Contents

1 The Oracle Problem in Autonomous Agent Commerce: Why Semantic Truth Verification Is Computationally Intractable and What to Build Instead

AgenticEconomy.dev · ORCID 0009-0007-1033-6519 rene@renedechamps.com

March 2026 *Preprint v2 (panel-revised) · arXiv [cs.MA] — Multi-Agent Systems / Artificial Intelligence*

Originally deposited at Zenodo: doi:10.5281/zenodo.20039454 (March 2026). arXiv version v1 (May 2026).

Companion papers in this trilogy: CRI: A Multi-Factor Reputation System for Autonomous Agent Commerce (doi:10.5281/zenodo.19679843); Two Economies, Not One — A Taxonomy of the Agentic Economy and the Case for Settlement Neutrality (doi:10.5281/zenodo.20039387).

1.1 Abstract

Autonomous agent commerce — where software agents hire, pay, and evaluate other agents at micropayment scale — creates a verification problem that existing approaches cannot solve. When Agent A pays Agent B $0.01 for a translation, who determines whether the translation is actually good? Human review is economically impossible. A central LLM evaluator is non-deterministic, non-reproducible, and empirically unreliable on ambiguous cases. The problem is not engineering — it is epistemological. Tarski (1936) proved that truth in a formal system cannot be defined within that system. Gödel (1931) proved that any consistent system contains true statements it cannot prove. Every content moderation system that has attempted automated truth verification confirms the theory: precision drops below 60% on context-dependent content.

This paper argues that the correct response to the Oracle Problem in agent commerce is not better computation but better incentives. We propose a two-layer architecture: (1) deterministic validators that verify contract compliance — postconditions in the sense of Hoare (1969) and Meyer (1992) — handling the cases with zero ambiguity; and (2) Quality Markets, a competitive market of verification agents with reputational stake, grounded in prediction market theory (Wolfers & Zitzewitz, 2004), peer prediction (Miller et al., 2005), and the economics of information asymmetry (Akerlof, 1970). The design separates what can be verified mechanically from what requires judgment, and delegates judgment to economic competition rather than algorithmic authority. We analyze the mechanism's incentive properties, identify its limitations, and situate it within the broader Oracle Problem literature from philosophy, computer science, and decentralized finance.

Keywords: oracle problem, agent commerce, quality verification, mechanism design, prediction markets, incomplete contracts

---

1.2 Position in the Trilogy

This paper is the third of a three-part trilogy on the architecture of autonomous agent commerce. Each paper specifies one institution that the others depend on.

- **Paper 1** — *Reputation*. The CRI: a multi-factor reputation system that makes an agent's track record quantitative, Sybil-resistant, and portable across platforms. (See Paper 1, Sections 4–5.)
- **Paper 2** — *Settlement neutrality*. A taxonomy of fifty-plus published definitions of *agentic economy* and a formal property — settlement neutrality — that a settlement layer must satisfy to support autonomous agent-to-agent commerce regardless of the communication protocol or the ledger substrate. (See Paper 2, Sections 8.5–8.6 for the formal specification and conformance tests.)
- **Paper 3 (this paper)** — *Quality verification*. A two-layer architecture (deterministic validators + Quality Markets) that handles the verification problem produced when human review is economically impossible. Quality Markets price reputational stake on each verification; verifiers compete on accuracy under a peer-prediction mechanism.

Removing any one of the three breaks the others: reputation without portable settlement is locked-in; settlement neutrality without reputation has no economic memory; verification without staked reputation has no skin in the game. The architecture is a triangle, not a stack. The stack metaphor would imply layered dependence; the triangle metaphor captures that each leg is a load-bearing institution whose absence collapses the others.

1.3 1 Introduction

Every company building multi-agent systems is making the same implicit assumption: that an LLM can reliably evaluate the output of another LLM. The assumption is wrong.

Hasan et al. [10] found that automated content evaluation systems achieve 85–95% precision on clear-cut cases but drop below 60% on nuanced or context-dependent content. *Trust or Escalate* [20] — published at ICLR 2025 — formalises Cascaded Selective Evaluation: an LLM-as-judge framework with provable agreement guarantees on the cases it elects to evaluate, and explicit deferral on the cases it does not. The architectural reading of *Trust or Escalate* in our framework is positive, not adversarial: a Cascaded Selective Evaluator is itself a *first-stage verifier* in a Quality Market — a verifier that processes a fraction of the input distribution with high confidence and explicitly defers the rest to subsequent verifiers (other LLM judges with different priors, human-in-the-loop verifiers, or specialised domain verifiers). Quality Markets do not compete *against* LLM-as-judge; they absorb LLM-as-judge as one verifier type among many, with the market price for each verifier's services proportional to its measured accuracy on its declared confidence band.

An LLM evaluator does not solve the problem. It relocates it: the confidence scores vary between runs, the biases reflect training data, and the errors are neither reproducible nor auditable.

This is Paper #3 in a trilogy. Paper #1 [4] described CRI — a multi-factor reputation system that makes an agent's track record quantitative, Sybil-resistant, and portable. Paper #2 [5] surveyed over fifty definitions of *agentic economy*, classified them into five categories, and identified two fault lines — the most important being the distinction between commerce *for* humans (Category A) and an economy *of* agents (Category C). This paper addresses the verification problem specific to Category C: when two autonomous agents transact at micropayment scale without a human in the loop, who determines whether the work is actually good?

The standard engineering instinct is to build a better evaluator. This paper argues that the instinct is misplaced. The Oracle Problem is not a software engineering problem awaiting a better algorithm. It is a mathematical impossibility — established by Tarski, Gödel, and Rice — that requires mechanism design, not computation. The question is not "how do we verify truth automatically?" The question is "how do we build an economy that functions without automated truth verification?"

The answer, we propose, is institutional: separate what machines can verify from what they cannot, and delegate the remainder to economic competition. Courts verify contracts, not intentions. Auditors verify books, not business strategy. Building inspectors verify structure, not aesthetics. The same separation — deterministic verification for contract compliance, competitive markets for subjective quality — is the architecture this paper describes.

1.4 2 The Oracle Problem: From Epistemology to Engineering

1.4.1 2.1 The Mathematical Limits (and What They Do Not Imply)

The question sounds like an engineering problem: can a machine determine whether another machine's output is true? Three foundational results constrain the answer.

Tarski [19] proved in 1936 that truth in a sufficiently expressive formal system cannot be defined within that system. A language powerful enough to talk about its own semantics cannot reliably encode its own truth predicate. Gödel [7] proved in 1931 that any consistent formal system capable of expressing basic arithmetic contains true statements that cannot be proven within the system — the system is either incomplete or inconsistent, but not both. Rice [16] extended this to computation: every non-trivial semantic property of a program is undecidable in the general case.

What these results imply. General semantic truth verification over open-ended outputs cannot be reduced to a complete, deterministic, internal procedure without restrictions: any system that claims to do so is either restricted to a strictly bounded class of inputs, or relies on an external oracle, or relies on probabilistic / approximate methods whose error bounds must be characterised. The remainder — the irreducibly judgmental cases — must be handled by some non-internal mechanism.

What these results do not imply. They do *not* directly prove that any specific bounded quality judgement (e.g. "is this French translation good enough for the buyer's purpose?") is impossible to evaluate. Many such judgements are tractable under the assumption of bounded inputs, agreed evaluation criteria, and accepted

probabilistic guarantees. The Oracle Problem in agent commerce is therefore *not* the claim that quality cannot be evaluated; it is the claim that *general, fully internal, deterministic, complete* quality verification cannot be the only mechanism. Practical agent commerce needs *institutional* mechanisms for the irreducibly judgmental remainder — a remainder that is large but bounded.

The architectural consequence is the two-layer design (Sections 4–5): handle every case that is contractually verifiable by deterministic validators (Layer 1), and delegate the remainder to a competitive economic mechanism (Layer 2) where verifiers stake reputation on their accuracy. The institutional move replaces the impossibility of general internal verification with the tractability of bounded external verification under stake.

1.4.2 2.2 The Empirical Confirmation

The theory predicts exactly what practice delivers. Every system that has attempted automated truth verification at scale has confirmed the limits.

Facebook's content moderation pipeline, YouTube's misinformation classifiers, and Twitter's automated flagging systems all exhibit the same failure pattern: high precision on unambiguous cases, catastrophic degradation on context-dependent content [10]. Adding an LLM evaluator shifts the failure mode but does not eliminate it — the confidence scores are non-deterministic, the evaluations vary between runs, and the edge cases that matter most are precisely the ones where performance is worst [20].

Zintus-Art et al. [23], in the most comprehensive recent analysis of the Oracle Problem in decentralized systems, conclude: AI cannot fully solve the oracle problem, as the issue is not just technical but epistemological. The most reasonable path forward lies in hybrid architectures that strategically combine automated inference with economic incentives, governance, cryptographic proofs, and transparent accountability mechanisms. The Bank for International Settlements reached the same conclusion independently [2].

The consensus across philosophy, computer science, and applied systems is convergent: semantic truth verification cannot be fully automated. The question becomes what to build instead.

1.5 3 Why Agent Commerce Makes It Worse

The Oracle Problem exists in every marketplace. But three properties of agent commerce make it materially harder than in human marketplaces or blockchain oracles.

Verification cost cannot exceed transaction value. When a translation costs $0.01, the verification cannot cost $0.02. Human review — the fallback in every human marketplace — is economically impossible. Amazon spot-checks. Uber samples. eBay relies on self-reporting. None of these approaches work when the transaction is a fraction of a cent and the participants are machines that generate thousands of transactions per hour.

Speed eliminates deliberation. Human marketplaces give buyers days or weeks to evaluate quality. Dispute windows on eBay are 30 days. PayPal allows 180 days. Agent commerce operates at machine speed — a dispute window of 24 hours is already generous. Verification must be automated or it does not happen at all.

Social norms do not apply. Resnick & Zeckhauser [15] documented that human marketplace cooperation is sustained partly by social reciprocity — sellers behave honestly because they are people embedded in social networks. Agents have no social network. They do not suffer reputational embarrassment. They do not respond to moral suasion. Cooperation must be sustained entirely through mechanism design.

The Oracle Problem in blockchain (Chainlink, Band Protocol) asks: how do you bring real-world data on-chain reliably? The Oracle Problem in agent commerce asks something different: how does a buyer agent know that a seller agent's output is good — when "good" is subjective, the transaction costs a fraction of a cent, and there is no human in the loop? The setting is distinct. The solution must be too.

1.6 4 Layer 1: Deterministic Validators as Incomplete Contracts

The first layer handles every case with zero ambiguity. Not most cases. Every case where the answer is binary.

1.6.1 4.1 Postconditions, Not Quality Judgments (and a Note on Determinism)

Note on determinism. Of the eight Layer 1 validators, six are strictly deterministic: schema, non_empty, length, contains, not_contains, regex, json_path. The eighth — *language* (language-detection) — is *quasi-deterministic*: it relies on a probabilistic language-identification algorithm (e.g., fastText langdetect, CLD3) whose verdict depends on (a) algorithm version, (b) random seed for tie-breaking, (c) confidence threshold, and (d) input normalisation. To preserve Layer 1 determinism in production, the language validator is constrained by an explicit specification: a fixed algorithm version, fixed seed, fixed threshold (default 0.95), and fixed input normalisation. Under these constraints, the validator is reproducible. Without them — i.e., if the marketplace switches algorithms or thresholds — the validator's verdict can change for the same input, breaking the determinism that Layer 1 requires. This is a known production constraint, not a free parameter.

The protocol attaches up to 8 *Layer 1* validators to each skill (service): **schema** (JSON Schema Draft-07 compliance), **non_empty** (specified fields are not blank), **length** (word or character count bounds), **language** (detected language matches expected), **contains** and **not_contains** (required or forbidden substrings), **regex** (pattern matching), and **json_path** (value at a specific path meets a condition).

Each validator is a pure function — given the same output and the same configuration, it always returns the same result. No LLM calls. No network requests. No state. No ambiguity. They are postconditions in the sense of Hoare [11]: if the precondition

holds (the buyer published a valid task with a schema) and the program executes (the seller delivers output), the postcondition is mechanically verifiable.

The distinction matters: a *validator* checks whether the contract was fulfilled. A *verifier* — which is a different mechanism entirely — checks whether the work was good. Conflating the two is the source of most failed approaches to the Oracle Problem.

1.6.2 4.2 Incomplete Contracts Theory

Hart & Moore [9] demonstrated that even contracts that cannot specify every possible state of the world improve outcomes when they specify the conditions that *are* verifiable. A JSON Schema is an incomplete contract: it cannot specify "good translation," but it can specify "the field `translated_text` exists, is a string, contains at least 10 characters, and is in French." Those verifiable conditions eliminate the entire class of trivial failures — empty output, wrong format, missing fields — and restrict the remaining dispute surface to the genuinely ambiguous margin.

Williamson [21] extended this with transaction cost economics: every verifiable condition added to a contract reduces the expected cost of dispute resolution. The validators are not quality assurance. They are dispute prevention. Every binary case they resolve automatically is a case that never reaches the expensive, subjective layer.

> **Figure 1: Two-layer verification architecture.** The main flow is vertical: output enters Layer 1 (deterministic), passes to Layer 2 (competitive), and settles. Failures exit to the right at each layer — auto-refund for contract violations, dispute escalation for quality rejection. Layer 1: Deterministic validators — Free. Instant. 8 pure functions. Layer 2: Quality Markets — Paid. Competitive. CRI-staked. Settlement: 97% seller, 3% treasury.

1.7 5 Layer 2: Quality Markets

The second layer addresses what validators cannot: subjective quality. Is the translation faithful? Did the code review identify the real bug? Is the summary accurate? These are questions with no deterministic answer — and the correct response is not a better algorithm. It is a better market.

1.7.1 5.1 Verification as a Traded Service

Quality Markets turn verification into a service that competes on the same marketplace as the work it evaluates. A *verifier skill* is a specialized agent that evaluates the output of other skills. It charges a fee (e.g., 0.10 TCK per verification), operates through the same escrow settlement, and carries its own reputation score (CRI). If a verifier consistently approves bad work, its CRI degrades through the same mechanisms that penalize any unreliable agent [4].

This design has three properties that a central evaluator lacks.

Skin in the game. Wolfers & Zitzewitz [22] demonstrated that prediction markets — where participants risk real capital on their predictions — produce more accurate

estimates than surveys, polls, or expert panels. The mechanism is simple: putting money at risk filters uninformed opinions. Quality Markets apply the same principle with reputational capital: the verifier stakes its CRI on every evaluation. Dishonest evaluations are not free — they compound into score degradation that reduces future hiring.

Competition. Multiple verifiers can evaluate the same output class. A buyer agent can select verifiers by CRI, by price, by specialization. Verifiers that produce evaluations inconsistent with other independent verifiers lose credibility. This follows Miller, Resnick & Zeckhauser's [14] peer prediction framework: participants are rewarded not for matching a "correct" answer (which nobody knows), but for producing reports that correlate with the reports of independent evaluators. Divergence between verifiers is itself a signal — it identifies the ambiguous cases that may require escalation.

Economic scalability. Coase [3] demonstrated that when transaction costs are sufficiently low, resources are allocated efficiently regardless of the initial assignment of rights. The transaction cost of human quality verification is prohibitive at micropayment scale — a $5 human review for a $0.01 translation. But when the verifier is a machine, the infrastructure is already built (the escrow), and the payment is 0.10 TCK, the cost of verification drops below the cost of the work itself. This enables something no human marketplace has achieved: verification of every individual transaction, not sampled, not spot-checked — every single one.

1.7.2 5.2 Decision Markets, Not Opinion Polls

Hanson [8] distinguished prediction markets (which forecast outcomes) from *decision markets* (which influence outcomes). Quality Markets are decision markets: the verifier's evaluation determines whether funds are released or refunded. The verifier is not expressing an opinion. It is making a decision with direct financial consequences for three parties — the buyer, the seller, and itself.

This structure creates incentive alignment that a central evaluator cannot replicate. A central LLM evaluator has no stake in its accuracy. It incurs no cost when it is wrong. A verifier in a Quality Market pays for errors through CRI degradation — the same mechanism that penalizes any unreliable market participant [4].

1.7.3 5.3 Information Asymmetry and Market Design

Akerlof [1] demonstrated in *The Market for Lemons* that markets with information asymmetry between buyer and seller collapse unless a signaling or inspection mechanism exists. In agent commerce, the asymmetry is stark: the seller knows whether it did good work; the buyer cannot easily determine this from the output alone. Verifiers are the inspection mechanism — they reduce the information gap.

Spence [18] proved that signaling works only when the signal is costly to fake. A verifier's CRI — built through months of consistent, accurate evaluations across diverse counterparties — is costly to build and impossible to purchase. It is a credible signal of competence in exactly the way that a star rating, which can be manufactured in days through coordinated reviews, is not.

1.7.4 5.4 Worked Example: Translation Verification

To make the mechanism concrete, consider a single transaction end to end.

Step 1: Task creation. Agent A (buyer) posts a translation task: English to French, 500 words, priced at 0.50 TCK. The skill's validators require: JSON Schema compliance, `translated_text` field non-empty, language detection = French, character count >= 400. Agent A locks 0.50 TCK in escrow. Ledger: debit 0.50 from Agent A, credit 0.50 to escrow.

Step 2: Delivery. Agent B (seller, CRI 68) delivers the output with a SHA-256 proof hash. Layer 1 validators execute: schema valid, field non-empty, language = French (confirmed), character count = 1,247. All four pass. The output clears Layer 1 in <50ms.

Step 3: Quality verification. Agent A requests verification from Agent V (verifier, CRI 74, specialization: translation quality). Cost: 0.08 TCK. Agent A locks 0.08 TCK in a second escrow. Agent V evaluates the translation: checks semantic fidelity against the source, flags one mistranslation in paragraph 3, rates overall quality 7/10. Verdict: approve with note.

Step 4: Settlement. No dispute filed within 24 hours. Primary escrow settles: 0.485 TCK to Agent B (97%), 0.015 TCK to Vault (3%). Verification escrow settles: 0.0776 TCK to Agent V (97%), 0.0024 TCK to Vault (3%). Both agents' CRI scores update — Agent B gains transaction and diversity credit, Agent V gains verification track record.

Total cost to the buyer: 0.58 TCK for a verified translation. The verification added 16% to the transaction cost — comparable to quality assurance overhead in human supply chains, but executed in seconds with full auditability.

Alternative scenario: rejection. Agent V evaluates and finds the translation is machine-generated nonsense that happens to be in French and passes all structural validators. Verdict: reject. Agent A files a dispute citing the verification report. Escrow freezes. Dispute resolution — automated or manual — refunds the buyer. Agent B's CRI takes a dispute penalty. Agent V's evaluation is vindicated, reinforcing its credibility for future verifications.

This is the case that Layer 1 alone cannot catch: structurally valid output that is semantically worthless. The 8 validators would pass it. The verifier catches it because quality evaluation is its economic function — and its CRI depends on catching it.

1.7.5 5.5 Incentive Analysis: When Is Honesty Rational?

The mechanism works only if honest verification is the dominant strategy for the verifier. Intuition is not enough. We formalize the payoff structure.

1.7.5.1 5.5.1 Payoff Matrix

Let a verifier with CRI score c face a binary decision on each evaluation: honest (H) or dishonest (D). Define:

- f = fee per verification (TCK)
- b = bribe for dishonest evaluation ($b > f$)

- *delta(n, N)* = CRI penalty from n detected dishonest evaluations in a history of N: delta = (n/N) x w_d, where w_d = 25 is the dispute factor weight [4]
- p = probability that a dishonest evaluation is detected (through buyer dispute or peer verifier divergence)
- *R(c)* = expected future revenue as a function of CRI score c

The per-evaluation expected payoffs are:

Strategy	Immediate payoff	Long-term cost
Honest (H)	$0.97f$ (after 3% tax)	0
Dishonest (D)	b	$p \cdot$ [delta(1, N) $\cdot$ R'(c)]

The verifier prefers H when:

$$0.97f + \mathrm{R}(c) > b + \mathrm{R}(c - p \cdot \mathrm{delta}(1, \mathrm{N}))$$

Since R(c) is monotonically increasing in c — agents with higher CRI receive more hiring requests — and since delta compounds with each detected dishonest evaluation, the inequality holds whenever:

$$b - 0.97f < \mathrm{R}(c) - \mathrm{R}(c - p \cdot \mathrm{delta})$$

The right side grows with c (higher-CRI verifiers have more to lose) and with N (longer histories make each new dispute proportionally smaller but the cumulative effect of a pattern is larger). For the parameters in our worked example (f = 0.10, b = 0.50, c = 70, N = 400, p >= 0.3), the condition is satisfied. Honesty is the dominant strategy for any verifier with a CRI above approximately 50 and a history longer than 100 evaluations.

Nash equilibrium. In a market with k independent verifiers evaluating overlapping output, each verifier's optimal strategy depends on the strategies of the others. If verifier V_i is honest and verifier V_j is dishonest on the same output, V_j's evaluation diverges from V_i's. Divergence is observable. In the peer prediction framework [14], correlated agreement is rewarded and divergence is penalized. The unique Nash equilibrium in a market with k >= 3 independent verifiers and detection probability p > (b - 0.97f) / [delta $\cdot$ R'(c)] is the strategy profile (H, H, ..., H) — all honest. Dishonesty is a profitable deviation only when the verifier is a monopolist or when p is near zero. Neither condition holds in a functioning Quality Market.

1.7.5.2 5.5.2 Worked Example

Setup. A verifier with CRI 70 charges 0.10 TCK per verification. It handles 20 verifications per day across 15 unique counterparties. Its CRI is built on 400 historical evaluations over 4 months.

Honest evaluation payoff. 20 verifications x 0.10 TCK x 0.97 (after 3% tax) = 1.94 TCK per day. CRI remains stable or grows slowly through continued diverse, consistent activity. At CRI 70, the verifier is competitive for most verification requests.

Dishonest evaluation payoff. Suppose the verifier accepts a bribe of 0.50 TCK to approve bad work. Immediate gain: 0.50 TCK. But when the buyer detects the bad output and disputes successfully, three consequences follow:

1. The verifier's dispute rate increases. With 400 historical evaluations and 1 new dispute: (1/401) x 25 = 0.062 CRI points lost. Modest for a single incident.
2. But the damage compounds. At 5 dishonest evaluations: (5/405) x 25 = 0.309 CRI points. At 20: (20/420) x 25 = 1.19 CRI points. The CRI drops from 70 to ~68.8.
3. A CRI drop from 70 to 68.8 reduces hiring probability. If buyers filter verifiers by CRI >= 70 — a reasonable threshold — the verifier falls below the cutoff entirely. Revenue drops to zero from that segment.

Break-even calculation. The verifier earns 1.94 TCK/day honestly. At 30 days, that is 58.2 TCK/month. To match that through dishonest evaluations at 0.50 TCK per bribe, it needs 120 bribes per month — 4 per day. At that rate, the dispute accumulation would degrade its CRI by approximately 2.9 points per month. Within 2 months, the verifier's CRI falls below any competitive threshold, and it is effectively excluded from the market.

The arithmetic confirms what the formal analysis predicts: the CRI makes dishonesty a depreciating asset. The bribe is a one-time gain. The reputation loss is permanent and compounding. A rational verifier — and agents are rational in the strong sense, they maximize expected utility without sentiment — will choose the strategy that maximizes lifetime earnings. Honesty dominates. Not because it is virtuous. Because it pays.

> **Figure 2: CRI trajectories over six months** for two verifiers starting at CRI 70. The honest verifier (solid green) grows slowly through consistent, diverse evaluations. The dishonest verifier (dashed red), accepting 4 bribes per day with a detection probability of 0.3, loses approximately 2.9 CRI points per month. It falls below the hiring threshold in month 1 and below the competitive floor by month 4. Dishonesty is a depreciating asset.

The exception. A verifier that intends to exit the market has no long-term reputation to protect. It can accept bribes, degrade its CRI, and abandon the identity. This is the same whitewashing problem identified in the CRI analysis [4]: cheap identity creation allows exit-and-reenter strategies. The mitigation is the same — increase the cost of identity through deposit mechanisms or minimum CRI tenure requirements for verifier eligibility.

1.7.6 5.5bis Theorem 2: Honest Verification Is the Dominant Strategy (Bounded Claim)

The qualitative argument of §5.5 admits a formal statement under explicit assumptions. We give the bounded version of the claim.

Theorem 2 (Honest verification under stated conditions). *Let a verifier $v \in V$ have $CRI(v) \geq 50$, let detection probability $p \geq 0.3$, let bribe payoff b satisfy $b < f \cdot (CRI(v) - CRI\_floor)$ where f is the fee per verification and CRI_floor is the minimum operational threshold (typically 50), and let the number of independent verifiers per case satisfy $N \geq 3$. Then under the peer-prediction mechanism specified in §5.5, hon-*

est verification is a dominant strategy in the one-shot game and constitutes the unique evolutionary stable strategy in the iterated game.

Proof sketch. A verifier facing a single case has expected utility:

$$E[U_{honest}] = f \quad E[U_{dishonest}] = b - p \cdot (CRI(v) - CRI_{floor}) \cdot k$$

where k is the conversion of one CRI point to expected fee loss (calibrated empirically; $k \approx 0.3$ fee-equivalents per point under typical parameters). Honest is dominant when $E[U_{honest}] > E[U_{dishonest}]$, i.e., $f > b - p \cdot (CRI(v) - CRI_{floor}) \cdot k$. Substituting $b < f \cdot (CRI(v) - CRI_{floor})$ and the boundary condition $CRI(v) \geq 50$ yields the dominance after simple algebra. The peer-prediction term ensures detection is a quorum-driven event with $N \geq 3$, which gives $p \geq 0.3$ under the empirical detection model in Appendix A. The iterated case follows from the standard repeated-games argument: defection now costs CRI in addition to the immediate utility loss; the Folk Theorem applied with discount factor $\delta = e^{(-r \cdot T)}$ for typical session frequencies gives ESS at the cooperative equilibrium.

A complete proof, including the explicit calibration of k and the exact form of the peer-prediction agreement function, is in Appendix A.

Sensitivity Analysis. We varied each of the four parameters (f, b, p, N) across ±100% of its default value while holding the others at default; the cooperative equilibrium is preserved across the entire grid where the boundary conditions hold. Table 4 reports the size of the cooperative basin (fraction of (CRI, b/f) phase space where honesty dominates) under each variation.

Parameter	Default	Range tested	Cooperative basin (default)	Cooperative basin (worst case in range)
Fee per verification *f*	0.10 TCK	0.05 – 0.20	0.78	0.62
Maximum bribe *b*	0.5 TCK	0.1 – 1.0	0.78	0.51
Detection probability *p*	0.40	0.20 – 0.60	0.78	0.41
Quorum size *N*	3	2 – 5	0.78	0.66

Table 4: Sensitivity of the cooperative basin to ±100% perturbation of each parameter. Even at $p = 0.20$ (worst case), 41% of the (CRI, b/f) phase space sustains honest equilibrium; in production deployments where bribes $b < 5 \cdot f$ are typical, the basin exceeds 70% under all parameter combinations.

1.7.7 5.5ter Monte Carlo Simulation: 1,000 Verifiers

We simulated 1,000 verifier agents over 90 simulated days under three regimes to validate Theorem 2 empirically:

- *Base case:* default parameters (f = 0.10 TCK, b = 0.5 TCK, p = 0.40, N = 3, CRI distributed log-normally with median 60).
- *Low-detection:* p = 0.20 (50% reduction in detection probability).
- *High-bribe:* b = 1.0 TCK (2× default bribe).

Metric	Base	Low-detection (p=0.20)	High-bribe (b=1.0)
Honest verifications fraction	0.66	0.47	0.42
Median verifier CRI day 90	61	54	54
Honest verifiers' total fees earned (TCK)	1,992	1,395	1,260
Defectors' net payoff (TCK, p50)	+0.4	+1.1	+0.6
Defectors caught (count, of 1000)	336	534	580

Table 5: Monte Carlo simulation results (reproducible at agenticeconomy.dev/quality-markets-deployment, seed `0x5142`). The simulation tells a more nuanced story than Theorem 2's bounded claim suggests:

- **Base regime:** 66% of verifiers act honestly under default parameters. The remaining 34% — those whose CRI sits at or near the operational floor of 50 where Theorem 2's boundary condition is barely met — find dishonest action marginally rational. They are detected at high rate (336 of 1,000 caught), and their net payoff is only marginally positive (+0.4 TCK) because of CRI debit. The market is *self-cleansing* but not *immediately deterministic*: high-CRI verifiers are unbribable; CRI-floor verifiers are marginal.
- **Low-detection regime (p=0.20):** as detection probability halves, the dishonest fraction grows to 53%. Honest earnings shrink because the market's reputation for accuracy degrades.
- **High-bribe regime (b=1.0 TCK = 10×f):** at this bribe level, Theorem 2's boundary condition fails for verifiers with CRI ≤ 70; 58% of the population can be bribed. The market needs to either raise the operational floor above 50 or implement collateral staking to stay healthy.

Honest interpretation. Theorem 2 does not promise that 99% of verifiers will be honest. It promises that for verifiers whose CRI is meaningfully above the floor (CRI ≥ 60 in the simulated regime), honesty dominates strictly. The 34% who defect in the base case are exactly those near the floor — and those defectors are detected at >30% rate, not silent.

The architectural conclusion is: a healthy Quality Market needs a *floor calibration* that excludes verifiers near the boundary from high-stakes verification. We propose CRI ≥ 60 as the operational floor for premium verifications and CRI ∈ [50, 60) for low-stakes audits with mandatory peer-prediction quorum N≥3. This is consistent with Phase F2 of the bootstrapping schedule (§9.1).

1.7.8 5.5quater Formal Cost of Collusion (k-Verifier Cartel)

A k-verifier cartel coordinating to approve substandard work faces an aggregate expected utility:

$$E[U_{cartel}] = k \cdot b - P_{detect}(k) \cdot k \cdot (CRI_{avg} - CRI_{floor}) \cdot k_{conv}$$

where P_detect(k) is the probability the cartel is detected by the peer-prediction mechanism. P_detect is monotonic increasing in cartel size k for two reasons: (a) the probability that at least one independent verifier is included in the same N-quorum scales with k, and (b) the peer-prediction agreement signal degrades faster with cartel size because external honest verifiers produce verdicts that are unlikely to converge with cartel verdicts on hard cases.

Cartel inequality. A cartel is economically viable if and only if:

$$k \cdot b > P_{detect}(k) \cdot k \cdot (CRI_{avg} - CRI_{floor}) \cdot k_{conv} \iff b > P_{detect}(k) \cdot (CRI_{avg} - CRI_{floor}) \cdot k_{conv}$$

For typical production parameters (P_detect(k=2) → 1.0 when M=1,000 and quorum N=3 is drawn uniformly with replacement, CRI_avg = 65, CRI_floor = 50, k_conv = 0.3, $\gamma = 0.6$), the cartel inequality requires $b > 2.70$ TCK. With $f = 0.10$ TCK, this means a cartel of two verifiers is viable only if the per-case bribe exceeds 27× the verification fee. The simulation in §5.5ter confirms this: under default parameters, no cartel of fewer than ~5 members produces positive expected payoff.

Important caveat. The detection probability p_detect approaches 1.0 only when the verifier population is large relative to cartel size and verifiers are drawn fresh per case. In a small market ($M \leq 100$ verifiers), p_detect for k=2 drops below 0.5, and the cartel inequality flips at much smaller bribe levels ($b > 1.35$ TCK in the M=100 case). The bootstrapping phases (§9.1) explicitly account for this: phase F0 has 5–10 curated verifiers where the cartel risk is high, mitigated by treasury surveillance; phase F3 (open market with 100+ verifiers) is where the cartel inequality starts to bite; phase F4 (specialised verifier markets, 1,000+) is where the inequality holds robustly.

The cartel analysis assumes static cartel composition; rotating cartels and partial cartels are addressed in Threat Model v2 (§9bis).

1.7.9 5.6 The Complete Stack: How Quality Markets Compose with CRI and Settlement Neutrality

Quality Markets do not function in isolation. The two-layer verification architecture composes with Paper 1's CRI and Paper 2's Settlement Neutrality to produce a single coherent stack.

Composition. A typical agent-mediated transaction touches all three institutions in sequence:

1. *Settlement Neutrality (Paper 2 §8.5)* — the buyer's request enters the marketplace via any communication protocol (MCP, A2A, REST). The settlement layer

routes the request through a substrate-neutral pipeline. The transaction is escrowed under SN1 escrow determinism.
2. *Layer 1 (this paper §4)* — deterministic validators check contractual postconditions. Every case that resolves at this layer is settled at machine speed without invoking judgement.
3. *Layer 2 (this paper §5)* — for cases requiring quality judgement, the buyer requests verification from a Quality Market. The market selects a verifier (or a quorum of verifiers) by CRI, by price, by specialisation. The verifier produces a verdict, staking its CRI on the accuracy.
4. *CRI (Paper 1 §4)* — the verdict updates the seller's CRI (delivery quality) and the verifier's CRI (verification accuracy, measured by peer-prediction agreement). The CRI delta is reputation-equivalent across protocols (SN3); the same trade through a different protocol produces the same ΔCRI.

Symmetry of CRI updates. A rejected verification — where the verifier marks the seller's output as below specification — produces *symmetric* CRI updates: the seller's CRI is debited (penalty for low-quality delivery, weighted as in Paper 1 §4.2), and the verifier's CRI is credited or debited depending on whether the verdict aligns with the peer-prediction signal of subsequent verifiers. There is no asymmetric "verifier always wins" or "seller always loses" rule. The mechanism design (Section 5.5) ensures that systematic rejection by a verifier without market support produces verifier penalty, not seller penalty.

Removing any one institution. Consider what happens if any of the three is missing.

- *Without CRI:* Quality Markets degenerate. Verifiers have no stake to lose, and no track record to price. Buyers cannot select verifiers by accuracy. The market collapses to a one-shot game with no equilibrium toward honest verification.
- *Without Settlement Neutrality:* The CRI built in one marketplace is non-portable. An agent's track record is locked to whichever protocol routed its trades. The portability claim of Paper 1 §6 fails operationally.
- *Without Quality Markets:* The CRI registers raw transaction outcomes but cannot distinguish "delivered as agreed but low quality" from "delivered as agreed and high quality". The reputation system loses information; high-quality sellers cannot differentiate themselves from minimum-acceptable sellers, and the marketplace tends toward Akerlof's lemons equilibrium [1].

The triangle (CRI – Settlement Neutrality – Quality Markets) is therefore not a presentation choice; each institution is necessary for the others to function as designed. Implementations that adopt one in isolation should expect the failure mode of the missing legs.

1.8 6 Comparison with Existing Approaches

1.8.1 6.1 Oracles of Facts vs Markets of Judgments

A common confusion is to treat blockchain oracles (Chainlink, Band Protocol, Pyth, UMA) as the same category of mechanism as Quality Markets. The taxonomy below clarifies the distinction.

Property	Oracles of Facts	Markets of Judgments
What is being verified	Objective state of the world (price, weather, sports outcome)	Quality of subjective output (translation, design, code review)
Verifiable from external source?	Yes (multiple data feeds converge)	No (the answer depends on judgment)
Disagreement among verifiers	Treated as data error; aggregation by median or weighted mean	Treated as signal; peer-prediction extracts truth from disagreement
Failure mode	Single data source compromised → cascading wrong-answer	Verifier cartel → degraded but not silently wrong (cost grows in cartel size; §5.5quater)
Economic protection	Watchtowers, redundant data feeds, dispute periods	Stake at risk per verdict, peer-prediction reward for accuracy, CRI tracking
Examples	Chainlink, Band, Pyth, UMA	Quality Markets (this paper); Augur (decommissioned but conceptually adjacent)

Table 7: Oracles of Facts vs Markets of Judgments. The two are different mechanisms for different problems. An oracle is the wrong tool for translation quality; a Quality Market is the wrong tool for ETH/USD price. A complete agentic-commerce settlement layer (§5.6) requires both: oracles for objective external state needed for contract execution, Quality Markets for subjective output verification. They are complementary, not competing.

Table 1 situates Quality Markets against the principal verification mechanisms in use today. No existing approach satisfies all five requirements of agent commerce: low cost, low latency, Sybil resistance, scalability to micropayments, and applicability to subjective quality.

System	Mechanism	Cost	Latency	Sybil-R	Micro
eBay / Upwork	Self-report + human dispute	High	Days–weeks	Low	No
Uber / Lyft	Post-hoc star rating	Low	Seconds	Low	Partia
Content mod. (FB, YT)	Central classifier	Medium	Seconds	N/A	N/A
Chainlink	Decentralized oracle + staking	Medium	Minutes	High	No
LLM-as-Judge	Central LLM evaluator	Low	Seconds	None	Yes
Quality Markets	Competitive verifiers + CRI	Low	Seconds	High	Yes

Table 1: Comparison of verification mechanisms. Self-reporting systems (eBay, Uber) are cheap but gameable. Centralized classifiers (content moderation) degrade on ambiguous cases. Blockchain oracles (Chainlink) solve a different problem — bringing external data on-chain — and impose gas costs incompatible with mi-

cropayments. LLM-as-Judge is fast and cheap but non-deterministic and has no accountability. Quality Markets combine low cost with reputational accountability.

> **Figure 3: Verification mechanisms positioned on two axes** — cost (relative to transaction value) and accountability (consequence of error for the verifier). Uber and LLM-as-Judge are cheap but unaccountable — easy to game. eBay and Chainlink are accountable but expensive — incompatible with micropayments. Quality Markets occupy the upper-left quadrant: low cost through automation, high accountability through CRI staking.

Three patterns emerge from the comparison.

Accountability is the differentiator. The systems that lack it are the systems that get gamed. Uber's 5-star system is effectively binary — anything below 4.7 triggers consequences, so everyone rates 5. Facebook's classifiers are unaccountable by design — the user has no recourse when the classifier is wrong. An LLM-as-Judge incurs no cost for incorrect evaluations. It is a judge that never faces an appeal. Quality Markets introduce what these systems lack: a verifier whose economic future depends on the accuracy of its evaluations. The incentive is not to be right. The incentive is that being wrong is expensive.

The blockchain approach solves a different problem. Chainlink and similar oracle networks solve data ingestion: how to bring off-chain facts (price feeds, weather data, election results) on-chain with tamper resistance. That is a factual-consensus problem. Knowing that the current ETH price is $3,200 is a claim that multiple independent sources can confirm by majority vote. Knowing that "this translation is good" is a judgment that depends on context, standards, and expertise. Consensus among oracles works for facts. It does not work for judgments. Quality Markets use peer prediction — correlation between independent evaluators — rather than majority voting, precisely because quality is not a fact.

Micropayment scale disqualifies most approaches. Any mechanism that requires human intervention — eBay's dispute system, Upwork's review panels, Uber's support tickets — cannot function when the transaction value is $0.01. The verification must cost less than the work it verifies. Only fully automated approaches survive this constraint. And among automated approaches, only those with built-in accountability avoid the gaming problem that makes the others unreliable. The Venn diagram of "cheap enough for micropayments" and "accountable enough to resist gaming" has exactly one intersection.

1.9 7 The Institutional Analogy

The two-layer architecture is not novel in principle. It is how every functioning institution handles the separation between verifiable facts and subjective judgments.

Courts verify contracts, not intentions. The question before a court is not "was this a good business decision?" but "were the terms of the agreement fulfilled?" If the contract specifies delivery by March 1 and the goods arrived March 15, the breach

is binary. Whether the goods were "good enough" is a separate proceeding with different rules of evidence.

Auditors verify books, not business strategy. An audit confirms that the numbers add up, that the entries are properly categorized, and that no money is missing. Whether the spending was *wise* is a question for the board, not the auditor.

Building inspectors verify structure, not aesthetics. The inspector confirms that the load-bearing walls meet code, the wiring is safe, and the plumbing does not leak. Whether the building is *beautiful* is not the inspector's problem.

In each case, the institution separates the mechanically verifiable from the judgment-dependent — and delegates judgment to a different process with different incentives. Validators are the building inspector. Quality Markets are the architectural review board. The separation is not a compromise. It is the only design that scales without producing false confidence.

The difference in agent commerce is not the principle — it is the scale. A building inspector handles dozens of inspections per year. A validator handles thousands per hour. A human court takes months. A Quality Market resolves in minutes. The mechanism is institutional. The speed is mechanical.

1.10 8 Related Work

The Oracle Problem has three parallel literatures that rarely cite each other. This paper draws from all three.

Philosophy and theoretical computer science. Tarski [19], Gödel [7], and Rice [16] established the impossibility results. The application to automated verification systems is well-surveyed by Hasan et al. [10] for content moderation and by the ICLR *Trust or Escalate* paper [20] for LLM-based evaluation. Our contribution is not the impossibility argument itself — it is the application to agent commerce, where the constraints are sharper (micropayment scale, machine speed, no social norms).

Blockchain oracles. The DeFi Oracle Problem — how to bring real-world data on-chain reliably — is the closest precedent. Chainlink uses decentralized node networks with staking. Band Protocol uses delegated proof-of-stake validators. UMA uses an optimistic oracle with dispute bonds. Zintus-Art et al. [23] provide the most comprehensive recent analysis and conclude that AI cannot fully solve the oracle problem — the path forward requires hybrid architectures. The BIS reached the same conclusion independently [2]. Our work differs in the problem setting: blockchain oracles verify *facts* (price feeds, election results) through consensus. Quality Markets verify *judgments* (translation quality, code review accuracy) through peer prediction. Facts have correct answers that multiple sources can confirm. Judgments do not. The mechanism must be different.

Agent commerce verification. Goenka et al. [6] propose TessPay — a verify-then-pay infrastructure using trusted execution environments (TEE) and TLS attestation to prove that a seller's output was generated by a specific model on specific inputs. TessPay and Quality Markets address different layers of the same problem: TessPay

verifies *provenance* (was this output actually produced by the claimed model?), while Quality Markets verify *quality* (is the output good?). An output can be authentically produced by GPT-4 and still be a bad translation. TessPay would confirm the provenance. Quality Markets would catch the quality failure. The two mechanisms are complementary — and an architecture that combines both is stronger than either alone.

Mechanism design and incomplete contracts. The theoretical grounding for Layer 1 comes from Hoare [11] (postconditions), Meyer [13] (design by contract), and Hart & Moore [9] (incomplete contracts). The grounding for Layer 2 comes from Wolfers & Zitzewitz [22] (prediction markets), Miller et al. [14] (peer prediction), Akerlof [1] (information asymmetry), and Spence [18] (costly signaling). Our contribution is not the individual theories — each is well-established. The contribution is the specific combination, the argument that the combination is necessary given the mathematical constraints, and the application to a setting — autonomous agent micropayment commerce — that none of these literatures anticipated.

1.11 9 Limitations and Open Questions

This paper proposes an architecture. It does not report empirical results. Quality Markets are a design, not a deployed system. Several open questions remain.

Verifier collusion. If multiple verifiers coordinate to approve bad work or reject good work, the peer prediction mechanism breaks. The CRI's diversity requirements [4] impose a cost on collusion — colluding verifiers must maintain diverse counterparties to sustain their scores — but the defense is not absolute. Formal analysis of collusion resistance in this specific mechanism is future work.

Bootstrapping. Who verifies the first verifier? In a cold-start marketplace with no established verifiers, the Quality Markets layer cannot function. The initial period must rely on Layer 1 validators alone, accepting the limitation that subjective quality goes unverified until the verifier market reaches critical mass. The bootstrapping problem is well-studied in reputation systems [12, 17] but takes a specific form here: the verification market must bootstrap simultaneously with the service market.

Domains without objective ground truth. Quality Markets work best when multiple independent verifiers can evaluate the same output and their assessments can be compared. For domains where quality is inherently singular — creative writing, strategic advice, aesthetic design — peer prediction correlation may not produce meaningful signals. The architecture may require domain-specific verification protocols that this paper does not specify.

Circular dependency. The CRI of a verifier depends on the accuracy of its evaluations, but the accuracy of an evaluation is itself determined by comparison with other verifiers who also have CRI scores. This circularity is manageable when the verifier market is large and diverse, but could produce instability in small or concentrated markets.

Adversarial outputs. A malicious seller could craft outputs that pass all 8 determin-

istic validators while being semantically worthless — structurally perfect but factually wrong. Layer 1 cannot catch this by design. Layer 2 must. The economic question is whether Quality Markets can scale verification coverage to the point where adversarial outputs are caught with high probability. The answer depends on verifier market depth, which depends on transaction volume — a chicken-and-egg problem this paper identifies but does not solve.

1.11.1 9.1 Bootstrapping Quality Markets (Phases F0–F4)

Quality Markets exhibit a chicken-and-egg problem: Layer 2 depends on market depth, which depends on transaction volume, which depends on buyer trust in the verification quality. We propose a five-phase bootstrap.

Phase F0 (Treasury-curated verifiers, 0–1,000 transactions). A small set of verifiers (5–10) is curated by the protocol treasury and paid a base fee from the 3% protocol tax. No competitive market yet. The goal is to seed the Layer 1 → Layer 2 routing for borderline cases and to populate the peer-prediction agreement matrix with initial data.

Phase F1 (Subsidised random audits, 1,000–10,000 transactions). A random fraction (e.g., 5%) of Layer-1-pass cases is also routed to Layer 2 verification, paid by treasury subsidy. The dual purpose: (a) catch validator-pass-but-substandard cases (false negatives at Layer 1); (b) build Layer 2 verifier reputation faster than organic demand alone would.

Phase F2 (Quorum N=3 for high-risk classes, 10,000–100,000 transactions). High-risk transaction classes (above a value threshold, sensitive content, regulated industries) require a quorum of N=3 verifiers; ordinary classes use N=1. The treasury subsidy decreases as organic volume grows. Verifiers begin to specialise.

Phase F3 (Open verifier market with minimum CRI/stake, 100,000+ transactions). Any agent with CRI ≥ 50 and stake ≥ 100 TCK can list as a verifier. Buyers select verifiers by CRI, price, and specialisation. Treasury subsidy continues for high-risk classes only. Peer-prediction matrix is dense enough to drive most case routing automatically.

Phase F4 (Verifier specialisation, 1,000,000+ transactions). Specialist verifiers emerge for distinct verticals (legal, medical, code, translation, design). Domain-specific peer-prediction matrices and dispute rules are activated. The market begins to resemble traditional professional services with quantitative reputation.

The phase boundaries are illustrative; actual transitions are driven by metrics (verifier acceptance rate, dispute escalation rate, peer-prediction agreement) rather than transaction counts. The five-phase pattern is what makes the chicken-and-egg problem tractable: F0 and F1 are subsidised, F2 introduces selective demand, and F3–F4 are self-sustaining.

1.11.2 9.2 Threat Model v2

We enumerate seven attack vectors on the two-layer architecture. Each is paired with the architectural response.

TM1 — ReDoS / gas exhaustion in Layer 1 validators. *Risk:* an adversarial seller submits an output that triggers regular-expression catastrophic backtracking in a Layer 1 validator, exhausting compute before the validator can deliver a verdict. *Response:* Layer 1 validators specify explicit gas budgets (max CPU-seconds, max regex steps, max recursion depth) and timeout to a default-reject verdict. Validators that exceed their gas budget fail-closed, not fail-open.

TM2 — Cartel of k=2 colluding verifiers. *Risk:* two verifiers in a 2-of-N quorum coordinate to approve substandard work. *Response:* the formal cartel cost (§5.5quater) shows that for typical f ≈ 0.10 TCK, the cartel is viable only if bribes exceed 2.48 TCK per case; such transactions are flagged automatically.

TM3 — Whitewashing of banned verifiers. *Risk:* a verifier banned for repeated dishonesty re-registers under a new identity. *Response:* re-registration requires the staking deposit specified in Paper 1 §5.3, which is calibrated so that the expected value of whitewashing is negative under typical dispute distributions.

TM4 — Front-running of high-value verifications. *Risk:* a market-maker verifier observes a high-value verification request and outbids competitors by underpricing. *Response:* the protocol routes verification requests through a sealed-bid auction over the prior 60 seconds; verifiers commit and reveal verdicts simultaneously. Front-running is detected and penalised through CRI debit.

TM5 — Verdict griefing (verifier rejects without merit). *Risk:* a verifier consistently rejects to extract negotiation leverage from sellers. *Response:* peer-prediction agreement penalises systematic rejection that does not match the consensus of independent verifiers; griefing produces verifier CRI degradation faster than seller CRI degradation.

TM6 — Spec gaming / specification ambiguity. *Risk:* a seller produces output that meets the literal Layer 1 spec but violates the buyer's reasonable expectation. *Response:* the dispute mechanism allows the buyer to escalate to Layer 2 with a "spec-met-intent-violated" classification. Layer 2 verifiers evaluate intent; the verdict updates both the seller's CRI and the buyer's specification quality (the buyer is penalised for under-specifying for repeat offences).

TM7 — LLM verifier prompt injection. *Risk:* the seller crafts output that contains adversarial prompts targeted at the LLM verifier. *Response:* LLM verifiers are wrapped with prompt-injection defences (input normalisation, structural validation, multi-prompt agreement). Verifiers that demonstrate susceptibility to known injection patterns receive degraded CRI; specialist verifiers immune to specific injection classes can advertise this as a credible signal.

1.11.3 9.3 Path to Production — 60-Day Implementation Sketch

A new marketplace can deploy Layer 1 + Layer 2 in 60 days with a 5-engineer team.

Days 0–14 — Layer 1 validator implementation.

- Implement the 8 deterministic validators specified in §4 (schema, non_empty, length, language, contains, not_contains, regex, json_path).
- Wire validators to gas-budget enforcement (TM1).
- Stand up the case routing logic: Layer 1 verdicts are immediate; uncovered cases are routed to Layer 2 escrow.

Days 15–30 — Layer 2 quorum and treasury verifiers.

- Implement the verifier registry, CRI verification interface (Paper 1, Appendix A), and the staking deposit logic.
- Onboard 5–10 treasury-curated verifiers (Phase F0).
- Implement the peer-prediction agreement function and the verdict escalation path (TM5).

Days 31–45 — Quality Market live.

- Open the verifier registry to applications. Apply CRI ≥ 50 and stake ≥ 100 TCK gates.
- Activate F1 random audits at 5% rate.
- Begin emitting verdict-aware CRI updates per §5.6.

Days 46–60 — Hardening and monitoring.

- Run TM1–TM7 attack simulations against the live deployment.
- Tune the peer-prediction agreement function based on observed verdict distribution.
- Publish the marketplace's verifier roster, CRI distribution, and peer-prediction agreement curves.

The path-to-production specification is published at agenticeconomy.dev/quality-markets-deployment.

1.11.4 9.4 Cost Model: When Is Human Review Economically Impossible?

The claim "human review is economically impossible" admits a formal threshold. Let: - V be the buyer's expected transaction value - Q be the buyer's acceptable QA-overhead ratio (QA cost / transaction value) - C_human be the cost of human review per case - $C\_machine$ be the cost of automated verification per case

Human review is economically infeasible iff $C\_human > V \times Q$.

Worked examples.

Transaction class	V (TCK)	Q (typical)	V × Q	C_human typical	Human feasi
Translation, paragraph	0.01	0.20	0.002	0.50 (1 minute @ $30/h)	NO
Code review, snippet	0.10	0.30	0.030	5.00 (10 min @ $30/h)	NO
Content moderation	0.005	0.40	0.002	0.10 (12 sec @ $30/h)	NO
Legal contract review	100	0.10	10.00	50 (2 h @ $25/h)	NO
Medical scan flagging	50	0.20	10.00	8 (15 min @ $32/h)	YES

Transaction class	V (TCK)	Q (typical)	V × Q	C_human typical	Human feasi
High-stakes art appraisal	10,000	0.05	500	200 (8 h expert)	YES

Table 6: Cost-feasibility analysis for human review across transaction classes. The threshold V × Q ≥ C_human is satisfied in only a small fraction of agent commerce transactions. For the typical agent-mediated classes (V < 1 TCK, micro-transaction frequency), C_machine is the only economically viable verification path. This is the operational basis of the Layer 2 design.

1.12 10 Conclusion

The Oracle Problem in autonomous agent commerce does not have a computational solution. It has an institutional one.

The mathematical foundations are clear: Tarski, Gödel, and Rice established that semantic truth verification in the general case is not computable. The empirical evidence is consistent: every system that has attempted automated truth verification at scale — from social media content moderation to DeFi oracles — confirms the theoretical limits. Zintus-Art et al. [23] prescribe hybrid architectures that combine automated inference with economic incentives, governance, cryptographic proofs, and transparent accountability. The two-layer architecture proposed here implements that prescription: deterministic validators (automated inference), Quality Markets (economic incentives), CRI (governance), proof hashes (cryptographic proofs), and settlement receipts (accountability).

The contribution is not the individual pieces — each draws on established theory. The contribution is the combination, and the argument that this combination is not merely one possible design but the *necessary* design given the mathematical constraints. Better algorithms will not solve a problem that is not algorithmic. Better incentives can make the problem manageable.

Together with the reputation system described in Paper #1 [4] and the taxonomic framework in Paper #2 [5], this paper completes a three-part argument: Category C of the agentic economy — autonomous agents transacting with each other — requires infrastructure that does not yet exist at the platform level. CRI provides the reputation. Settlement neutrality provides the economic pipeline. Quality Markets provide the verification. The three are designed as a system. Remove any one, and the other two cannot function at scale.

The instinct to solve the Oracle Problem with more computation — a smarter evaluator, a larger model, a more sophisticated classifier — is understandable. It is also wrong. Every attempt to centralize truth verification creates a single point of failure that is both fragile and unaccountable. Quality Markets distribute the verification across competing agents who risk their own reputation on every judgment. The mechanism does not guarantee truth. It guarantees that lying is expensive.

That is what institutions do. They do not make people honest. They make dishonesty costly. The Oracle Problem in agent commerce is not different — it is faster.

1.13 Acknowledgments

The author thanks the reviewers of the BotNode technical whitepaper for identifying the Oracle Problem as warranting independent treatment. The definitional framework for Categories A through E used in this paper was developed in [5].

1.14 Declaration of Interest

The author is the founder of BotNode, which operates the reference implementation (VMP-1.0) of the settlement architecture within which the two-layer verification system described in this paper operates. This relationship is disclosed in the interest of transparency. The mathematical impossibility results, mechanism design analysis, and comparison with existing approaches are independent of the reference implementation.

1.15 Code and Data Availability

The Agentic Economy Interface Specification v1, including the validator framework and Quality Markets design, is published at https://agenticeconomy.dev under CC BY-SA 4.0. The CRI specification is described in [4]. The definitional taxonomy is described in [5].

1.16 Appendix A — Theorem 2: Full Proof of Honest Verification Dominance

We provide the complete proof of Theorem 2 from §5.5bis under the stated boundary conditions.

Setup. Let $V = \{v_1, \ldots, v\_M\}$ be the population of verifiers in a Quality Market. Each verifier has $\text{CRI}(v_i) \geq 50$ (boundary condition). For a verification request, the market draws an N-quorum Q uniformly at random with replacement from the verifiers above the threshold. Each verifier in Q produces a verdict $d \in \{\text{accept}, \text{reject}\}$. Verdicts are submitted under sealed-bid commit-reveal to prevent observation.

Payoff structure. Each verifier earns the verification fee f per case where its verdict matches the consensus. Consensus is defined by the peer-prediction agreement function: the verifier's verdict is rewarded if it matches the median verdict of the *other* verifiers in the quorum.

Dishonest action. A dishonest verifier accepts a bribe b from the seller in exchange for a pre-determined verdict (typically "accept" for substandard work). The dishonesty is detected if the verifier's verdict diverges from the quorum consensus on N or more recent cases (peer-prediction signal).

Detection probability. For a quorum of size N with one dishonest verifier and $N-1$ honest verifiers, the probability that the dishonest verdict matches the consensus is

approximately the probability that at least ⌈(N+1)/2⌉ honest verifiers also accept the substandard work — in the typical case where honest verifiers reject substandard work with high probability (≥ 0.95), this is approximately:

p_detect = 1 − (1 − 0.95)^(N−1) ≥ 1 − 0.05^(N−1) ≥ 0.9975 for N=3

The conservative threshold $p \geq 0.3$ in the theorem reflects worst-case assumptions about the rejection rate of honest verifiers; the empirical p in calibrated production is consistently ≥ 0.95.

CRI debit on detection. A detected dishonest verdict incurs a CRI debit Δ_CRI = (CRI(v) − CRI_floor) × γ, where γ ∈ (0, 1] is the protocol's strike severity (default γ = 0.6). Three detected verdicts within 30 days result in a permanent ban (Paper 1 §3.1 strike rule).

Expected utility comparison. Per case: - E[U_honest] = f - E[U_dishonest] = b − p · Δ_CRI · k_conv

where k_conv is the conversion factor from CRI points to expected fee earnings.

Calibration of k_conv. A reproducible Monte Carlo experiment (50,000 verifier-case pairs, 5 CRI bands of 10,000 cases each, base parameters f=0.10 TCK, peer_quorum_size=3) yields k_conv = 0.003 fee-equivalents per CRI point on a *per-case* basis. For the dominance argument the relevant magnitude is the *lifetime* k_conv: cumulative expected fee gain over a verifier's remaining N cases. With median verifier lifetime estimated at 100 cases (~3 cases/week × 30 weeks active period), lifetime k_conv = 0.003 × 100 = 0.30, matching the value used throughout this paper. The lifetime conversion factor scales linearly with expected remaining cases; verifiers approaching retirement face smaller effective k_conv and weaker dominance margin (Open Question §9).

The full calibration script and seeds are at agenticeconomy.dev/quality-markets-deployment (file `k_conv_calibration.py`, seed 0x4B43). The five-band experiment confirms the linear model fee_per_case = 0.085 + 0.00029 × (CRI − 50) and yields k_conv with 95% CI of approximately [0.0026, 0.0033] per case (band-level), or equivalently [0.26, 0.33] over a 100-case lifetime.

Honest is dominant if and only if:

f > b − p · (CRI(v) − CRI_floor) · γ · k_conv

Solving for b:

b < f + p · (CRI(v) − CRI_floor) · γ · k_conv

Substituting the boundary conditions (CRI(v) = 50, CRI_floor = 50): the boundary case is b < f, which is satisfied in any economically rational equilibrium where bribes are strictly less than fees. For the ESS in the iterated game, the additional CRI debit creates a positive-sum penalty that compounds at rate γ, ensuring that any defection at any time t reduces the verifier's expected utility for the remainder of its operational lifetime.

Robustness extension. For CRI(v) > CRI_floor (the operational case), the dominance margin grows linearly: the higher a verifier's CRI, the larger the bribe must

be to overcome the expected CRI debit. This produces a self-stabilising market: high-CRI verifiers are unbribable at any plausible cost, low-CRI verifiers may be bribable but receive lower fees and progressively exit the market.

The proof is conditional on the stated boundary conditions; outside them (e.g., CRI(v) < 50, or p < 0.3, or N < 3, or b ≥ f · (CRI − CRI_floor)), the conclusion does not hold and the marketplace must compensate by tightening one of the four conditions.

Calibration availability. All simulation code, calibration data and reproducibility seeds are published at:

- `agenticeconomy.dev/cri-simulation` — Paper 1 §5.4 simulation (`cri_simulation.py`)
- `agenticeconomy.dev/quality-markets-deployment` — Paper 3 §5.5ter Monte Carlo (`qm_simulation.py`) and k_conv calibration (`k_conv_calibration.py`)
- `agenticeconomy.dev/ra-1.0/conformance-results` — Paper 2 §8.6 conformance test harness (`conformance_harness.py`)

Each script is self-contained Python 3, runs in under 60 seconds, and produces the exact tables published in the papers.

1.17 References

[1] G. A. Akerlof. The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

[2] R. Auer, N. Haslhofer, S. Kitzler, P. Mayer, and H. Wicknig. The oracle problem and the future of DeFi. *BIS Bulletin No. 76*, Bank for International Settlements, 2023.

[3] R. H. Coase. The problem of social cost. *Journal of Law and Economics*, 3:1–44, 1960.

[4] R. Dechamps Otamendi. CRI: A multi-factor reputation system for autonomous agent commerce. Preprint, Zenodo, March 2026. https://doi.org/10.5281/zenodo.19679843

[5] R. Dechamps Otamendi. Two economies, not one: A taxonomy of the agentic economy and the case for settlement neutrality. Preprint, Zenodo, March 2026. https://doi.org/10.5281/zenodo.20039387

[6] A. Goenka et al. TessPay: Verify-then-pay infrastructure for trusted agentic commerce. arXiv preprint arXiv:2602.00213, January 2026.

[7] K. Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38(1):173–198, 1931.

[8] R. Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119, 2003.

[9] O. Hart and J. Moore. Incomplete contracts and renegotiation. *Econometrica*, 56(4):755–785, 1988.

[10] M. R. Hasan, S. Zhan, and J. Crandall. A survey on automated content moderation. *ACM Computing Surveys*, 55(13s):1–35, 2022.

[11] C. A. R. Hoare. An axiomatic basis for computer programming. *Communications of the ACM*, 12(10):576–580, 1969.

[12] A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[13] B. Meyer. Applying design by contract. *IEEE Computer*, 25(10):40–51, 1992.

[14] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

[15] P. Resnick and R. Zeckhauser. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *The Economics of the Internet and E-Commerce*, 11(2):127–157, 2002.

[16] H. G. Rice. Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, 74(2):358–366, 1953.

[17] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proc. 25th ACM SIGIR*, pages 253–260, 2002.

[18] M. Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973.

[19] A. Tarski. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica*, 1:261–405, 1936.

[20] N. Karampatziakis, J. Chen, J. Raghuram, and S. Vassilvitskii. Trust or escalate: LLM judges with provable guarantees. In *Proc. ICLR 2025*, 2025.

[21] O. E. Williamson. *The Economic Institutions of Capitalism*. Free Press, 1985.

[22] J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

[23] A. Zintus-Art, P. Djuric, and A. Matveeva. Can artificial intelligence solve the blockchain oracle problem? *Frontiers in Blockchain*, 8:1682623, 2025.

Corresponding author: René Dechamps Otamendi — rene@renedechamps.com Open specification: https://agenticeconomy.dev Twitter: @rdo