it specifies the criteria that the constraints in Section 4 and the proofs in Section 5 operationalize.

Figure 1 presents four analytical modes of digital economic coordination rather than a claim that every system follows a linear historical sequence. The first records human negotiation and contracting. The second adds platform-based discovery and matching while humans remain the economic actors. The third automates market actions through principal- or platform-owned accounts, to which identity, budget, reputation, and liability remain attached. In the fourth, the software agent itself acquires persistent identity, bounded authority, settlement capability, reputation, and an evidence-producing accountability path. Generic multi-agent systems, algorithmic markets, and gig platforms do not require all five criteria and therefore need not constitute agentic economies under Definition 1.

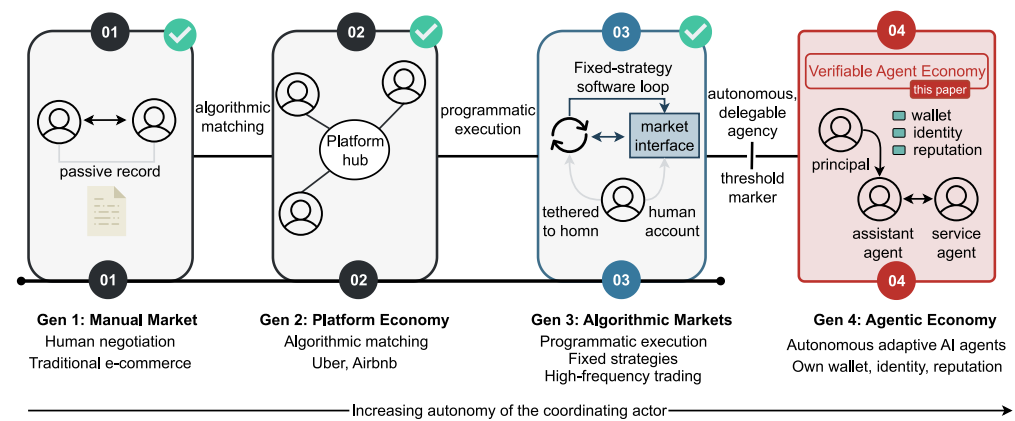


Figure 1. Four analytical modes of digital economic coordination. An agentic economy arises when a software agent acquires persistent identity, bounded delegated authority, settlement capability, reputation, and an accountable evidence trail.

3.2. Economic Departures and Formal Design Requirements

The agentic economy makes explicit institutional functions that conventional exchange often assigns to people, organizations, or law. Four departures motivate the constraints in Section 4. First, agency theory treats hidden action under misaligned incentives; autonomous agents add opaque policies and specification-sensitive behavior, which motivates delivery incentive compatibility based on verifiable conduct and persistent consequences. Second, a digital receipt can establish delivery without establishing contracted quality, so verifiable settlement must connect payment to evidence while accounting for verification cost. Third, autonomous execution may precede effective *ex post* remedies, which makes machine-enforceable budget and authority limits necessary. Fourth, digital identities can be recreated and adverse histories discarded, so Sybil resistance must make identity multiplication and reputation reset economically consequential.

One additional departure informs the research agenda without creating another formal constraint. Rational-choice models generally presume a stable objective, whereas language-model agents can display context-dependent behavior [54]. The framework therefore disciplines observable conduct rather than presuming that latent objectives are known. Questions concerning agents' labor-like and capital-like properties are reserved for the research roadmap and do not enter the present guarantees.

3.3. The Agentic Economy Stack

The agentic economy stack (shown in Figure 2) locates the criteria and their enforcement mechanisms within three functional planes. Coordination covers policy, discovery, negotiation, matching, and strategic action; trust covers identity, reputation, and behavior verification; and settlement covers wallets, delegated limits, and conditional payment. Human integration provides override, escalation, accountability, and distributional transparency, while governance controls admission, pricing, sanctions, and rule revision. Table 2 links the six layers to their governing theories and principal failure modes.

The formal constraints cross these layers: identity and admission support Sybil resistance; policy, verification, and reputation support delivery incentive compatibility; verification and payment support verifiable settlement; and wallets, governance, and human oversight support budget and authority safety. Actor criteria and layer indices remain distinct. Identity (C1), authority (C2), payment (C3), and reputation (C5) map primarily to L3, L1, L2, and L4, while verifiable behavior (C4) spans execution and settlement evidence. This mapping connects the reference architecture to the formal problem rather than treating the stack as a descriptive taxonomy.

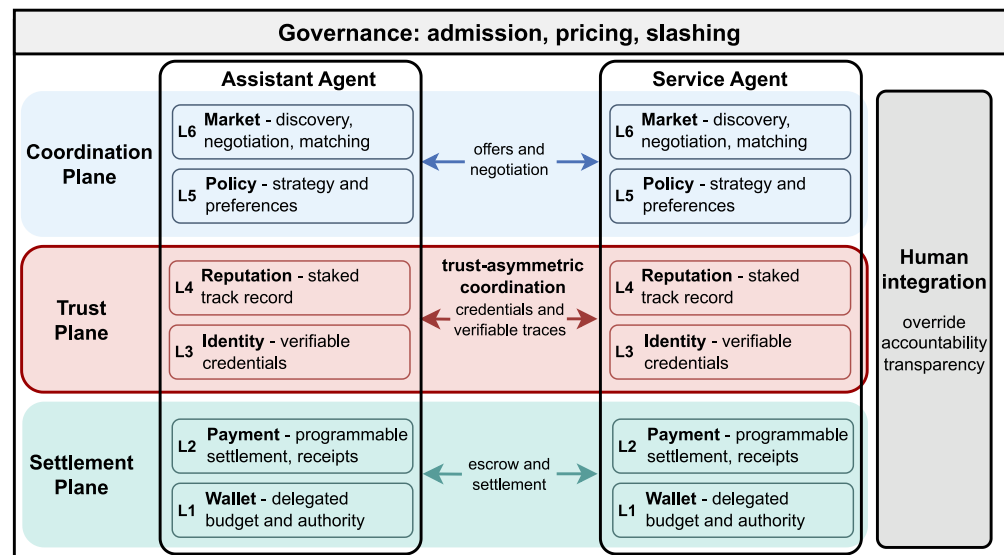


Figure 2. The agentic economy stack. Autonomous agents interact through coordination, trust, and settlement planes under cross-cutting human integration and an enclosing governance frame. The emphasized region identifies the failure surfaces formalized by the trust-asymmetric coordination problem.

Table 2. The six layers of the agentic economy stack, their functional planes, governing theories, and principal failure modes. Human integration and governance cut across all layers.

Layer	Plane	Governing Theory	Primary Failure Mode
L1 Wallet	Settlement	Agency theory	Over-delegation, authority misuse
L2 Payment	Settlement	Fair exchange, transaction-cost economics	Payment and quality decoupling
L3 Identity	Trust	Institutional economics	Sybil, shallow registration
L4 Reputation	Trust	Information economics	Reputation forgery, whitewashing
L5 Policy	Coordination	Rational choice, game theory	Manipulated or unstable preferences
L6 Market	Coordination	Mechanism design, repeated games	Collusion, manipulation

3.4. Scope and Boundary Conditions

Three clarifications delimit the framework. First, criteria (C1) to (C5) are cumulative, but criterion (C5) requires a persistent state rather than an established history: reputation is initialized at admission and updated through attributable conduct. Off-chain operation remains in scope when consequential actions preserve the same identity, evidence, authority, settlement, and sanction semantics. Second, economic-actor status grants neither legal personhood nor direct liability; it denotes an operational role under delegated authority. Non-pecuniary objectives remain within scope only when the agent controls delegated resources and enters value-bearing transactions. Third, the extension to embodied agents is conceptual and is not established by the results of this paper. No theorem, proposition, or corollary below is stated over physical action, the constraint set gains no additional member for it, and the numerical study contains no embodied experiment. The narrower claim is that criteria (C1) to (C5) and the four constraints can be expressed for embodied agents if a trustworthy physical evidence chain supplies what a digital artifact supplies in the present setting. Whether such a chain can be established, and at what cost, remains open and is developed in Section 7.3.

4. System Model and Problem Formulation

We model delegated exchange among consumer principals, assistant agents, and service agents as a repeated Bayesian interaction with partially observable service type, effort, and delivery. The formulation separates four requirements: delivery incentives, evidence-conditioned settlement, delegated-authority enforcement, and resistance to identity multiplication.

4.1. Agents, Types, and the Interaction Game

Assistant agents $\mathcal{A}$ act for consumer principals and service agents $\mathcal{S}$ for provider principals, with $\mathcal{A} \cap \mathcal{S} = \emptyset$. They interact at $t = 1, 2, \dots$ in a repeated market, so reputation can affect future opportunities.

Definition 4 (Agent as economic player). *An agent is a tuple $i = (\rho_i, \psi_i, u_i, \mathcal{I}_i, B_i, R_i, \tau_i)$, where ρ_i is its persistent identity (C1), ψ_i its policy, u_i its utility function, $\mathcal{I}_i$ its information set, B_i its delegated budget and authority envelope (C2, C3), R_i its reputation state (C5), and τ_i the attributable evidence generated by consequential actions (C4). The tuple operationalizes the economic-actor criteria of Definition 2.*

Each period begins with task context c , promised quality $\bar{q}$, a delegated envelope, and principal valuation $v(\cdot)$. Discovery identifies candidates [55]; negotiation selects an offer; execution produces output o at private cost $C_i(o)$ and trace τ_i ; and settlement verifies the specified evidence, transfers payment, applies sanctions, and updates reputation. Before contracting, the principal commits the settlement-relevant value bounds or the rule from which they are derived. Subjective value outside that commitment is not priced by the mechanism.

Service agent i privately observes type $\theta_i = (q_i, C_i)$ drawn from common prior F , where q_i denotes its feasible quality and C_i its cost function. The principal observes delivery only through mechanism-exposed evidence, and the service agent does not observe the principal's full valuation. Trust asymmetry therefore combines delegated control with unequal information.

Assumption 1 (Self-interest and rationality). *Each service agent selects policy ψ_i to maximize expected discounted utility under common prior F , information set $\mathcal{I}_i$, and discount factor $\delta \in (0, 1)$. Beliefs update by Bayes' rule on histories reached with positive probability. Honesty is not assumed; compliant delivery must instead be sequentially rational.*

This best-response model is an idealization; Section 6.7 separately tests agents that learn from realized rewards. Let $\ell_i \geq 0$ denote realized slashing loss and $z \in \{0, 1\}$ indicate whether verification is invoked, with z_t its value in period t . A service agent receiving payment p obtains

$$u_i = p - C_i(o) - \ell_i, \quad (1)$$

while the principal obtains

$$u_p = v(o) - p - c_v z. \quad (2)$$

Reputation affects continuation utility; verification consumes resources, while residual mispricing enters the exposure criterion below.

4.2. Settlement, Value, and Verification

A receipt establishes an observable event but need not identify the quality or value relevant to settlement.

Definition 5 (Receipt, value, and verification). Let O be the output space and $\text{Receipt} : O \rightarrow \mathcal{R}$ map each output to settlement-visible digital evidence, such as an execution attestation, content hash, or signed service receipt. Receipt r induces the fiber

$$\mathcal{F}(r) = \{o \in O : \text{Receipt}(o) = r\}. \quad (3)$$

The committed settlement value satisfies $v(o) > 0$. Outputs in the same fiber remain indistinguishable to bare settlement even when their values differ. At cost $c_v \geq 0$, a verifier partitions each receipt fiber into residual fibers. For residual fiber r' , let $[v_{\min}(r'), v_{\max}(r')]$ be its value interval, with $v_{\min}(r') > 0$. The verifier's worst-case multiplicative spread is

$$s(c_v) = \sup_{r'} \frac{v_{\max}(r')}{v_{\min}(r')} \geq 1. \quad (4)$$

The case $c_v = 0$ corresponds to the bare receipt; Assumption 2 states how additional verification affects this spread.

Verification evaluates a contracted claim linking $\bar{q}$, τ_i , and the receipt. Deterministic checks, sampling audits, semantic evaluators, external oracles, and human arbitration instantiate this interface with different error rates and costs.

Definition 6 (Identity, reputation, and stake). Creating a fresh identity costs $\gamma \geq 0$. Each service agent carries reputation R_i , updated from attributable evidence τ_i , and posts stake $S_i \geq 0$ subject to slashing after verified misconduct. Identity governs admission, reputation affects future matching, and stake creates an immediate loss upon detection. The trust rule specifies whether reputation transfers across contexts and whether a sanctioned identity may re-enter.

Definition 7 (Mechanism). A mechanism is a tuple $M = (\mu, \pi, \phi) \in \mathcal{M}$, where matching rule μ selects service agents from task context, offers, and reputation; payment rule π specifies verification conditions and transfers; and trust rule ϕ governs admission, reputation, staking, and slashing. The mechanism acts on observable histories and admissible evidence, not directly on private type or effort.

4.3. The Trust-Asymmetric Coordination Problem

Mechanism M induces a repeated Bayesian game. A profile is *compliant* when selected service agents deliver the promised quality and participants follow the identity, payment, and authority rules. The mechanism must make compliant delivery sequentially rational while satisfying the four constraints in Figure 3.

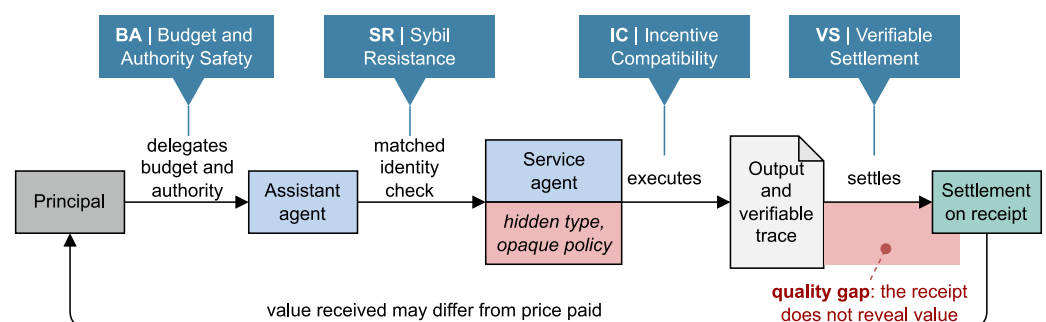


Figure 3. Trust-asymmetric coordination. A principal delegates bounded resources to an assistant agent that contracts with a service agent whose type and effort remain partially hidden. The four constraints protect delivery incentives, settlement, delegated authority, and identity.

- *Delivery Incentive Compatibility (IC)*: Given compliant play by others, delivering the promised quality maximizes each selected service agent's expected discounted payoff at every information set. This is an *ex post* delivery-discipline condition under a given detection technology, not incentive compatibility in the mechanism-design sense; it does not imply truthful revelation of types, costs, prices, or capabilities. Verifiable traces expand the contractible signal available to deter hidden-action deviations [8,49].
- *Verifiable Settlement (VS)*: For residual interval $[v_{\min}, v_{\max}]$ and receipt-conditioned price P , exposure is $\max\{P/v_{\min}, v_{\max}/P\}$. Writing $q(M)$ for worst-case exposure over residual fibers, the payment rule releases funds only against admissible evidence and must satisfy $q(M) \leq \beta(c_v)$ for a verification-dependent bound β . Verification therefore trades resource cost against unresolved exposure between buyer and seller [50,51].
- *Budget and Authority Safety (BA)*: On every realizable history, spending and actions routed through the mechanism remain within the delegated envelope B_a of assistant agent $a \in \mathcal{A}$. The constraint covers the enforced execution path; it does not govern external actions that bypass the constrained wallet or authority gate.
- *Sybil Resistance (SR)*: Net of identity cost γ , stake exposure, and reputation rule ϕ , no agent increases expected utility by operating an additional identity. A fresh identity therefore provides no profitable reputation reset, duplicate access, or evasion of prior sanctions.

Four deviations remain outside this constraint set. First, the matching rule does not make submitted prices or capability claims truthful; misrepresentation is disciplined only after execution through verification, slashing, and reputation. Second, *IC* does not govern assistant-side conduct. An assistant that overpays within a loose envelope remains bounded by *BA* and the exposure bound in Section 5.3, but its selection decision is not made incentive compatible. The same boundary excludes conflict between the assistant and its own principal: a misformulated instruction or a prompt-injection attack can redirect an assistant that is executing the mechanism faithfully, and neither is represented in the game [5,9]. Budget and authority safety still caps what such an assistant can spend through the enforcement path but not whether that spending serves the principal. Third, collusion among service agents is not modeled. Coordinated pricing is a documented risk [37–39] and remains an open problem in Section 7.2. Coordinated reputation inflation, in which a group of agents transacts internally to raise one another's scores, is excluded on the same ground, and covert coordination can be difficult to detect from settlement evidence alone [40]. Fourth, conduct that satisfies the verified claim while defeating principal intent remains residual because verification evaluates specified evidence, not intent.

Expected discounted surplus under compliant play is

$$W(M) = \mathbb{E} \left[\sum_{t=1}^{\infty} \delta^{t-1} (v(o_t) - C_{i_t}(o_t) - c_v z_t) \right], \quad (5)$$

where i_t denotes the service agent selected in period t .

Payments, refunds, and redistributed stake cancel from this expression; actual resource destruction or administrative cost must be subtracted separately.

Problem 1 (Trust-asymmetric coordination problem). *The trust-asymmetric coordination problem is*

$$\begin{aligned} & \underset{M \in \mathcal{M}}{\text{maximize}} && W(M) \\ & \text{subject to} && \text{IC, VS, BA, and SR.} \end{aligned} \quad (6)$$

A mechanism is feasible when all four constraints hold. A feasible mechanism is exposure-optimal when its receipt-conditioned payment rule minimizes residual exposure over all receipt-conditioned payment rules. Section 5 constructs M^* , which is feasible when the stated deterrence, participation, identity-cost, and enforcement conditions hold, and establishes its exposure optimality within that rule class. It does not claim that M^* solves the unrestricted welfare-maximization problem.

5. Coordination Mechanism

This section constructs $M^* = (\mu^*, \pi^*, \phi^*)$ from credential-gated admission, reputation-weighted matching, constrained escrow, and verification-conditioned settlement. The exposure result holds within the class of receipt-conditioned payment rules; the remaining guarantees depend on the stated incentive, identity-cost, and enforcement conditions.

5.1. The Mechanism

Trust rule ϕ^* : Admission requires a credential binding ρ_i to a principal and refundable stake $S_i \geq \underline{S}$. Reputation $R_i \in [0, 1]$ starts at zero. Verified compliance gives $R_i^+ = \min\{1, R_i + \eta\}$; a detected shortfall gives $R_i^+ = \max\{0, R_i - \eta\}$ and destroys fraction σ of S_i . Updates require attributable evidence, and unslashed stake is refundable at orderly exit.

Matching rule μ^* : Each service agent submits an offer containing a settlement ceiling p_i and promised quality. Admissibility requires valid credentials, sufficient stake, task compatibility, and p_i within the principal's delegated budget. The mechanism selects

$$i^* \in \operatorname{argmax}_{i \in \mathcal{E}(c)} \{R_i - \lambda p_i\}, \quad (7)$$

where $\mathcal{E}(c)$ is the admissible set and $\lambda > 0$ controls the tradeoff between reputation and price. This is a scoring rule, not a truthful mechanism: agents may shade prices or overstate capability. The ceiling p_{i^*} bounds pre-execution exposure, while a capability claim rejected by the verifier triggers refund, slashing, and reputation loss. Offer-stage truthfulness is not claimed.

Payment rule π^* : The wallet reserves $p_{i^*} \leq B_a$ and validates the action against delegated authority. Verification at budget c_v produces $[v_{\min}, v_{\max}]$, and settlement is admissible only if $P^* = \sqrt{v_{\min} v_{\max}} \leq p_{i^*}$. If the verifier accepts the evidence, P^* is released to the service agent and the escrow remainder is returned to the principal; rejection returns the escrow to the principal and activates sanctions. The ceiling bounds exposure before execution, while P^* prices residual uncertainty at settlement.

Assumption 2 (Verification technology). At budget c_v , the verifier partitions each receipt fiber into residual fibers whose value spread is bounded by $s(c_v)$, where $s : \mathbb{R}_{\geq 0} \rightarrow [1, s_0]$ is nonincreasing, $s(0) = s_0$, and $s(c_v) \rightarrow 1$ as verification strength increases. Cost c_v is incurred per verified settlement. The scalar spread summarizes residual uncertainty rather than a particular implementation. With false-negative rate $\epsilon(c_v)$, the effective spread is

$$s_\epsilon(c_v) = s(c_v) + \epsilon(c_v)(s_0 - s(c_v)), \quad (8)$$

which interpolates between the verified and unresolved bounds. Residual spread, shortfall detection $d(c_v)$, and false-positive probability $\phi(c_v)$ remain distinct verification properties. When false negatives affect residual ambiguity, the exposure results below use $s_\epsilon(c_v)$ in place of $s(c_v)$.

Assumption 2 treats verification as an interface rather than presupposing a particular technology. Table 3 instantiates the three verification quantities for four verifier classes and identifies when the interface is empty, meaning that no admissible claim can be evaluated and the mechanism has nothing to price. Section 6.5 measures these quantities for a sampling-audit verifier.

Table 3. Verifier classes, their verification primitives, and the conditions under which no admissible claim can be evaluated.

Verifier Class	Residual Spread $s(c_v)$	Detection $d(c_v)$, False Positives $\varphi(c_v)$	Interface Empty When
Deterministic check (hash, attestation, schema)	Equals 1 for the verified predicate and remains unconstrained for untested properties	Approaches $d = 1$ and $\varphi = 0$ only when the check and its attestation are trustworthy	Value depends on properties outside the check
Sampling audit (test suite, human spot check)	Value spread induced by the confidence bounds on the audited pass rate; interval width decreases at order $k^{-1/2}$	Determined by the sampling and acceptance rules; controlling φ requires calibration on compliant cases	No sampleable case distribution exists, or outcomes do not identify compliant conduct
Semantic evaluator (model-based judge)	Value interval induced by calibrated uncertainty in the judge's assessment	Estimated against labeled contracted claims; φ may arise from systematic judge error	The quality predicate is not operational, or the judge shares the agent's relevant failure modes
Human arbitration	Value interval implied by admissible evidence and adjudicator uncertainty, limited by contract specificity	Must be estimated from adjudication performance; cost and delay enter the verification budget	The obligation is not adjudicable, or delay exceeds the settlement window

Algorithm 1 specifies the ordering of authority validation, execution, verification, and transfer; Figure 4 locates those operations in the infrastructure. Blockchain provides one auditable implementation: ERC-8004-style registries can instantiate ϕ^* [16,25], escrow contracts can implement stake and slashing, and account-abstraction policies can enforce π^* [7]. The results also apply to non-blockchain or hybrid substrates that reproduce the same admission, evidence, authority, settlement, and sanction semantics and satisfy Assumption 2. Zero-knowledge and zkML proofs are possible evidence technologies but are not implemented in the evaluation.

Algorithm 1 Per-period execution of M^*

Require: admitted agents with identities ρ_i , stakes S_i , and reputations R_i ; task context c ; delegated envelope B_a of assistant agent a ; verification budget c_v

- 1: each admitted service agent i submits an offer $(p_i, \bar{q})$
- 2: form $\mathcal{E}(c)$ using credential, stake, task, authority, and settlement-ceiling checks
- 3: compute $sc_i \leftarrow R_i - \lambda p_i$; select $i^* \in \operatorname{argmax}_{i \in \mathcal{E}(c)} sc_i$ $\triangleright \mu^*$
- 4: reserve p_{i^*} and validate the action against B_a $\triangleright \pi^*; BA$
- 5: agent i^* executes the task, emitting output o and attributable trace τ_{i^*}
- 6: invoke the verifier at budget c_v and obtain $[v_{\min}, v_{\max}]$ $\triangleright \pi^*; VS$
- 7: **if** the verifier accepts the evidence **then**
- 8: release $P^* = \sqrt{v_{\min} v_{\max}}$ to i^* ; return the remainder to the principal; increase R_{i^*}
- 9: **else**
- 10: return the escrow to the principal; slash σS_{i^*} ; decrease R_{i^*} $\triangleright \phi^*$
- 11: **end if**

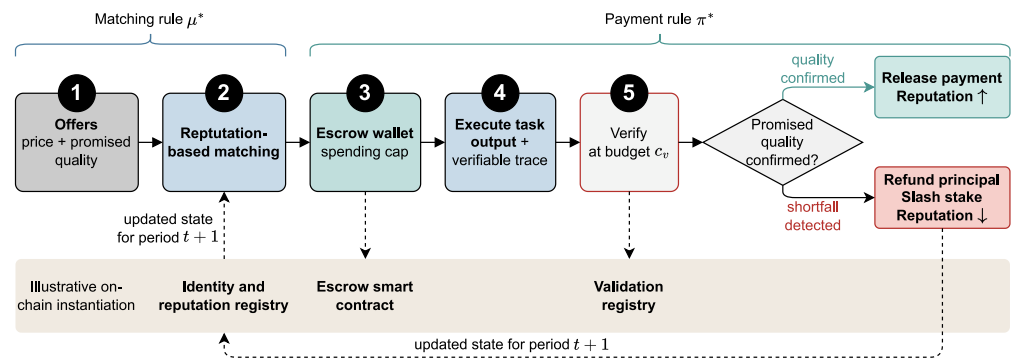


Figure 4. Per-period mechanism flow. Arrows indicate the order of operations in Algorithm 1. Verifiable registries, constrained escrow, and validation services instantiate the trust and settlement steps without making blockchain a definitional requirement.

5.2. Criterion-to-Mechanism Mapping

Table 4 records where the five economic-actor criteria enter the mechanism; the formal guarantees are established separately below.

Table 4. Economic-actor criteria and their corresponding mechanism components.

Criterion (Definition 2)	Mechanism Component
(C1) Identity	Credential-gated admission
(C2) Authority	Constrained wallet and authority envelope
(C3) Payment	Escrow and receipt-conditioned settlement
(C4) Verifiable behavior	Verification at the selected budget, with geometric-mean release
(C5) Reputation	Stake, reputation update, and score matching

5.3. Residual Verification Exposure

For residual fiber r' , let its value interval be $[v_{\min}(r'), v_{\max}(r')]$ and spread $s_{r'} = v_{\max}(r')/v_{\min}(r')$. A receipt-conditioned payment rule assigns price $P(r') > 0$. Buyer and seller exposures are

$$q_b(P; r') = \frac{P(r')}{v_{\min}(r')}, \quad q_s(P; r') = \frac{v_{\max}(r')}{P(r')}, \quad (9)$$

and the worst-case residual exposure of a mechanism M with price rule P is $q(M) = \sup_{r'} \max\{q_b(P; r'), q_s(P; r')\}$. This multiplicative criterion treats buyer overpayment and seller underpayment symmetrically and is invariant to the monetary unit.

Theorem 1 (Residual exposure bound). *For every residual fiber r' and receipt-conditioned price $P(r')$,*

$$\max\{q_b(P; r'), q_s(P; r')\} \geq \sqrt{s_{r'}}. \quad (10)$$

Equality holds exactly at $P^*(r') = \sqrt{v_{\min}(r')v_{\max}(r')}$. Hence π^* minimizes exposure pointwise over all receipt-conditioned payment rules and satisfies

$$q(M^*) = \sup_{r'} \sqrt{s_{r'}} \leq \sqrt{s(c_v)}. \quad (11)$$

If false negatives enlarge the effective spread to $s_e(c_v)$, the corresponding bound is $q(M^*) \leq \sqrt{s_e(c_v)}$. Equality holds when some residual fiber attains the applicable spread bound.

Proof. For each r' ,

$$q_b(P; r') q_s(P; r') = \frac{P(r')}{v_{\min}(r')} \frac{v_{\max}(r')}{P(r')} = s_{r'}. \quad (12)$$

The maximum of two positive numbers with fixed product $s_{r'}$ is minimized when they are equal. Thus $\max\{q_b, q_s\} \geq \sqrt{s_{r'}}$, with equality when $P(r')/v_{\min}(r') = v_{\max}(r')/P(r')$, which gives $P^*(r') = \sqrt{v_{\min}(r')v_{\max}(r')}$. Taking the supremum preserves this pointwise optimality over the residual fibers. $\square$

Remark 1. The factor $\sqrt{s}$ also appears in the hidden-cost impossibility result of Ling et al. [6], while comprehension gating bounds exposure through robustness audits [24]. Here the fiber contains unresolved output values; the contribution is the verification-coupled price, not the arithmetic and geometric mean identity.

Corollary 1 (Verifiable settlement). Setting $\beta(c_v) = \sqrt{s(c_v)}$, or $\beta(c_v) = \sqrt{s_\epsilon(c_v)}$ when false negatives enlarge residual ambiguity, gives $q(M^*) \leq \beta(c_v)$, so M^* satisfies VS. In the latter case, the bound is nonincreasing and converges to 1 when $\epsilon(c_v)$ is nonincreasing and tends to zero. Because verification cost enters welfare, budget selection remains a separate tradeoff between exposure and resource consumption.

5.4. Delivery Incentive Compatibility, Safety, and Sybil Resistance

The remaining guarantees depend on the continuation value of reputation, execution-time authority checks, and the cost of identity renewal.

Proposition 1 (Delivery incentive compatibility). Let g_i denote agent i 's one-period gain from delivering a shortfall, $d(c_v) \in (0, 1]$ the probability that the verifier detects it, and $\Delta_i \geq 0$ the per-period reduction in expected future payoff after the resulting reputation loss. Compliant delivery is sequentially optimal whenever

$$d(c_v) \left(P^* + \sigma S_i + \frac{\delta}{1 - \delta} \Delta_i \right) \geq g_i. \quad (13)$$

A detected shortfall returns the escrow to the principal, so the deviating agent forfeits P^* in addition to the slashed stake and continuation loss. If the condition holds for every selected type and beliefs are consistent, compliant delivery satisfies IC. The proposition governs delivery by selected service agents; it does not cover offer-stage claims, assistant-side conduct, or coordinated deviations. With false-negative rate $\epsilon(c_v)$, $d(c_v) = 1 - \epsilon(c_v)$. The condition shows how detection, stake, patience, and reputation-dependent continuation value jointly deter shortfalls.

Proof. A shortfall produces immediate gain g_i . Detection causes forfeiture of P^* under the escrow refund in Algorithm 1, slashing loss σS_i , and discounted continuation loss $\delta \Delta_i / (1 - \delta)$. Expected deviation gain is therefore

$$g_i - d(c_v) \left(P^* + \sigma S_i + \frac{\delta}{1 - \delta} \Delta_i \right), \quad (14)$$

which is nonpositive under the stated condition. The one-shot deviation principle establishes sequential optimality; strict inequality makes compliant delivery the unique modeled best response. $\square$

Corollary 2 (Deterrence without continuation value). *If the continuation term is unavailable, so $\Delta_i = 0$, compliant delivery remains sequentially optimal whenever*

$$d(c_v)(P^* + \sigma S_i) \geq g_i. \quad (15)$$

For a worst-case gain $g_{\max}$, deterrence is restored by

$$\underline{S} \geq \max\left\{0, \frac{g_{\max}}{\sigma d(c_v)} - \frac{P^*}{\sigma}\right\}. \quad (16)$$

This result covers thin markets, where limited substitution makes Δ_i approach zero, and terminal periods, where an exiting agent has no continuation value. Deterrence then rests on the forfeited price and slashed stake; Section 6.6 measures the resulting tradeoff.

Proposition 2 (Budget and authority safety). *Under M^* , every mechanism-mediated payment and action lies within the delegated envelope B_a on every realizable history; hence M^* satisfies BA. Coverage follows the enforcement path, not the settlement venue: an action is covered whenever it transits the constrained wallet and authority gate, whether settlement is on-chain or off-chain. External effects that bypass this gate are outside the guarantee, and restricting an agent to the covered path remains a deployment requirement.*

Proof. Rule π^* reserves and releases funds only after the constrained wallet verifies the payment ceiling, while the authority gate rejects actions outside the principal's encoded mandate. Because these checks precede execution, no mechanism-mediated path exceeds B_a . $\square$

Proposition 3 (Sybil resistance). *Let V_0 denote the incremental gross value of an additional fresh identity and q_{slash} its equilibrium probability of verified misconduct before reputation accumulation. Creating that identity is nonprofitable whenever*

$$\gamma + \sigma \underline{S} q_{\text{slash}} \geq V_0. \quad (17)$$

Under this condition, M^* satisfies SR. The inequality concerns the incremental value of identity multiplication, including duplicate access and whitewashing, not the value of an agent's original identity.

Proof. A fresh identity begins at $R_i = 0$ and has incremental net value

$$V_0 - \gamma - \sigma \underline{S} q_{\text{slash}}. \quad (18)$$

The stated condition makes this value nonpositive. If $q_{\text{slash}} = 0$, identity cost γ alone must dominate V_0 ; refundable stake deters only identities exposed to detectable misconduct. $\square$

Although V_0 is not directly observable, registry data provide the conservative bound $V_0 \leq \bar{u} \bar{m} / (1 - \delta)$, where $\bar{u}$ is mean settled profit per match and $\bar{m}$ is mean matches per period. Admission parameters can be calibrated against this upper bound.

A delayed Sybil behaves compliantly while building reputation and defects on exit. The build phase delivers the contracted service and increases Δ_i , but the continuation term vanishes at exit. Corollary 2 then applies, and the continuously posted stake must satisfy Equation (15); the binding deterrent shifts from reputation to stake.

The framework does not assume one credential per actor. Each additional legitimate credential incurs γ and locks at least $\underline{S}$, so multiplication carries both an identity cost and a capital requirement. Issuers must preserve a positive γ through a scarce identity root, hardware attestation, or rate-limited issuance. If credentials are unlimited and $\gamma \rightarrow 0$, SR rests only on expected slashing exposure; this is an issuer-side deployment condition.

Proposition 4 (False positives, deterrence, and participation). *Let $\varphi(c_v) \in [0, 1)$ denote the probability that a compliant delivery is wrongly rejected, with the same escrow refund, slashing, and reputation loss as a correctly detected shortfall. Define $D_i = P^* + \sigma S_i + \frac{\delta}{1-\delta} \Delta_i$ as the total loss from rejection. Compliant delivery is sequentially optimal whenever*

$$(d(c_v) - \varphi(c_v)) D_i \geq g_i, \quad (19)$$

and, relative to an outside option normalized to zero, a compliant agent participates only while

$$\varphi(c_v) \leq \varphi_{\max} = \frac{P^* - C_i(\bar{q})}{D_i}, \quad (20)$$

where $C_i(\bar{q})$ is the cost of promised-quality delivery.

Proof. A compliant delivery is rejected with probability φ , while a shortfall is detected with probability d ; either rejection imposes D_i . Expected payoffs are $P^* - C_i(\bar{q}) - \varphi D_i$ for compliance and $P^* - C_i(\bar{q}) + g_i - d D_i$ for a shortfall. Their difference is $g_i - (d - \varphi) D_i$, which gives Equation (19). Requiring compliant payoff to be non-negative gives Equation (20). $\square$

Deterrence depends on $d - \varphi$, whereas participation depends separately on φ ; Proposition 1 is the case $\varphi = 0$. Raising stake enlarges D_i , which strengthens deterrence but lowers $\varphi_{\max}$. Section 6.5 quantifies this tradeoff.

Corollary 3 (Feasibility and exposure optimality). *Under Assumptions 1 and 2, suppose that Corollary 1 and Propositions 2 and 3 apply. If Proposition 1 holds when $\varphi(c_v) = 0$, or the deterrence and participation conditions of Proposition 4 hold when $\varphi(c_v) > 0$, then M^* satisfies IC, VS, BA, and SR. It is therefore feasible for Problem 1 and exposure-optimal within the class of receipt-conditioned payment rules. The result does not establish global welfare optimality, and feasibility remains conditional on the stated deterrence, participation, identity-cost, and enforcement conditions.*

6. Numerical Analysis

The numerical study assesses the behavior of M^* and the tradeoff between welfare and verification under the repeated-game model. It compares M^* with Receipt-Only and Verification-Only variants, evaluates Theorem 1 as an exact pricing-identity check, and tests whether the analytical incentive predictions persist under learning agents. The evaluation combines model simulations, learning-agent experiments, alternative verification laws, and an implementation-based sampling-audit verifier. These analyses assess the mechanism's internal consistency and robustness under the stated experimental conditions; they do not constitute field validation of a deployed agent economy. The public reproducibility package is identified in the Data Availability Statement.

6.1. Experimental Setup

The analytical best-response simulations use $n = 50$ agents, $T = 4000$ periods, and 100 seeds; the learning experiments use $T = 8000$ and 50 seeds, and the convergence analysis uses 200 seeds. Each private shirking gain g_i follows a heavy-tailed truncated normal distribution, so deterrence changes the honest fraction gradually rather than at a common threshold. Honest delivery has value 1.0 and cost 0.4, whereas a shortfall has value 0.25 and cost 0.1, giving $s_0 = 4$. The baseline specifies $s(c_v) = 1 + (s_0 - 1)e^{-c_v}$, verification cost $\alpha_{cv}c_v$ with $\alpha_{cv} = 0.05$ for budget c_v , and geometric-mean settlement. Outcomes include welfare, honest delivery, disputes, exposure, and theory-linked diagnostics. All stochastic results use fixed seeds; confidence bands represent variation across seeds, and convergence results report medians and interquartile ranges.

The sweeps vary verifier quality, stake, patience, identity renewal, and learning. Six variants isolate the mechanism components: M^* ; open and centralized Verification-Only; Receipt-Only; verification with reputation but no staking; and verification with staking but no reputation. The two Verification-Only variants differ only in routing efficiency and produce identical conduct, so their conduct curves are merged after equality checks.

6.2. Welfare and the Verification Tradeoff

Figure 5a shows that M^* reaches an interior welfare maximum of 0.56 at $c_v = 0.75$, with honest delivery of 0.99; beyond this point, verification cost dominates. This optimum illustrates Corollary 1 under the baseline calibration rather than identifying a universal budget. At $c_v = 0.5$, honesty under M^* rises from 0.78 to 1.00 as δ increases from 0.80 to 0.99, while the variant without reputation remains at 0.74 (Figure 5b). Receipt-Only reaches welfare of 0.15. At $c_v = 1.50$, M^* attains welfare of 0.53 and honesty of 1.00, compared with 0.37 and 0.66 under open Verification-Only. Open Verification-Only attains its own maximum later and lower, at welfare 0.38 and honesty 0.77 at $c_v = 2.25$. The differences isolate the contributions of stake and reputation from the routing factor and show that patience affects conduct only through continuation value.

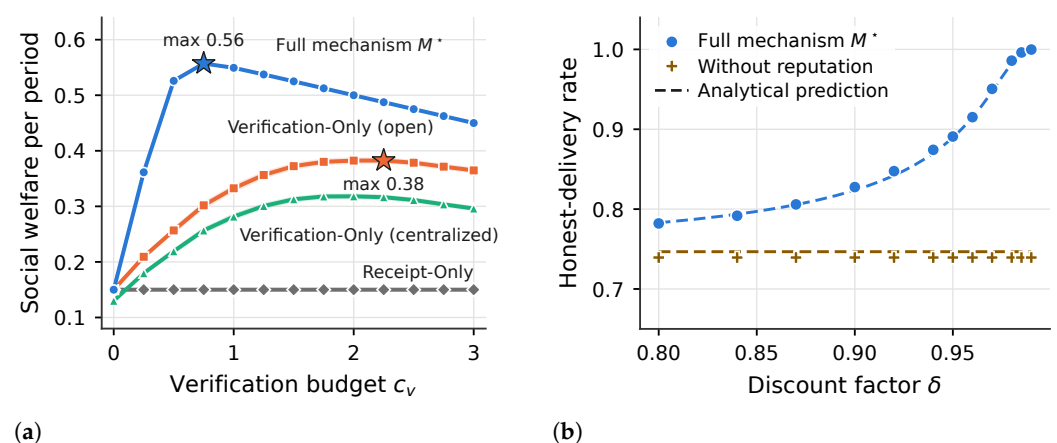


Figure 5. Welfare and continuation value. (a) M^* reaches its maximum welfare at $c_v = 0.75$. (b) At $c_v = 0.5$, honest delivery under M^* increases with the discount factor and tracks the analytical benchmark, whereas the variant without reputation remains nearly constant.

6.3. Residual Exposure

Figure 6 evaluates Theorem 1 exactly rather than through Monte Carlo estimation. Geometric-mean pricing attains exposure $\sqrt{s}$, whereas arithmetic-mean pricing gives $(1 + s)/2$: the respective values are 2.00 and 2.50 at $s = 4$, and 2.83 and 4.50 at $s = 8$ (Figure 6a). At fixed $c_v = 1$, false negatives widen the effective residual fiber. Geometric-mean pricing continues to attain $\sqrt{s_e(c_v)}$, while the excess under arithmetic-mean pricing

increases with the unresolved spread (Figure 6b). Disputes initially increase as partial verification detects shortfalls and disappear once deterrence makes compliant delivery optimal. Verification-Only produces more intermediate-budget disputes because it lacks reputation-based continuation consequences.

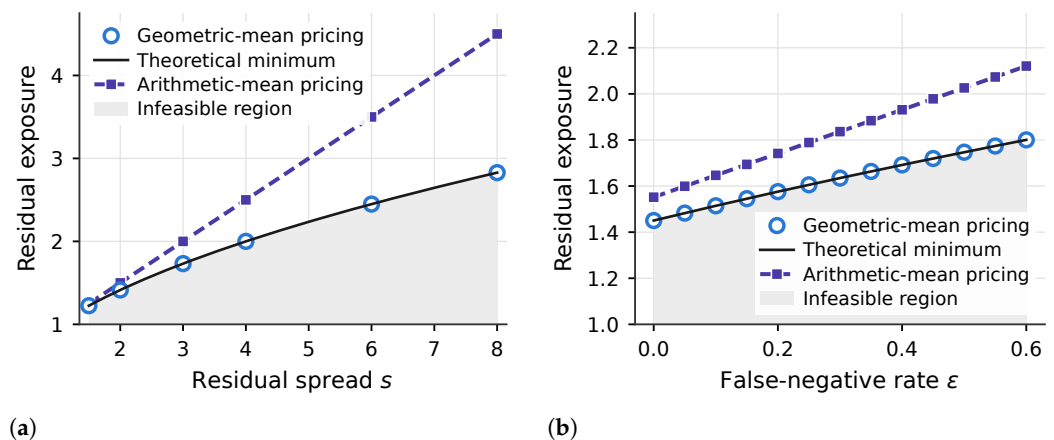


Figure 6. Residual exposure. (a) Geometric-mean pricing attains the analytical bound, whereas arithmetic-mean pricing increasingly exceeds it. (b) At $c_v = 1$, false negatives widen the effective fiber, but geometric-mean pricing attains the degraded bound $\sqrt{s_\epsilon(c_v)}$.

6.4. Delivery Incentive Compatibility and Sybil Resistance

At $c_v = 0.5$, honesty rises from 0.50 without stake to 0.80 at stake 0.40, and the analytical curve crosses 0.95 near stake 1.0 (Figure 7a). Its agreement with the simulation checks the implementation of Proposition 1. The variant without reputation requires more stake, isolating continuation value, while Receipt-Only remains at zero honesty because it cannot penalize shortfalls. As whitewashing rises to 0.8, M^* maintains welfare of 0.55, whereas welfare without staking falls from 0.48 to 0.36 and Verification-Only remains at 0.33 (Figure 7b). The stability of M^* remains conditional on the identity-cost inequality in Proposition 3; it does not imply that any positive stake deters every additional identity.

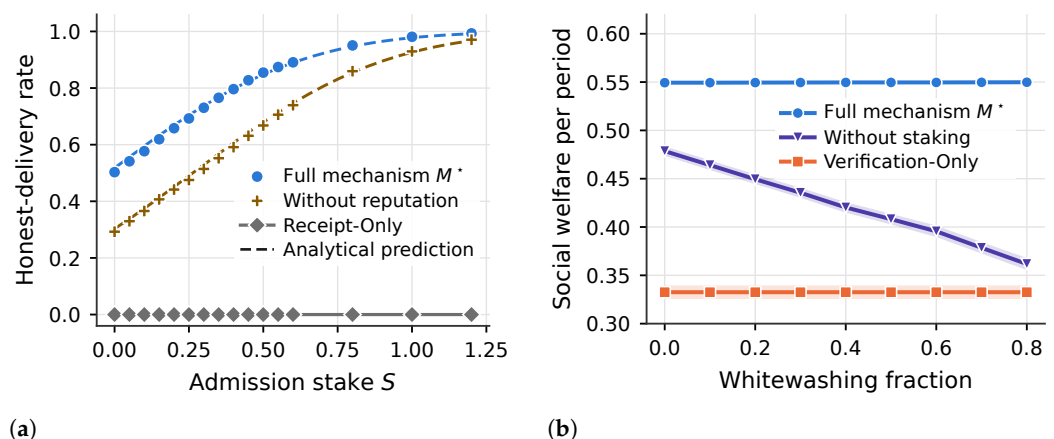


Figure 7. Delivery incentive compatibility and Sybil resistance. (a) Honest delivery rises with the admission stake and closely follows the analytical benchmark over 100 seeds. (b) Staking preserves welfare as the whitewashing fraction rises, whereas removing staking makes identity renewal profitable; Verification-Only provides the no-reputation reference level.

6.5. Verification Technology

The baseline functions satisfy $s(c_v) = 1 + (s_0 - 1)(1 - d(c_v))$, so residual spread tracks undetected mass. Alternative technologies replace d and derive s through this coupling. All five detection laws produce an interior optimum (Figure 8 and Table 5); its location

shifts from 0.75 to 1.50 under slow linear detection and coincides with the 0.75 switch point under threshold detection. Thus, the interior optimum reflects the tradeoff between cost and deterrence rather than the exponential specification, although its location remains technology dependent.

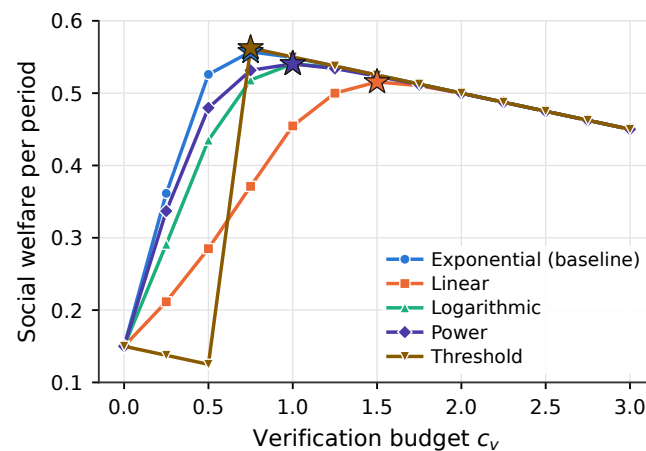


Figure 8. Welfare under five detection laws. Stars mark the interior welfare maxima.

Table 5. Peak welfare and honesty under alternative detection laws and the measured sampling-audit verifier.

Verification Technology	Welfare Peak	At c_v	Honesty at Peak
Exponential baseline	0.557	0.75	0.99
Linear	0.515	1.50	0.98
Logarithmic	0.540	1.00	0.98
Power	0.540	1.00	0.98
Threshold at $c_v = 0.75$	0.562	0.75	1.00
Measured sampling audit	0.563	0.50	0.97

The final row instantiates a sampling-audit verifier. Compliant and graded-shortfall implementations of a ranking specification are executed against an independent test bank, with budget purchasing k tests without replacement. A Wilson interval on the audited pass rate defines residual spread, and calibration on compliant implementations targets a false-positive rate of 0.05. The verifier neither accesses the model's $v_{\min}/v_{\max}$ oracle nor fits a curve to the results.

The measured spread falls from 4.00 at zero budget to 1.37 at $c_v = 3$, compared with the assumed 1.15 (Figure 9a). Detection exceeds the assumed value at $c_v = 0.5$ (0.51 versus 0.39) but falls below it at $c_v = 3$ (0.77 versus 0.95), because sampling can miss rare violations. Substituting these measurements preserves the interior optimum but shifts it from 0.75 to 0.50. The existence of the optimum is therefore robust in this experiment, whereas its location is verifier dependent.

At $c_v = 1$ and stake 0.60, welfare and honesty remain stable below the participation bound $\varphi_{\max} = 0.072$ but decline sharply when compliant agents withdraw above it (Figure 9b). Increasing stake from zero to 1.00 lowers false-positive tolerance from 0.127 to 0.056; stake therefore strengthens deterrence while potentially weakening participation. The sharp transition reflects the common outside option assumed in the model. Heterogeneous outside options would instead produce gradual exit beginning at the bound.

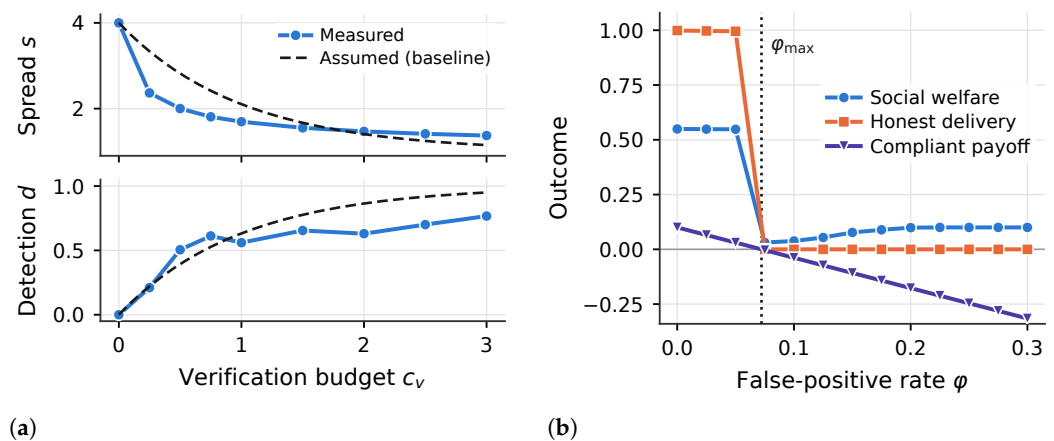


Figure 9. Measured verification technology. (a) Sampling-audit spread and detection compared with Assumption 2. (b) Welfare, honest delivery, and compliant payoff against ϕ at $c_v = 1$ and stake 0.60; the dotted line marks $\phi_{\max} = 0.072$.

Because the allowed-failure threshold in Table 6 is integer valued, the acceptance rule changes discretely. Detection can therefore decrease when the allowance increases, as between $k = 6$ and $k = 8$; this decrease follows from calibration to a fixed false-positive target rather than from Monte Carlo noise.

Table 6. Measured sampling-audit primitives. The budget purchases k tests; s is the mean residual spread derived from the Wilson interval; and d and ϕ are the realized detection and false-positive rates under an acceptance rule targeting $\phi = 0.05$.

c_v	k Tests	Allowed Failures	s Measured	d Measured	ϕ Measured
0.00	0	—	4.000	0.000	0.000
0.25	2	1	2.366	0.211	0.003
0.50	4	1	2.002	0.505	0.015
0.75	6	1	1.809	0.613	0.028
1.00	8	2	1.695	0.560	0.007
1.50	12	2	1.552	0.655	0.024
2.00	16	3	1.469	0.630	0.013
2.50	20	3	1.413	0.701	0.028
3.00	24	3	1.372	0.767	0.045

6.6. Market Thickness and Thin Markets

Thickness κ is the fraction of future profit lost after a reputation downgrade; $\kappa = 0$ represents monopoly. At $c_v = 0.5$ and stake 0.60, honesty decreases from 0.89 at $\kappa = 0.15$ to 0.74 under monopoly, while stake 1.00 raises monopoly honesty to 0.93 (Figure 10a). The curves track Corollary 2: forfeited payment and stake substitute for continuation value as the market thins.

Maintaining honesty of 0.95 requires the stake to rise from 0.53 in the thickest evaluated market to 1.10 under monopoly (Figure 10b). Thin markets therefore roughly double the evaluated cost of deterrence rather than making deterrence infeasible. A market opening with few participants must rely more heavily on stake-backed deterrence, with the stake floor set by Corollary 2, until the market thickens. Which other instruments can substitute for continuation value during that phase remains an open question. Because analytical agents respond to mechanism deterrence rather than to one another, n and T govern sampling, not competition; κ is the model's incentive-relevant market-structure parameter.

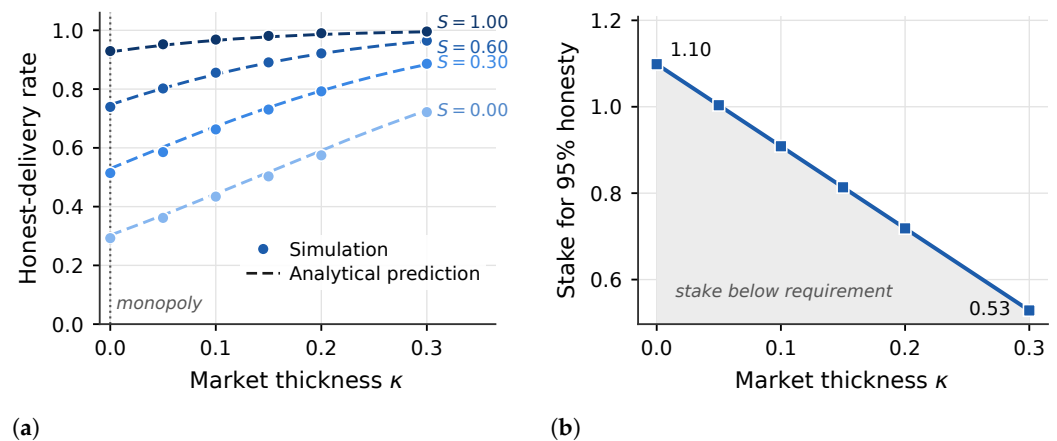


Figure 10. Market thickness and deterrence at $c_v = 0.5$. (a) Honest delivery across κ and four admission stakes; dashed curves show Corollary 2. (b) Stake required to maintain honesty of 0.95.

6.7. Robustness to Learning Agents

Epsilon-greedy Q-learning tests adaptive behavior without revealing Proposition 1 to the agents. At $c_v = 1$, honesty under M^* rises from approximately 0.85 at period 250 to 0.99 at period 8000; Verification-Only rises from 0.76 to 0.99, whereas Receipt-Only falls from 0.43 to 0.04 (Figure 11a). The identical Verification-Only trajectories are merged after an equality check.

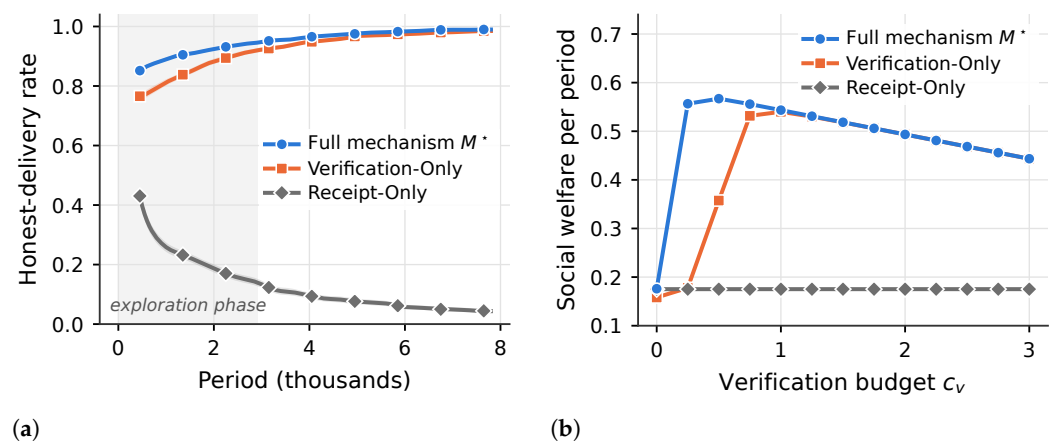


Figure 11. Learning-agent robustness. (a) Mean 250-period honesty trajectories with 95% confidence bands at $c_v = 1$. (b) Welfare preserves the mechanism ordering and an interior optimum.

Convergence is the first period after which 250-period moving honesty remains within ± 0.05 of final-quartile honesty for 500 periods. At $c_v = 1$, all 200 runs of M^* converge honestly after a median of 2777 periods (IQR 2498–3093), compared with 3469 (IQR 3023–3990) for Verification-Only. Receipt-Only converges dishonestly in 183 runs after a median of 3969 periods (IQR 3309–4640). Figure 11b preserves the welfare ordering: M^* peaks at 0.57 and $c_v = 0.50$, while Receipt-Only remains near 0.18. This lower optimum reflects homogeneous learners whose deterrence saturates earlier than in the heavy-tailed analytical population; the experiment tests adaptive behavior rather than reproducing the baseline calibration.

Three diagnostics connect the results to the model's economic foundations: the *moral-hazard residual*, $MHR = \Pr[\text{shirk} \wedge \neg \text{detected}]$; *residual coordination overhead*, $RCO = (\alpha_{cv}c_v + \Pr[\text{dispute}] \cdot 0.05)/P^*$; and illustrative *dispute-resolution time*, $DRT = \lceil \ln((s_0 - 1)/0.1)/(K_{\text{det}} \max(c_v, 0.1)) \rceil$ with $K_{\text{det}} = 1$. Under M^* , MHR falls from

1.00 at $c_v = 0$ to 0.00 at $c_v = 3$ (Figure 12a). At $c_v = 1.50$, its overhead is 0.15 and arises entirely from verification; Verification-Only retains a dispute overhead of approximately 0.028 at $c_v = 0.50$ (Figure 12b). Receipt-Only has zero measured overhead despite poor conduct and is omitted from the bars. By construction, DRT is identical across verifying mechanisms and falls from 14 rounds at $c_v = 0.25$ to 2 at $c_v = 3$; it is undefined for Receipt-Only.

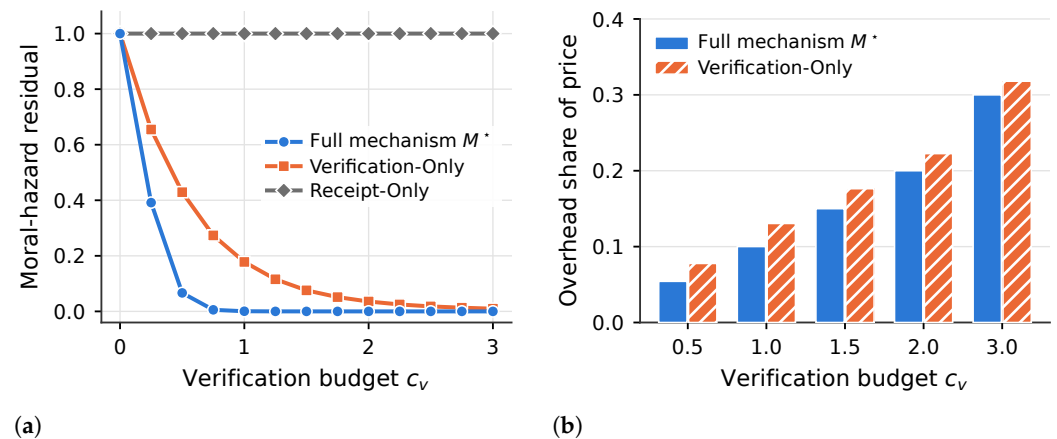


Figure 12. Theory-linked diagnostics. (a) Moral-hazard residual; Verification-Only curves are merged after equality checks. (b) Verification and dispute components of coordination overhead.

6.8. Component and Matching Ablations

Table 7 compares component removals at $c_v = 1$. Without verification, honesty is zero and welfare is 0.150. At this calibration, staking contributes 0.070 to welfare and 15.8 honesty points, whereas reputation contributes 0.008 and 1.9 points at stake 0.60. This smaller marginal effect does not make reputation dispensable: forfeited payment and stake already provide most deterrence, while the matching ablation isolates the continuation value created by reputation. The differences are calibration specific rather than a general ranking of components.

Table 7. Component ablation at $c_v = 1$, averaged over 100 seeds. Each ablated row disables one enforcement component of M^* .

Variant	Verify	Stake	Rep.	Welfare	Honesty
Full mechanism M^*	yes	yes	yes	0.549	0.999
Verification + reputation, no staking	yes	no	yes	0.479	0.841
Verification + staking, no reputation	yes	yes	no	0.541	0.980
Verification-Only, open matching	yes	no	no	0.333	0.517
Verification-Only, centralized routing	yes	no	no	0.282	0.517
Receipt-Only	no	—	—	0.150	0.000

Because analytical runs select agents uniformly, a separate experiment measures continuation value under each matching rule (Table 8). A reputation slash reduces the focal agent's selection share from 0.024 to 0.003 under score matching and to zero under reputation-only matching. Price-only and random matching leave selection unchanged, giving $\Delta_i = 0$. Reputation-sensitive rules add 5.0 honesty points without stake but only 0.6 at stake 0.60, where stake substitutes for continuation value. Price-only matching pays 0.40, compared with 0.50 under score matching; score matching therefore trades higher payment for better conduct rather than dominating both dimensions.

Table 8. Matching-rule ablation at $c_v = 1$. Selection shares are measured before and after one reputation slash; Δ_i is the resulting per-period continuation loss.

Matching Rule	Share, Intact	Share, Slashed	Δ_i	Honesty, $S = 0$	Honesty, $S = 0.6$
Score, $R_i - \lambda p_i$	0.0241	0.0029	0.0021	0.582	0.982
Reputation only	0.0252	0.0000	0.0025	0.592	0.983
Price only	0.0241	0.0241	0.0000	0.532	0.976
Random	0.0250	0.0250	0.0000	0.532	0.976

Per-period centralized welfare is $\omega v(o) - C(o) - \alpha_{cv}c_v$, with stipulated routing efficiency $\omega = 0.92$ and $\omega = 1$ otherwise. The discount represents centralized dispatch; M^* instead bears verification cost $\alpha_{cv}c_v$. The main comparison assigns $\omega = 1$ to both M^* and open Verification-Only. Removing the centralized discount raises its peak welfare from 0.318 to 0.382, still below the peak of 0.557 for M^* ; the ordering is unchanged across $\omega \in \{0.88, 0.92, 0.96, 1.00\}$.

6.9. Sensitivity Analysis

Sensitivity sweeps vary n , T , s_0 , slashing, stake, patience, verification cost, and dispute overhead. Across these sweeps, M^* retains the highest welfare, the verification optimum remains interior, geometric-mean pricing attains the exposure bound, and staking prevents whitewashing loss. Table 9 reports one-at-a-time results for M^* . Changing n or T affects only Monte Carlo precision because analytical agents respond independently to deterrence; this invariance is not evidence of competition robustness. Market structure instead enters through κ and the matching rule. Larger s_0 increases the optimal budget, while greater patience, higher stake, or smaller spread reduces the required verification. The verification-cost coefficient changes peak welfare but not its grid location.

Table 9. One-at-a-time sensitivity analysis for M^* : peak welfare and honesty.

Parameter	Values	Welfare Peak (Value at c_v)	Honesty at Peak
n	20, 50, 100	0.56 at 0.75; 0.56 at 0.75; 0.56 at 0.75	0.99; 0.99; 0.99
T	4000, 20,000	0.56 at 0.75; 0.56 at 0.75	0.99; 0.99
s_0	2, 4, 8	0.57 at 0.50; 0.56 at 0.75; 0.50 at 1.50	1.00; 0.99; 0.96
σ	0.5, 1.0	0.54 at 1.00; 0.56 at 0.75	0.98; 0.99
stake	0.3, 0.6, 1.0	0.54 at 1.00; 0.56 at 0.75; 0.57 at 0.50	0.98; 0.99; 0.98
δ	0.90, 0.95, 0.99	0.55 at 1.00; 0.56 at 0.75; 0.57 at 0.50	0.99; 0.99; 1.00
α_{cv}	0.025, 0.05, 0.1	0.58 at 0.75; 0.56 at 0.75; 0.52 at 0.75	0.99; 0.99; 0.99
dispute overhead	0.025, 0.05, 0.1	0.56 at 0.75; 0.56 at 0.75; 0.56 at 0.75	0.99; 0.99; 0.99

7. Discussion and Future Research

The results support a bounded interpretation of verifiable agent exchange. Geometric-mean settlement is exposure-optimal within the class of receipt-conditioned payment rules, while joint satisfaction of the four coordination constraints remains conditional on the stated assumptions and parameter inequalities (Corollary 3). The mechanism therefore does not eliminate trust. It reallocates trust from individual counterparties to credential issuers, verification processes, authority gates, and settlement rules. This reallocation makes dependencies explicit, but it also makes verifier quality and enforcement-path coverage central points of failure.

7.1. Implications

For human principals, Proposition 2 limits mechanism-mediated spending and actions to an encoded authority envelope. Theorem 1 complements this control by bounding

multiplicative mispricing within an unresolved value fiber by $\sqrt{s(c_v)}$. Neither result establishes that an authorized action is prudent or that the exchanged asset has high intrinsic value. Rather, they bound what the agent may do and how severely residual uncertainty can distort settlement. Verification budget is consequently a design variable rather than an objective to maximize: additional verification narrows exposure but consumes resources, which explains the interior welfare optimum observed in the numerical analysis.

For digital artifacts, settlement-visible evidence and the delivered object can coincide. Stronger provenance for data, models, compute, and agent-generated outputs can therefore narrow the range of values consistent with the available evidence. Its economic benefit lies in reducing pricing uncertainty, not in creating value or guaranteeing quality. This distinction also marks the boundary of the formal results. Once an agent acts in the physical world, evidence no longer consists of the delivered artifact itself but of records describing an external event. Section 7.3 develops this limitation.

Reallocating trust from counterparties to infrastructure carries its own exposure. Proposition 4 treats a verifier that is honest but noisy, so its errors are random and bounded by measurable rates. Capture of, or collusion with, a credential issuer or verifier is a different failure, because the evidence itself becomes adversarial rather than imprecise. Failure of an authority gate or settlement contract is different again, because it invalidates the enforcement path that Proposition 2 assumes. No rate parameter in the present model represents either case. Governing these dependencies requires the separation of duties, independent attestation, contract assurance, and accountable infrastructure governance rather than stronger conditions on the mechanism [47].

Deployment on a blockchain adds transaction overhead that the welfare model does not price. Let ΔG denote the expected incremental on-chain cost per transaction relative to Receipt-Only, excluding the verification cost already carried by c_v . It covers the amortized gas of credential, escrow, and stake operations together with the probability-weighted gas of audit and arbitration calls. Table 7 gives a welfare advantage of $0.549 - 0.150 = 0.399$ per unit of high-quality delivery value, which the calibration normalizes to one. Under proportional scaling, the net modeled advantage remains positive whenever that value exceeds $\Delta G / 0.399 \approx 2.51\Delta G$. Because geometric-mean settlement pays $P^* = \sqrt{1.00 \times 0.25} = 0.50$ in the same units, the equivalent condition on the settlement amount is $P^* > 1.25\Delta G$. Both ratios are specific to the Receipt-Only comparator and to this calibration, and a stronger alternative narrows the margin. Settlement latency remains outside the welfare model.

7.2. Research Roadmap

Figure 13 organizes seven research areas that follow from limitations of the analyzed mechanism. They identify extensions of the framework rather than the results established in this study.

Agency theory: Classical agency theory models hidden effort but generally treats the agent's objective as given. For language-model agents, the operational objective may also be opaque, and specification gaming may satisfy a literal contract while defeating its intended purpose. Future work should identify incentive structures and process evidence that reveal compliance with the principal's objective rather than only with an observable output specification.

Transaction-cost economics: Agents can reduce search and bargaining costs without proportionately reducing the cost of verifying delivery [50,51]. If verification becomes the dominant transaction cost, organizational boundaries may increasingly follow the obligations that are difficult to verify rather than those that are difficult to negotiate. Testing this proposition requires transaction-level cost evidence from deployed agent markets.

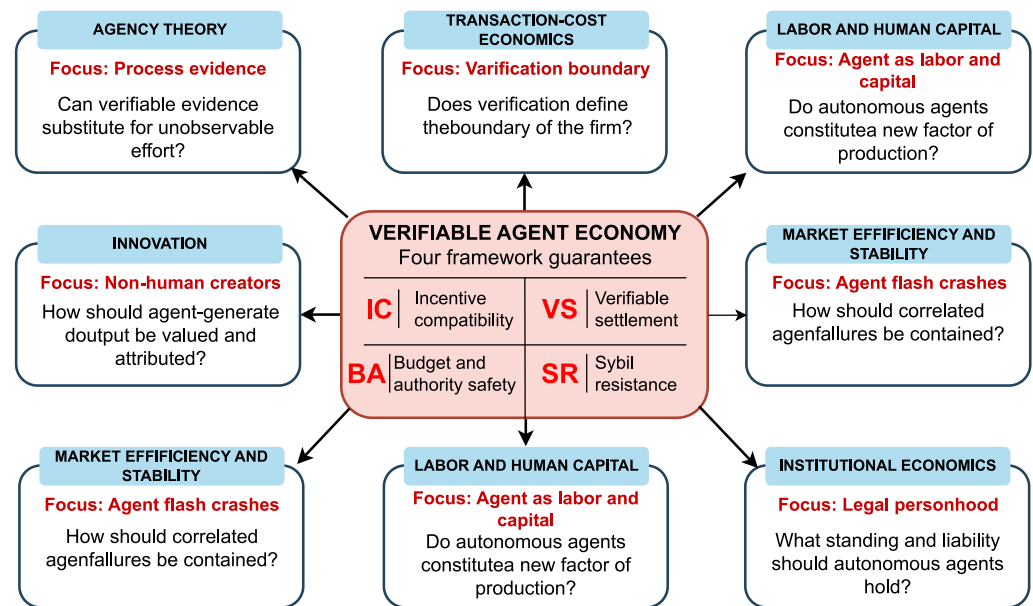


Figure 13. Research roadmap linking the framework's four requirements to open questions in agency theory, transaction-cost economics, institutional economics, labor theory, market stability, innovation, and distributed learning.

Institutional economics and legal personhood: Technically bounded authority does not confer legal standing, and a slashed stake is a private enforcement instrument rather than a legal remedy. Liability, jurisdiction, and the assets against which an autonomous agent can be held accountable therefore remain unresolved [53]. Future research must determine whether responsibility remains with the principal, shifts to service providers, or requires a distinct legal status for autonomous agents.

Labor and human capital: Autonomous agents combine labor-like production with capital-like reproducibility. This dual character complicates the allocation of ownership returns, bargaining power, and factor shares. Empirical analysis will require explicit records of agent ownership, contribution, and reputation before those distributions can be measured.

Market efficiency and stability: The present model analyzes individual deviations and does not characterize aggregate dynamics. Agents trained on shared data or using similar pricing rules may produce correlated behavior, with implications for price informativeness and shock propagation [56,57]. Future models should address tacit collusion, correlated cascades, circuit breakers, decentralized-finance failures [58], and network diffusion [59].

Innovation: Low-cost algorithmic experimentation may accelerate the creation of products, prices, and contract forms, while making attribution more difficult. Provenance evidence could record how agents combine purchased inputs, but it cannot by itself determine ownership or allocate the resulting returns. The open problem is how such evidence should interact with contractual and institutional rules for machine-generated innovation.

Distributed and federated learning: Agent economies may require policy improvement across organizations that cannot pool proprietary data. Credentials, staking, and reputation can govern participation, but they do not ensure secure aggregation, valid model updates, or privacy. Future work should connect these properties to the four coordination constraints, potentially through proofs of prescribed update procedures under explicit computational and verifier assumptions.

7.3. Embodied Agents and the Physical Evidence Chain

Embodied agents expose the clearest boundary between what the framework can express and what its results establish. Consider a delivery robot contracted to move a temperature-controlled package under the obligation “delivered to address A before time $\bar{t}$, with container temperature never above $8\text{ }^{\circ}\text{C}$ ”. The framework’s components can be mapped to this transaction: an operator-bound credential establishes identity; an authority envelope restricts route, time, and expenditure; escrow conditions payment on a verification receipt; and the verified outcome updates reputation. Encoding such an envelope requires a representation of agent behavior in space and time, for which space–time-aware behavioral models offer one starting point [60]. Unlike a digital artifact, however, the delivery must be represented by signed position and time traces, a sealed-container temperature log, and a recipient acknowledgment produced by attested sensing modules. If these records are trustworthy, c_v can represent sampling and audit depth, $[v_{\min}, v_{\max}]$ can represent the range consistent with the sampled trace, and sensor drift can enter Proposition 4 as a false-positive rate. This mapping demonstrates expressibility, not a formally proven extension.

The unresolved problem is the authenticity of the physical evidence. In the digital setting, the delivered artifact can itself supply the settlement evidence. In the physical setting, the evidence is a claim about the world produced by devices that the acting agent may own or influence. Additional readings from a compromised thermometer do not close this gap. The extension therefore requires an assumption absent from Assumption 2: sensing must be rooted in hardware or institutions that the acting agent cannot manipulate. Trusted execution environments, tamper-evident logs, and independent sensing are possible foundations, but each introduces cost and an additional trust dependency. Future research must determine when such an evidence chain is economically justified and which physical obligations should remain outside autonomous settlement when it cannot be established.

8. Conclusions

Autonomous transactions require explicit rules for evidence, authority, identity, and settlement rather than assumed trust, and those rules have to be designed into the economic infrastructure rather than supplied after a failure. This paper turned that requirement into a formal object: a definition with testable criteria, a layered reference model, a coordination problem with four constraints, and a mechanism whose guarantees are each stated at the strength it earns. The exposure bound is unconditional within its rule class, while delivery incentive compatibility and Sybil resistance hold under stated deterrence and identity-cost conditions. Keeping the two apart is the framework’s practical content. It tells a designer which properties follow from the mechanism itself, and which have to be bought with stake, patience, or verification budget in a particular market. The next step is to instantiate the mechanism on deployed identity, wallet, verification, and settlement primitives and test it with real agents and tasks, which is the evidence needed to establish where the model’s assumptions hold, where implementation cost dominates, and where the formal account must be revised.

Author Contributions: Conceptualization and methodology, M.M.K. and Q.Q.; validation, Q.Q.; investigation, resources, and data curation, M.M.K.; writing—original draft preparation, M.M.K.; writing—review and editing, Q.Q.; visualization, M.M.K.; supervision, funding acquisition, and project administration, Q.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Basic and Applied Basic Research Foundation of Guangdong Province, grant number 2023TQ07A264.

Data Availability Statement: All code, data, and scripts underlying the reported figures and tables are publicly available at <https://github.com/karimmd/agent-economy> (accessed on 27 September 2026). Any further material is available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rothschild, D.M.; Mobius, M.M.; Hofman, J.M.; Dillon, E.W.; Goldstein, D.G.; Immorlica, N.; Jaffe, S.; Lucier, B.; Slivkins, A.; Vogel, M. The Agentic Economy. *Commun. ACM* **2026**, *69*, 39–42. [CrossRef]
2. Zhu, S.; Sun, J.; Nian, Y.; South, T.; Pentland, A.; Pei, J. The Automated but Risky Game: Modeling Agent-to-Agent Negotiations and Transactions in Consumer Markets. In *Proceedings of the Natural Language Processing Workshop*; ACL: Stroudsburg, PA, USA, 2025; p. 16. [CrossRef]
3. Li, Y.; Wang, L.; Wang, K.; Yang, Z.; Wang, K.; Guan, Z.; Gao, J. A402: Binding Cryptocurrency Payments to Service Execution for Agentic Commerce. *arXiv* **2026**, arXiv:2603.01179.
4. Huang, D.; Chua, J.K.; Wang, Z. Beyond Task Success: Measuring Workflow Fidelity in LLM-Based Agentic Payment Systems. In *Trends and Applications in Knowledge Discovery and Data Mining (PAKDD 2026 Workshops)*; Lecture Notes in Computer Science; Springer: Singapore, 2026; Volume 16603, pp. 357–362. [CrossRef]
5. Li, Z.; Wang, Q.; Wang, Z. Five Attacks on x402 Agentic Payment Protocol. *arXiv* **2026**, arXiv:2605.11781.
6. Ling, S.; Chen, Y.; Huang, Y.; Zhou, Y.; Du, Y.; Wu, L.; Wang, C. Free-Riding the Agentic Web: A Systematic Security Analysis of x402 Payments. *arXiv* **2026**, arXiv:2605.30998.
7. Yu, J.; Zhou, S.; Yin, H.; Seong, B. PASS: A Provenanced Access Subaccount System for Blockchain Wallets. *arXiv* **2026**, arXiv:2604.22602.
8. Holmstrom, B. Moral Hazard and Observability. *Bell J. Econ.* **1979**, *10*, 74–91. [CrossRef] [PubMed]
9. Deng, Z.; Guo, Y.; Han, C.; Ma, W.; Xiong, J.; Wen, S.; Xiang, Y. AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. *ACM Comput. Surv.* **2025**, *57*, 1–36. [CrossRef]
10. Bandi, A.; Kongari, B.; Naguru, R.; Pasnoor, S.; Vilipala, S.V. The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. *Future Internet* **2025**, *17*, 404. [CrossRef]
11. Karim, M.M.; Khan, S.; Van, D.H.; Liu, X.; Wang, C.; Qu, Q. Transforming Data Annotation with AI Agents: A Review of Architectures, Reasoning, Applications, and Impact. *Future Internet* **2025**, *17*, 353. [CrossRef]
12. Tomasev, N.; Franklin, M.; Leibo, J.Z.; Jacobs, J.; Cunningham, W.A.; Gabriel, I.; Osindero, S. Virtual Agent Economies. *arXiv* **2025**, arXiv:2509.10147.
13. Xu, M. The Agent Economy: A Blockchain-Based Foundation for Autonomous AI Agents. *arXiv* **2026**, arXiv:2602.14219.
14. Gondauri, D.; Batiashvili, M. The Agentic Economy: Humans, AI Agents, Robots, and the Measurable Transition toward Distributed Economic Action. *arXiv* **2026**, arXiv:2605.18935.
15. Jin, Y.; Wu, S.; Chen, C.; Bao, L.; Yang, X.; Chen, J. The Web4 Agent Economy: A Large-Scale Empirical Study of the Landscape, Challenges, and Opportunities. *arXiv* **2026**, arXiv:2606.25876.
16. Mafrur, R.; Khusumanegara, P. From Agent Identity to Agent Economy: Measuring the Operational Readiness of ERC-8004 AI Agents. *arXiv* **2026**, arXiv:2606.12128.
17. Yang, K.; Zhai, C. Ten Principles of AI Agent Economics. *arXiv* **2025**, arXiv:2505.20273.
18. Chen, Y.; Chen, J.; He, C.; Li, Y.; Ji, Y.; Wu, Y.; Yang, D.; Diao, L.; Shou, L.; Zhang, H.; et al. Token Economics for LLM Agents: A Dual-View Study from Computing and Economics. *arXiv* **2026**, arXiv:2605.09104.
19. Yang, Y.; Wen, Y.; Wang, J.; Zhang, W. Agent Exchange: Shaping the Future of AI Agent Economics. *arXiv* **2025**, arXiv:2507.03904.
20. Chaffer, T.J. Can We Govern the Agent-to-Agent Economy? *arXiv* **2025**, arXiv:2501.16606.
21. Lovén, L.; Saleh, A.; Farahani, R.; Murturi, I.; López, M.B.; Donta, P.K.; Dustdar, S. Real-Time AI Service Economy: A Framework for Agentic Computing Across the Continuum. *arXiv* **2026**, arXiv:2603.05614v1.
22. Ruan, A.; Zhang, X. AgentCity: Constitutional Governance for Autonomous Agent Economies via Separation of Power. *arXiv* **2026**, arXiv:2604.07007.
23. Yang, H.; Zhang, B.; Wang, N.; Guo, C.; Zhang, X.; Lin, L.; Wang, J.; Zhou, T.; Guan, M.; Zhang, R.; et al. FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models. *arXiv* **2024**, arXiv:2405.14767.
24. Baxi, R. The Comprehension-Gated Agent Economy: A Robustness-First Architecture for AI Economic Agency. *arXiv* **2026**, arXiv:2603.15639.
25. Ortega, D. AI on Arbitrum: Establishing an Agent Registry with ERC-8004. Arbitrum Foundation Blog, 2025. Informal Use of the Phrase “Verifiable Agent Economy”. Available online: <https://blog.arbitrum.foundation/ai-on-arbitrum-establishing-an-agent-registry-with-erc-8004-2/> (accessed on 27 September 2026).

26. Zhang, Y.; Xiang, Y.; Lei, Y.; Wang, Q.; Qiu, T.; Sun, Y.; Zarkov, S.; Yuen, T.H.; Deppeler, A.; Yu, J.; et al. SoK: Blockchain Agent-to-Agent Payments. *arXiv* **2026**, arXiv:2604.03733.
27. de Valois-Franklin, A.; Bogdan, A. RAILS: Verification-Native Clearing for Agentic Commerce. *arXiv* **2026**, arXiv:2606.08790.
28. Shen, Y.; Zhang, J.; Shao, Z.; Luo, W.; Wang, Y.; Chen, T.; Zheng, Z.; Chen, J. Web3 × AI Agents: Landscape, Integrations, and Foundational Challenges. *arXiv* **2025**, arXiv:2508.02773.
29. Bansal, G.; Hua, W.; Huang, Z.; Fournay, A.; Swearngin, A.; Epperson, W.; Payne, T.; Hofman, J.M.; Lucier, B.; Singh, C.; et al. Magentic Marketplace: An Open-Source Environment for Studying Agentic Markets. *arXiv* **2025**, arXiv:2510.25779.
30. Rochet, J.C.; Tirole, J. Platform Competition in Two-Sided Markets. *J. Eur. Econ. Assoc.* **2003**, *1*, 990–1029. [[CrossRef](#)]
31. Manning, B.S.; Horton, J.J. *General Social Agents*; NBER Working Paper 34937; National Bureau of Economic Research: Cambridge, MA, USA, 2026.
32. Immorlica, N.; Lucier, B.; Slivkins, A. Generative AI as Economic Agents. *ACM SIGecom Exch.* **2024**, *22*, 93–109. [[CrossRef](#)]
33. del Rio-Chanona, R.M.; Pangallo, M.; Hommes, C. Can Generative AI Agents Behave Like Humans? Evidence from Laboratory Market Experiments. *arXiv* **2025**, arXiv:2505.07457.
34. Korinek, A. *AI Agents for Economic Research*; NBER Working Paper 34202; National Bureau of Economic Research: Cambridge, MA, USA, 2025.
35. Dawid, H.; Harting, P.; Wang, H.; Wang, Z.; Yi, J. Agentic Workflows for Economic Research: Design and Implementation. *arXiv* **2025**, arXiv:2504.09736.
36. Hao, Y.; Xie, D. A Multi-LLM-Agent-Based Framework for Economic and Public Policy Analysis. *arXiv* **2025**, arXiv:2502.16879.
37. Calvano, E.; Calzolari, G.; Denicolo, V.; Pastorello, S. Artificial Intelligence, Algorithmic Pricing, and Collusion. *Am. Econ. Rev.* **2020**, *110*, 3267–3297. [[CrossRef](#)]
38. Klein, T. Autonomous Algorithmic Collusion: Q-Learning Under Sequential Pricing. *Rand J. Econ.* **2021**, *52*, 538–558. [[CrossRef](#)]
39. Fish, S.; Gonczarowski, Y.A.; Shorrer, R.I. Algorithmic Collusion by Large Language Models. *arXiv* **2024**, arXiv:2404.00806.
40. Motwani, S.R.; Baranchuk, M.; Strohmeier, M.; Bolina, V.; Torr, P.H.S.; Hammond, L.; de Witt, C.S. Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2024.
41. Affonso, F.M. Vertical Tacit Collusion in AI-Mediated Markets. *arXiv* **2026**, arXiv:2601.03061.
42. Abada, I.; Lambin, X. Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided? *Manag. Sci.* **2023**, *69*, 5042–5065. [[CrossRef](#)]
43. Keppo, J.; Li, Y.; Tsoukalas, G.; Yuan, N. On the Fragility of AI Agent Collusion. *arXiv* **2026**, arXiv:2603.20281.
44. Qi, M.; Zhu, T.; Zhang, L.; Li, N.; Zhou, W. Towards Transparent and Incentive-Compatible Collaboration in Decentralized LLM Multi-Agent Systems: A Blockchain-Driven Approach. *arXiv* **2025**, arXiv:2509.16736.
45. Xia, Y.; Wang, T.; Xu, W.; Zhang, S. DAO-Agent: Zero Knowledge-Verified Incentives for Decentralized Multi-Agent Coordination. *arXiv* **2025**, arXiv:2512.20973.
46. Qu, Q.; Liu, S.; Zhu, F.; Jensen, C.S. Efficient Online Summarization of Large-scale Dynamic Networks. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3231–3245. [[CrossRef](#)]
47. Alqithami, S. Autonomous Agents on Blockchains: Standards, Execution Models, and Trust Boundaries. *arXiv* **2026**, arXiv:2601.04583.
48. Karim, M.M.; Khan, S.; Qu, Q.; Muzammal, M.; Sharif, K.; Biswas, S. From Trust to Augmentation: A Comprehensive Survey on Synergistic Integration of Decentralized and Generative Intelligence. *Comput. Sci. Rev.* **2026**, *61*, 100936. [[CrossRef](#)]
49. Jensen, M.C.; Meckling, W.H. Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *J. Financ. Econ.* **1976**, *3*, 305–360. [[CrossRef](#)]
50. Coase, R.H. The Nature of the Firm. *Economica* **1937**, *4*, 386–405. [[CrossRef](#)]
51. Williamson, O.E. The Economics of Organization: The Transaction Cost Approach. *Am. J. Sociol.* **1981**, *87*, 548–577. [[CrossRef](#)] [[PubMed](#)]
52. Shahidi, P.; Rusak, G.; Manning, B.S.; Fradkin, A.; Horton, J.J. The Coasean Singularity? Demand, Supply, and Market Design with AI Agents. In *The Economics of Transformative AI*; NBER Working Paper 34468; National Bureau of Economic Research: Cambridge, MA, USA; University of Chicago Press: Chicago, IL, USA, 2025.
53. Riedl, M.O.; Desai, D.R. AI Agents and the Law. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*; AAAI: Washington, DC, USA, 2025; Volume 8, pp. 2189–2198.
54. Khan, M.A.; Amani, M.; Das, S.; Ghosh, B.; Wu, Q.; Gummadi, K.P.; Gupta, M.; Ravichander, A. In Agents We Trust, but Who Do Agents Trust? Latent Source Preferences Steer LLM Generations. *arXiv* **2026**, arXiv:2602.15456.
55. Yu, S.; Yang, M.; Qu, Q.; Shen, Y. Contextual-boosted Deep Neural Collaborative Filtering Model for Interpretable Recommendation. *Expert Syst. Appl.* **2019**, *136*, 365–375. [[CrossRef](#)]
56. Grossman, S.J.; Stiglitz, J.E. On the Impossibility of Informationally Efficient Markets. *Am. Econ. Rev.* **1980**, *70*, 393–408.
57. Hayek, F.A. The Use of Knowledge in Society. *Am. Econ. Rev.* **1945**, *35*, 519–530.

58. Ante, L. Autonomous AI Agents in Decentralized Finance: Market Dynamics, Application Areas, and Theoretical Implications. *Technol. Forecast. Soc. Change* **2026**, *228*, 124669. [[CrossRef](#)]
59. Chen, Y.; Qu, Q.; Ying, Y.; Li, H.; Shen, J. Semantics-aware Influence Maximization in Social Networks. *Inf. Sci.* **2020**, *513*, 442–464. [[CrossRef](#)]
60. Qu, Q.; Chen, C.; Jensen, C.S.; Skovsgaard, A. Space-Time Aware Behavioral Topic Modeling for Microblog Posts. *IEEE Data Eng. Bull.* **2015**, *38*, 58–67.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.