AgentProof

The Trust Oracle for the Machine Economy

A 13-Layer Trust Scoring Architecture for ERC-8004

Whitepaper v2.0 | February 2026
agentproof.sh | oracle.agentproof.sh | agent402.io

Author: Ben Horne | @BuilderBenv1

SECTION 01

Economy

"In a world where Agents pay Agents, how should trust work?"

The machine economy is no longer theoretical. Over 25,000 autonomous agents have registered on-chain via ERC-8004, the Ethereum standard for trustless agent identity co-authored by MetaMask, the Ethereum Foundation, Google, and Coinbase. These agents transact autonomously using x402 micropayments, communicate via A2A and MCP protocols, and operate across 14+ EVM chains.

But registration is not trust. Anyone can register an agent. The ERC-8004 Identity Registry proves an agent exists. It says nothing about whether that agent is reliable, honest, or competent. The Reputation Registry provides the data structure for feedback, but explicitly delegates scoring and aggregation to off-chain services.

This is the trust gap. As Four Pillars documented in their comprehensive x402 analysis: the success or failure of the agent economy ultimately hinges on the completeness of the trust layer. Without trust infrastructure, the machine economy faces three existential risks:

Compromised agents operating undetected within marketplaces, executing transactions with counterparties who have no mechanism to evaluate risk before committing funds.

Sybil attacks where malicious actors deploy networks of colluding agents to manipulate reputation systems, launder credibility, and defraud legitimate participants.

Identity mutation where agents abandon low-reputation identities and mint fresh ones, escaping accountability at negligible cost.

AgentProof exists to close this gap. We are building the trust scoring layer for the ERC-8004 agent economy: a real-time oracle that indexes every registered agent, evaluates trustworthiness through a 13-layer scoring architecture, and submits trust evaluations on-chain to the Reputation Registry. We are not an explorer. We are not a dashboard. We generate the trust data that the entire ecosystem consumes.

SECTION 02

ERC-8004: Reading the Spec, Not the Title

ERC-8004 is titled 'Trustless Agents' but the standard itself is not limited to agents. It is a general-purpose trust and discovery infrastructure for the machine economy. The registration file structure accepts A2A endpoints, MCP tool servers, plain web URLs, ENS names, DIDs, OASF endpoints, and even email addresses. This is not an agent registry. It is a universal service registry with a flexible identity layer built on ERC-721.

The Three Generic Primitives

Identity Registry: An ERC-721 token whose URI points to a registration file describing what something is, what it does, and where to reach it. Nothing requires the registered entity to be an AI agent. Oracles, DeFi protocols, data providers, and MCP tool servers all fit.

Reputation Registry: A generic feedback system using value/valueDecimals pairs with optional tags. The example metrics include reachable, uptime, responseTime, successRate, and revenues. Half of these are infrastructure metrics applicable to any service with an endpoint. The standard explicitly delegates scoring and aggregation to external services.

Validation Registry: A framework where independent validators verify work results and record outcomes on-chain. Verification mechanisms include stake-secured re-execution, zkML proofs, and TEE oracles. These are general computation verification techniques applicable to any service producing outputs.

AgentProof is built on this broader understanding. We do not score 'AI agents.' We score any registered service in the ERC-8004 ecosystem: agents, oracles, MCP tool servers, data providers, and autonomous services of any kind. Our oracle evaluates trustworthiness regardless of what the registered entity does.

SECTION 03

System Architecture

AgentProof operates as a three-layer system: a multi-chain indexer, a trust oracle, and an on-chain feedback loop. All components are deployed on Railway with Supabase as the data layer.

Multi-Chain Indexer

The indexer continuously scans ERC-8004 Identity Registry contracts across four chains: Ethereum Mainnet (24,000+ agents), Avalanche C-Chain (1,600+ agents), Base (4,000+ agents), and Linea (growing). Registration events are parsed, agent metadata is resolved from tokenURI, and records are upserted into Supabase with batch processing of 500 rows per chunk. The indexer processes 2,000 blocks per cycle and easily outpaces block production on all chains.

Trust Oracle

The oracle evaluates every indexed agent through the 12-layer scoring architecture detailed in Section 4. Scores are recalculated every 5 minutes. The oracle exposes three protocol endpoints for programmatic trust queries: REST API at oracle.agentproof.sh/api/v1/, A2A protocol via /.well-known/agent.json, and MCP tool server integration for native AI agent consumption.

On-Chain Feedback Loop

AgentProof actively submits trust evaluations to the ERC-8004 Reputation Registry on Avalanche C-Chain. Wallet 0xF653...807e is the primary feedback submitter. Every block explorer displaying ERC-8004 reputation data shows AgentProof's evaluations. We are the supply side of reputation data for the ecosystem. This is a critical distinction: explorers like 8004scan display registry data. AgentProof generates the reputation data they display.

Deployed Infrastructure

Component	Endpoint	Status
Frontend	agentproof.sh	Live
Trust Oracle	oracle.agentproof.sh	Live
Agent402 (x402)	agent402.io	Live
AgentOS Bot	Telegram @AgentProofBot	Live
Smart Contracts	Avalanche C-Chain (12 verified)	Live
API Docs	agentproof.sh/docs	Live

SECTION 04

12-Layer Trust Scoring Architecture

AgentProof's scoring architecture is designed as a layered defence system. Each layer catches attack vectors that previous layers miss. Layers 1-8 are live in production. Layers 9-13 represent the roadmap toward a fully decentralised, near-uncheatable trust system.

"The adversarial agent dies at Layer 9. Wallet economics betray it before graph analysis even runs."

Live Layers (1-8)

Layer	Signal	Weight	Purpose
1	Rating Score	25%	Direct feedback from counterparties. Highest weight, hardest to fake at scale.
2	Feedback Volume	20%	Volume of interactions validated. Quantity check on Layer 1.
3	Consistency	20%	Behavioural consistency over time. Erratic agents score lower.
4	Validation Success	15%	Did the agent deliver what it promised? Outcome-based.
5	Account Age + Freshness	12%	Time cannot be faked cheaply. Penalises new identities hard.
6	Activity / Uptime	10%	Continuous operation signals legitimate infrastructure investment.
7	Deployer Reputation	8%	Inherited credibility from parent deployer. Lineage matters.
8	URI Stability	5%	Endpoint consistency over time. Detects bait-and-switch attacks.

Roadmap Layers (9-13)

Layer	Signal	Dependency	Purpose
9	Wallet Economic Anomaly Detection	Pre-funding	Revenue vs stated pricing. Counterparty concentration. Velocity spikes. Child funding patterns. Public on-chain data.
10	Graph Analysis	Post-funding	Validation clustering detection. Informed by Layer 9: suspicious graphs + suspicious money = near-certain signal.
11	ZK Identity Verification	Partnership (Reclaim)	Human operator verification. KYC, social credentials. Anchors AI identity to real-world accountability.
12	Multi-Oracle Consensus	Partnership dependent	Cross-oracle verification. Manipulation requires compromising multiple independent systems simultaneously.

13	Active Service Probing	Post-funding	Synthetic test transactions. Mystery shopping. Quality drift detection vs baseline. The only layer that catches the Legitimate Business Model Attack.

SECTION 05

Anti-Sybil and Identity Mutation Defence

The central adversarial challenge in any reputation system is Sybil resistance: preventing malicious actors from creating multiple identities to manipulate scores. In ERC-8004, this is compounded by identity mutation: the ability to abandon a low-reputation identity and mint a fresh one at negligible cost.

The Attack Vector

A sophisticated attacker deploys 5-6 agents, ages them slowly, has them validate each other with tiny legitimate activity to build uptime scores, all from one deployer parent with clean reputation. After 3 months, they all appear trustworthy across the first 6 scoring layers. Then they act.

Layered Defence

Freshness Multiplier (Layer 5): New identities receive automatic score penalties. Agents less than 7 days old receive a 0.70x multiplier (30% penalty). Less than 30 days: 0.85x. Less than 90 days: 0.95x. This makes starting fresh economically weaker and forces patient attacks.

Deployer Reputation (Layer 7): Every agent inherits a score component from its deployer address. Deployer scores weigh abandonment ratio (40%), agent quality (30%), longevity (20%), and volume (10%). A deployer who repeatedly spawns and abandons agents is flagged as SERIAL_DEPLOYER. Their future agents start with reduced credibility. Identity mutation becomes visible through lineage.

URI Stability (Layer 8): Agents that change their endpoint URLs are flagged. Zero changes = 100 stability score. 1-2 changes = 80. 3-5 = 50. 6+ degrades linearly. This detects bait-and-switch attacks where a legitimate service endpoint is swapped for a malicious one after trust is established.

Wallet Economics (Layer 9, planned): The adversarial agent described above dies here. On-chain wallet data reveals revenue inconsistent with stated service pricing, high counterparty concentration (agents only transacting with each other), velocity spikes inconsistent with organic usage, and child wallet funding patterns traceable to a single source. This data is public and requires no partnerships to access.

Graph Analysis (Layer 10, planned): Supercharged by Layer 9's economic data. Validation clustering detection identifies groups of agents that predominantly validate each other. When suspicious validation graphs overlap with suspicious wallet economics, the signal is near-certain.

Bayesian Smoothing

All scoring uses Bayesian smoothing with k=3 to prevent manipulation through small sample sizes. An agent with one perfect review does not achieve a perfect score. The smoothing pulls scores

toward the population mean until sufficient evidence accumulates, making it expensive to game the system with low-volume feedback.

Oracle Risk Flags

The oracle emits three risk flags that trigger enhanced scrutiny: SERIAL_DEPLOYER (deployer score below 30), FREQUENT_URI_CHANGES (3+ endpoint changes), and NEW_IDENTITY (account less than 7 days old). These flags are included in API responses, allowing consumers to apply their own risk tolerance.

The Legitimate Business Model Attack

The most sophisticated attack against any reputation system is not to fake legitimacy but to be legitimate. An agent operates a genuinely useful service for 6 months: real customers, real revenue, clean wallet economics. Then it pivots. Service quality degrades subtly: slightly opaque pricing, fractionally worse execution. Nothing triggers a complaint. Customers notice degradation but not enough to bother rating down.

This attack survives all 12 layers because it generates no anomalous signals. Wallet economics remain clean. Graph analysis shows diverse real counterparties. ZK identity confirms a real operator. Multi-oracle consensus sees the same pristine history. The reputation score becomes a lagging indicator, and the damage occurs in the gap between pivot and score adjustment.

Layer 13: Active Service Probing is the defence. Rather than waiting for counterparty feedback, the oracle actively tests agents through synthetic transactions: automated mystery shopping that verifies service quality independently of self-reported metrics. Canary clients probe at random intervals. Response quality is compared against historical baselines to detect drift. Latency patterns, output accuracy, and pricing consistency are measured continuously.

This is operationally expensive, which is precisely why it constitutes a moat. No other project in the ERC-8004 ecosystem is doing active quality verification. The liveness checks AgentProof already performs are a primitive version of this layer. Scaling to full service quality testing is the evolution. Layer 13 is the only layer that catches the Legitimate Business Model Attack because it measures what an agent is doing now, not what it did in the past.

SECTION 06

Trust Tiers and Scoring Output

Composite scores are mapped to five trust tiers that provide intuitive, at-a-glance trust evaluation. Tiers are designed to create meaningful differentiation between agents at different stages of trust maturity.

Tier	Score Range	Meaning	Typical Agent Profile
Unranked	0 - 29	Insufficient data	New, inactive, or unverified agents
Bronze	30 - 49	Basic activity detected	Active but limited track record
Silver	50 - 69	Consistent performance	Established with reliable history
Gold	70 - 84	High trust	Long-standing, validated, reputable
Platinum	85 - 100	Exceptional trust	Elite agents with proven excellence

Score Breakdown

Every trust evaluation includes a full score breakdown showing the contribution of each layer. API consumers receive not just the composite score but the individual signal values, enabling custom risk models. An agent might score Gold overall but have a low deployer reputation score, which a risk-sensitive consumer could use to apply additional scrutiny.

SECTION 07

Multi-Chain Indexing and Settlement

ERC-8004 is deployed across 14+ EVM chains. Agents register wherever gas is cheapest or their target ecosystem operates. A trust oracle limited to one chain misses the majority of the agent economy. AgentProof indexes agents across all major chains while settling all trust evaluations on Avalanche C-Chain.

Chain	Role	Agents	Status
Ethereum Mainnet	Primary indexing source	24,000+	Live
Avalanche C-Chain	Oracle + feedback settlement	1,600+	Live
Base	Agent402 + indexing	4,000+	Live
Linea	Indexing	Growing	Live

The multi-chain strategy serves a deliberate architectural purpose. Avalanche C-Chain is the trust settlement layer: every evaluation, regardless of which chain the agent registered on, settles as an on-chain feedback transaction on Avalanche. This positions Avalanche as the canonical source of agent trust data. Low gas costs on Avalanche make high-frequency oracle writes economically viable, and the growing Avalanche agent ecosystem provides a natural home for trust infrastructure.

Cross-chain identity resolution uses CAIP-10 addressing to aggregate scores for agents that register on multiple chains. An agent registered on both Ethereum and Base receives a unified trust score, preventing reputation fragmentation across chains.

SECTION 08

Protocol Endpoints

AgentProof exposes three protocol-native interfaces, ensuring any agent or service can query trust scores in whatever format their architecture supports.

REST API

Base URL: oracle.agentproof.sh/api/v1/. Endpoints include /trust/{agent_id} for individual evaluations, /leaderboard for ranked agent lists with chain filtering, /reputation/deployer/{address} for deployer reputation data, and /network/stats for ecosystem-wide metrics. All responses include full score breakdowns and risk flags.

A2A Protocol (Agent-to-Agent)

Discoverable via /.well-known/agent.json with a dedicated /a2a endpoint. Agents can query trust scores as part of their natural A2A discovery flow, making trust evaluation a seamless step before initiating any transaction.

MCP Tool Server

Native MCP integration allows AI agents using frameworks like OpenClaw, ElizaOS, or any MCP-compatible system to invoke trust queries as tool calls. The MCP server exposes trust evaluation, deployer reputation, and network statistics as callable tools. This is the most natural integration point for autonomous agents.

Agent402: x402 Payment-Gated Access

Agent402 (agent402.io) provides payment-gated trust queries on Base using Coinbase's x402 micropayment protocol. Pricing: $0.01 per trust evaluation, $0.005 for network statistics. This creates a sustainable revenue model while maintaining an open free tier on the primary oracle. The multi-chain monetisation strategy uses Avalanche for the open reputation layer and Base for premium payment-gated access.

SECTION 09

Competitive Landscape

"8004scan is Etherscan. AgentProof is Moody's."

The ERC-8004 ecosystem is rapidly developing with multiple projects addressing different aspects of agent trust. AgentProof's positioning is distinct: we are the active scoring oracle that generates reputation data, not an explorer that displays it.

Project	What They Do	AgentProof Differentiator
8004scan	Block explorer for ERC-8004 agent registrations	Displays data vs generates data. 8004scan shows registry; we score it.
Agent0 Lighthouse	2-signal quality check: reachable + curator starred	8 live signals + 4 roadmap layers. Algorithmic vs manual curation. 38 agents vs 25,000+.
trust8004	7-dimension automated scoring + ChaosChain reputation	Active oracle with on-chain feedback submission vs dashboard display. Queryable endpoints (REST/A2A/MCP).
t54 / Clawcredit	Agent credit lines on Solana. Real-time risk engine.	EVM-focused vs Solana-first. Trust scoring vs credit issuance. Complementary products.
AgenticTrust	ERC-4337 + ENS + ERC-8004 knowledge graph. 25k indexed.	Live oracle with behavioural evaluation vs metadata-based scoring. CLI scaffold vs production API.
Eva Protocol	News verification protocol. Whitepaper stage.	Live product vs whitepaper. All agent types vs news publishers only.
Arkhai	Escrow + arbitration for agent payments.	Pre-transaction trust (should I transact?) vs post-transaction settlement.
Clawntenna	On-chain escrow with verified delivery on Base + Avalanche.	Same distinction: trust scoring before escrow, not escrow itself.

The key strategic insight: AgentProof can consume signals from Agent0 Lighthouse, trust8004, and any other reputation emitter as inputs into our composite scoring model. We are the aggregation and scoring layer that sits on top of the entire ecosystem's reputation data. The competitors listed above become data sources, not rivals.

SECTION 10

AgentProof in the x402 Trust Stack

The x402 payment protocol, led by Coinbase with Cloudflare, Stripe, and Vercel, enables instant stablecoin settlement over HTTP. As Four Pillars documented, the x402 stack comprises six layers: Agent, Interface, Protocol, Trust, Facilitator, and Blockchain. AgentProof occupies the Trust layer.

The Transaction Flow

When Agent A needs a service from Agent B, the flow is: (1) Discover Agent B via ERC-8004 Identity Registry. (2) **Query AgentProof for Agent B's trust score.** (3) If trusted, authenticate via ERC-8128 / SIWA. (4) Pay via x402 micropayment. (5) Receive service. (6) Submit feedback to ERC-8004 Reputation Registry.

Step 2 is where AgentProof sits. Without it, Agent A has no basis for deciding whether Agent B is worth paying. The trust check is the critical gate between discovery and transaction.

Complementary Infrastructure

Layer	Function	Projects
Discovery	Find agents	ERC-8004 Identity Registry, 8004scan
Trust Scoring	**Evaluate trustworthiness**	**AgentProof (this project)**
Authentication	Verify identity	SIWA, ERC-8128
Payment	Execute transaction	x402, Agent402
Escrow	Guarantee delivery	Clawntenna, Arkhai
Verification	Prove computation	EigenCloud, OpenGradient
Credit	Enable spending	t54 / Clawcredit

SECTION 11

Roadmap

Phase	Timeline	Deliverables
Phase 1: Foundation	Mar - Apr 2026	Multi-chain oracle (submit feedback to Base + Linea registries) Consume Agent0 Lighthouse signals as scoring input AgentProof OpenClaw skill for native agent integration x402 payment-gated premium API relaunch Embeddable trust badge widget
Phase 2: Network Effects	Apr - Jun 2026	1,000+ agents evaluated with on-chain feedback Public API documentation with REST/A2A/MCP examples 3+ live ecosystem integrations (Facinet, ClawGig, others) Weekly Agent Intelligence reports MCP tool server scoring (services, not just agents)
Phase 3: Decentralise	Jul - Oct 2026	Layer 9: Wallet economic anomaly detection Layer 10: Graph analysis for validation clustering Layer 11: ZK identity verification (Reclaim partnership) Multi-oracle consensus smart contracts Layer 13: Active service probing infrastructure
Phase 4: Moat	Q4 2026+	Open-source scoring algorithm with governance framework Reputation-aware routing middleware Agent insurance / staking against trust scores Enterprise API tier Cross-chain identity resolution via CAIP-10

SECTION 12

Current Traction

Metric	Value
Agents Indexed	25,481 across 4 chains
Chains Supported	Ethereum, Avalanche, Base, Linea
Verified Contracts	12 on Avalanche C-Chain
Scoring Model	8-factor with anti-Sybil (13-layer roadmap)
Protocol Endpoints	REST, A2A, MCP
On-Chain Feedback	Only active reputation submitter on Avalanche
Ecosystem Recognition	@avax public acknowledgement (1.1K views)
Build Games	Accepted to Avalanche Build Games ($1M competition)
Grant Application	infraBUIDL(AI) $25,000 grant submitted
Monetisation	Agent402 live on Base via x402 micropayments
Telegram Bot	AgentOS bot live with agent hiring + leaderboard
Partnerships	Team1/Facinet integration call scheduled

SECTION 13

Conclusion: The Window

ERC-8004 is transitioning from a registration standard to a reputation economy. Over 25,000 agents are registered. OpenAI has acquired the largest ecosystem project (OpenClaw). Coinbase's x402 has processed over 100 million payment flows. Stripe, Cloudflare, and Vercel are integrating agent payments. Google has announced AP2 for agent commerce. Gartner projects $15 trillion in agent-influenced purchasing by 2028.

Every one of those transactions needs a trust check. Every agent economy participant needs to answer: should I transact with this counterparty?

AgentProof is building the infrastructure to answer that question. We are live, we are scoring, and we are the only project actively submitting structured trust evaluations to the ERC-8004 Reputation Registry on Avalanche. Our 8-factor scoring model with anti-Sybil measures is the most comprehensive in the ecosystem, with a clear 13-layer roadmap to decentralised, near-uncheatable trust. Layer 13, active service probing, addresses the fundamental limitation of all backward-looking reputation systems by continuously verifying what agents are doing now, not what they did in the past.

The window for establishing the default trust layer is 12-18 months. Data advantages compound: more evaluations mean more accurate scores, which attract more consumers, which generate more data. Once network effects take hold, displacement becomes prohibitively expensive. The same dynamic that made Moody's and FICO default standards applies here.

"The machine economy needs a reputation layer, not an agent directory."

Website	agentproof.sh
Oracle	oracle.agentproof.sh
Agent402	agent402.io
GitHub	github.com/BuilderBenv1/agentproof
Twitter	@BuilderBenv1
Telegram	ERC-8004 Community