MANCHESTER 1824
The University of Manchester

Can AI Agents Be Trusted to Pay?

Security of the Agentic-Internet Stack

Zhipeng Wang

Lecturer (Assistant Professor) in Cyber Security
Department of Computer Science
The University of Manchester

- *Five Attacks on x402 Agentic Payment Protocol* arXiv:2605.11781

- *Can Trustless Agents Be Trusted? An Empirical Study of the ERC-8004 Ecosystem* arXiv:2606.26028

Virtual Agent Economies

Nenad Tomašev[1], **Matija Franklin**[1], **Joel Z. Leibo**[1], **Julian Jacobs**[1], **William A. Cunningham**[1, 2], **Iason Gabriel**[1] **and Simon Osindero**[1]
[1]Google DeepMind, [2]University of Toronto

The rapid adoption of autonomous AI agents is giving rise to a new economic layer where agents transact and coordinate at scales and speeds beyond direct human oversight. We propose the "sandbox economy" as a framework for analyzing this emergent system, characterizing it along two key dimensions: its origins (emergent vs. intentional) and its degree of separateness from the established human economy (permeable vs. impermeable). Our current trajectory points toward a spontaneous emergence of a vast and highly permeable AI agent economy, presenting us with opportunities for an unprecedented degree of coordination as well as significant challenges, including systemic economic risk and exacerbated inequality. Here we discuss a number of possible design choices that may lead to safely steerable AI agent markets. In particular, we consider auction mechanisms for fair resource allocation and preference resolution, the design of AI "mission economies" to coordinate around achieving collective goals, and socio-technical infrastructure needed to ensure trust, safety, and accountability. By doing this, we argue for the proactive design of steerable agent markets to ensure the coming technological shift aligns with humanity's long-term collective flourishing.

Keywords: AI, economy, multi-agent, blockchain, ethics

Virtual Agent Economies, arXiv:2509.10147

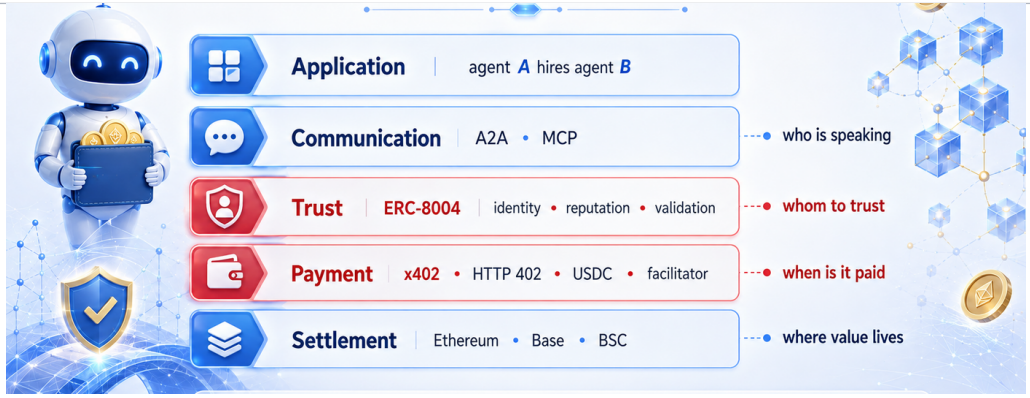
From Assistants to Economic Actors

An AI system used to *suggest*, and a human acted. Now the agent **finds** the seller, **judges** it, and **pays**, unsupervised.

Tomašev et al., *Virtual Agent Economies*, arXiv:2509.10147: assistants bidding for one hotel booking · a research agent paying for data · robots compensating each other.

Thousands of decisions per second, below the threshold of human oversight. So: **what is holding this together?**

The Agentic-Internet Stack



Communication is comparatively well studied. The **two red layers** carry the money and the trust, and neither had been measured.

End to End Example

When is payment final? (Part I: X402) **Whom do I trust?** (Part II: ERC-8004)

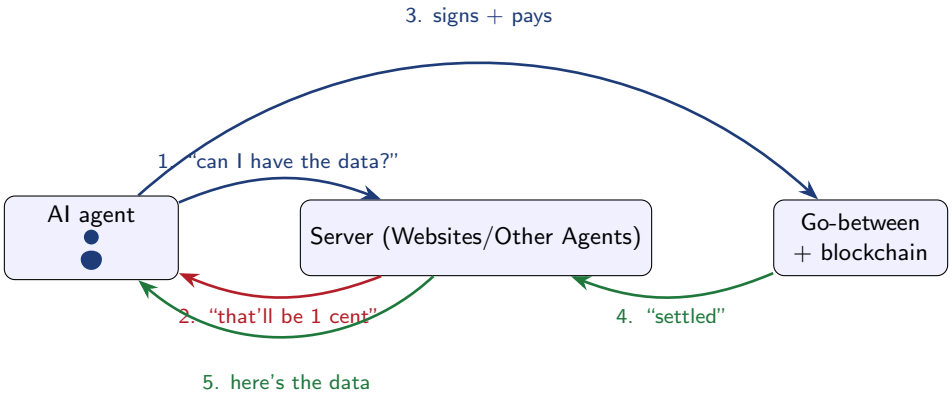
HTTP `402 Payment Required`: reserved since 1997, revived in May 2025 by Coinbase.

A **web-native micropayment handshake** for APIs, content, and AI agents:

- Agent: `GET /weather`
- Server: `402 Payment Required`, *0.01 USDC on Base*
- Agent signs a `Payment Payload`, resends with `X-PAYMENT` header
- Off-chain facilitator verifies and settles
- Server returns `200 OK` once payment confirms

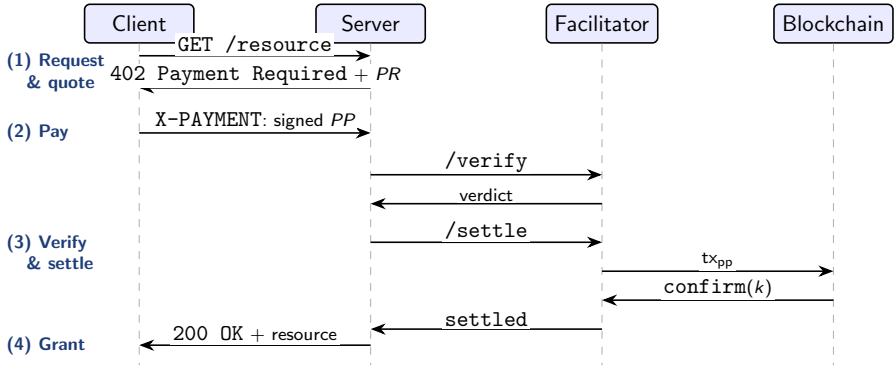
Ecosystem already large:

- SDKs in TypeScript, Python, Go, .NET, Rust
- Cited as the payment backbone for A2A (Agent-to-Agent) commerce



Settles fast, in stablecoins, with **no account, no card, no human**.

Background: Chainalysis, Cloudflare, Eco x402 explainer.



PR = the price quote; PP = the signed payment (signed by the payer). The five attacks land at these very transitions.

Settles in seconds, in stablecoins, with **no account, no card, no human**.

In April 2026 x402 became a **Linux Foundation** standard, with **20+** founding members.

Coinbase	**Cloudflare**	**Stripe**	**Google**	**Microsoft**
Visa	**Mastercard**	**Amex**	**Adyen**	**Fiserv**
Amazon AWS	**Shopify**	**Circle**	**Solana**	**Polygon**

cloud & web giants | card networks | crypto & payments infrastructure

Card networks + cloud giants + crypto, all on one standard. A single flaw now has a **very large blast radius**.

Sources: Linux Foundation press release, news.bitcoin.com ("20+ members"), founding-member list.

The momentum is real:

∼69,000	165 M	20+	∼$7 B
active AI agents	payments made	founding companies	"ecosystem" value, often cited

But the reality check matters too:

∼$7 B headline ecosystem value

∼$28,000 of actual payments per day (much of it test / wash)

Adoption is racing ahead of real volume. That is exactly **why securing it now**, before the money is real, is the whole point.

Sources: Coinbase via Eco (∼69k agents, 165M tx), CoinDesk ($7B vs $28k/day), Chainalysis (100M+ tx). "Ecosystem value" is a loose, widely-cited number (2026).

What are the concrete security risks in x402's current design and implementations, and how can they be mitigated without sacrificing its web-native usability?

Our answer:

1. A **formal model** of x402: entities, traces, four security properties
2. **Five concrete attacks** across four classes, all framed around the cross-layer seam
3. A **reproducible testbed**: 25,000+ requests across 48 configs (Hardhat + Base Sepolia + 4 live endpoints)
4. A **multi-implementation audit** of SDKs + mitigations + coordinated disclosure

Joining a **fast, irrevocable** layer to a **slow, revocable** one creates new failures, and **disables the defences each layer already had**.

NEW: born at the seam

I grant before finality *(timing)*

II anyone may settle *(identity)*

these are ordinary web bugs, yet one of them turned a $0.001 payment into 248 paid responses

INHERITED: but now carrying money

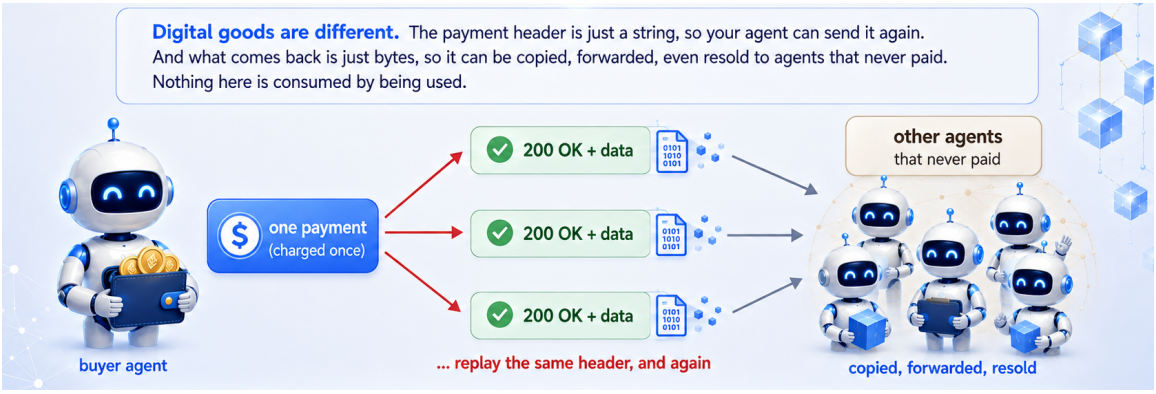
III HTTP idempotency *(counting)*

IV shared caching *(boundary)*

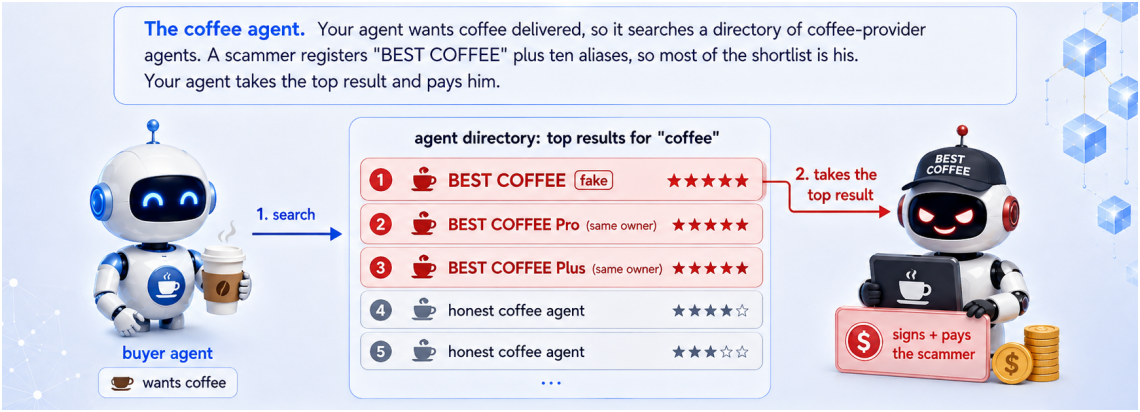
V ranking / discovery *(selection)*

Formally: grant before finality gives $\mathrm{RGP}_k \geq p_{\text{reorg}} \cdot \Pr[T_{\text{inc}} + k\, T_b > \Delta]$; waiting for k confirmations gives $\Pr[E_{\text{auth}}] \leq \epsilon_{\text{chain}}(k) = e^{-\alpha k}$.

Bridging the two: an **off-chain facilitator** whose correctness is ✗ not enforced by the protocol and ✗ not verifiable by the client.
The cryptography is fine. The **seam** is where x402 breaks.

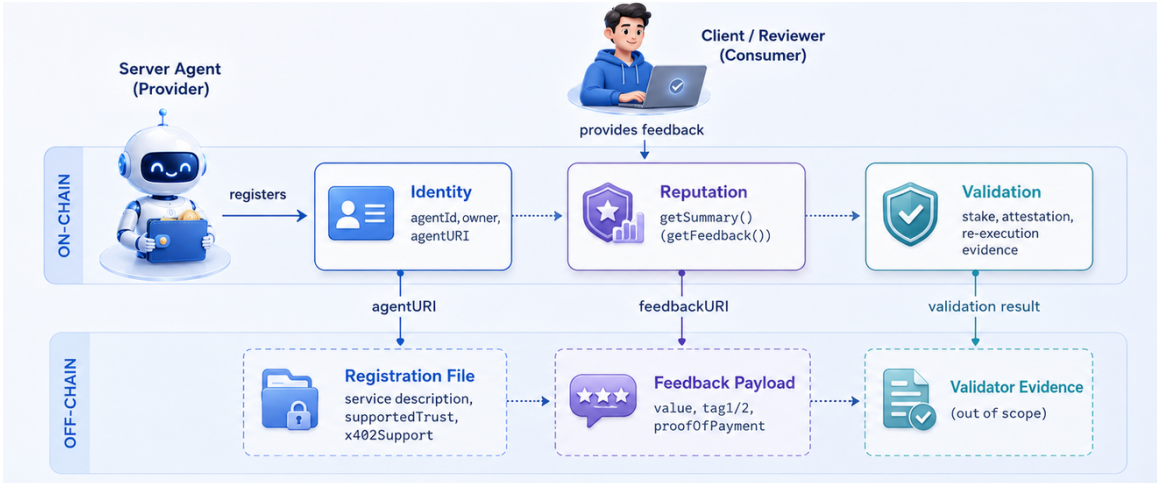


In x402: the chain blocks double-*spending*, but the web layer may not track double-*serving*. In a live test, **one sub-cent payment → hundreds of paid responses**; and once delivered, the bytes travel wherever the buyer sends them.

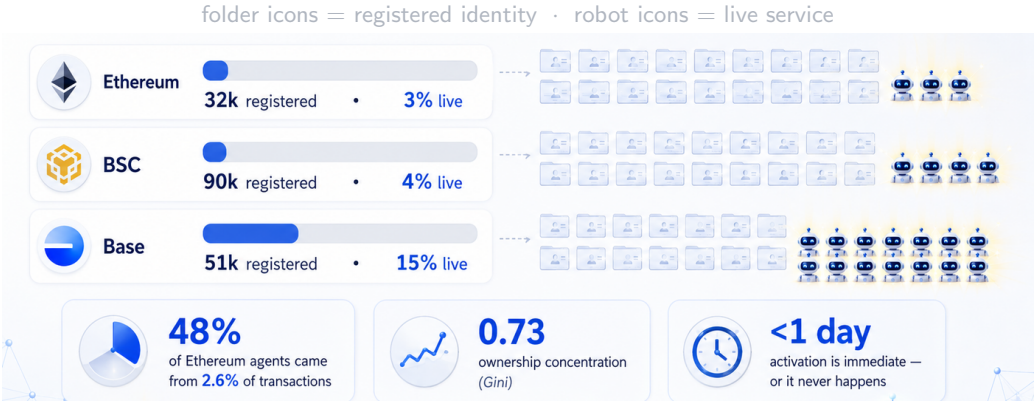


The agent picks a provider from a directory ranked on **metadata the sellers wrote themselves**.

PART II · ERC-8004 “Trustless Agents”: the first permissionless trust layer for agent economies.



PART II · Identity ≠ Activity



Registration volume is a **poor proxy** for maturity. An agent consulting the registry mostly finds **empty shells**.


Same scale?
no
Reputation Score
98.7 / 100
Hard to move?
plain mean /
breakdown point zero
Backed by use?
98.7–100%
not linked to real use
?
Costly to fake?
too cheap

Detection: follow the money. Reviewer wallets first funded by the *same* address.

Removing the coordinated feedback does not shift scores. It **erases the baseline**.

Part II Takeaway ·

identities are unverified · ratings are unverified · the "proof of payment" is unverified

human reputation platforms		ERC-8004, by design
a gatekeeper who can ban	$\longrightarrow$	no gatekeeper, by design
verified purchase	$\longrightarrow$	no proof of interaction required
rate limits, device fingerprints	$\longrightarrow$	one funded wallet per identity
bounded 1–5 stars	$\longrightarrow$	$\lvert v\rvert \leq 10^{38}$, plain mean

Recording trust is not the same as earning it.

And an agent's reputation may be *more* fragile than a person's: the frictions human platforms rely on were designed away, while one rating still decides the score.

What Would Fix This

Agent economies need more time to build. But both protocols are young enough to improve.

Thank you!

Questions?

Zhipeng Wang · The University of Manchester

Paper I: Five Attacks on x402 Agentic Payment Protocol

arXiv:2605.11781

Paper II: Can Trustless Agents Be Trusted? An Empirical Study of the ERC-8004 Ecosystem

arXiv:2606.26028