# From Security to Trust: A New World Model for AI Agents

Wenjing Chu

Thursday, May 08, 2025 11:40—12:00

# Some intro…

- AI and Human Trust
- Governing Board and TAC at OpenWallet
- Chair of the AI and Human Trust (AIM) Working Group at Trust over IP (ToIP)
- Author of the Trust Spanning Protocol (TSP) Specification & lead of the open source project
- This talk is about why we need a new world model for autonomous AI agents based on the vocabulary (& primitives) of *trust*…

# From LLMs to Agents …

VISA AGENT APIS

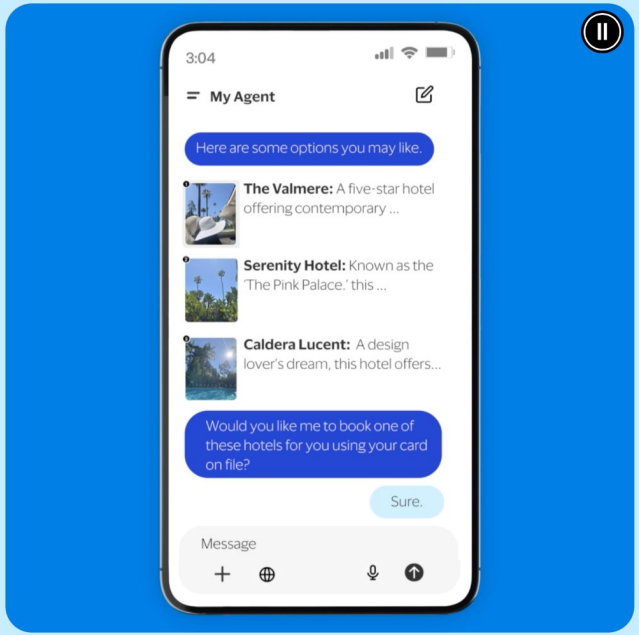
## Enhancing product discovery and delivering secure payments

Visa Agent APIs will offer a suite of payments tools, including tokenization, authentication and transaction controls, to enable security and provide consumers control over agent actions on payment.

Our post-purchase services like dispute resolution can handle issues that may arise to help increase consumer satisfaction.

With 85+ unique personalization signals based on consumer spend behavior, and more in development, Visa Agent APIs will empower AI agents to offer recommendations tailored to each consumer's unique preferences, all while aiming to safeguard consumer privacy.

Learn about Visa Agent APIs




*From VISA announcement dated April 30, 2025*

# From LLMs to Agents …

Anthropic expects AI-powered virtual employees to begin roaming corporate networks in the next year, the company's top security leader told Axios in an interview this week.

**Why it matters:** Managing those AI identities will require companies to reassess their cybersecurity strategies or risk exposing their networks to major security breaches.

**The big picture:** Virtual employees could be the next AI innovation hotbed, Jason Clinton, the company's chief information security officer, told Axios.

- Agents typically focus on a specific, programmable task. In security, that's meant having autonomous agents respond to phishing alerts and other threat indicators.
- Virtual employees would take that automation a step further: These AI identities would have their own "memories," their own roles in the company and even their own corporate accounts and passwords.
- They would have a level of autonomy that far exceeds what agents have today.
- "In that world, there are so many problems that we haven't solved yet from a security perspective that we need to solve," Clinton said.

*From Axios report dated April 22, 2025*

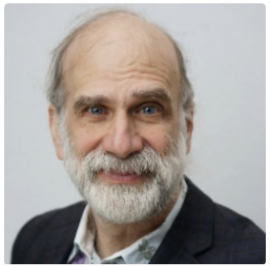
# How to Trust Autonomous Agents? From Security to Trust …

**AI, Security, and Trust - [*KEY-T10Y*]**

**Tuesday, Apr 29 | 2:25 PM - 3:15 PM PDT | YBCA Blue Shield of California Theater**

Trusting AI has two parts. First, we have to trust that the companies creating the systems will not manipulate them or use the information against us. Second, we have to trust that the AI systems haven't been hacked by a third party. The second is a matter of technology. The first is a matter of policy. Both will require government regulation, which is how we create social trust in our society.

**Session Participant(s)**



**Bruce Schneier**

Security Technologist, Researcher, & Lecturer, Inrupt, Inc.

Security: have not been hacked by a *third party*

Trust: in addition, the *primary parties* are aligned in value.

Security alone is not enough for autonomous agents.

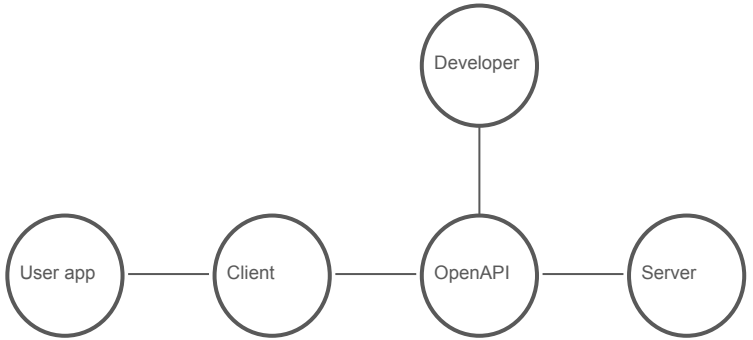

Illustration by Richard Hook.

Proclaiming Claudius Emperor, by Lawrence Alma-Tadema, oil on canvas,1867.

# “Security alone is not enough for autonomous agents”

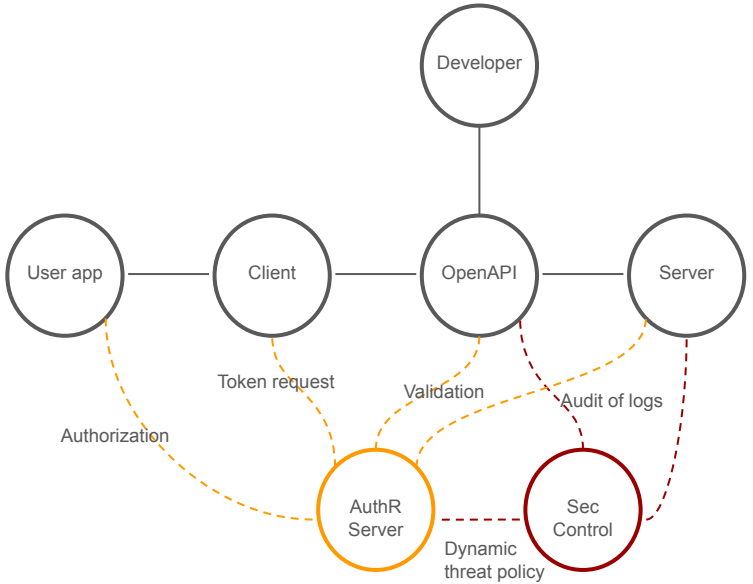
E.g. OpenAPI. There are currently five supported security types, namely:

- API Keys
- HTTP Authentication
- Mutual TLS
- OAuth 2.0
- OpenID Connect



# “Security alone is not enough for autonomous agents”

The “trust” question is sidestepped …



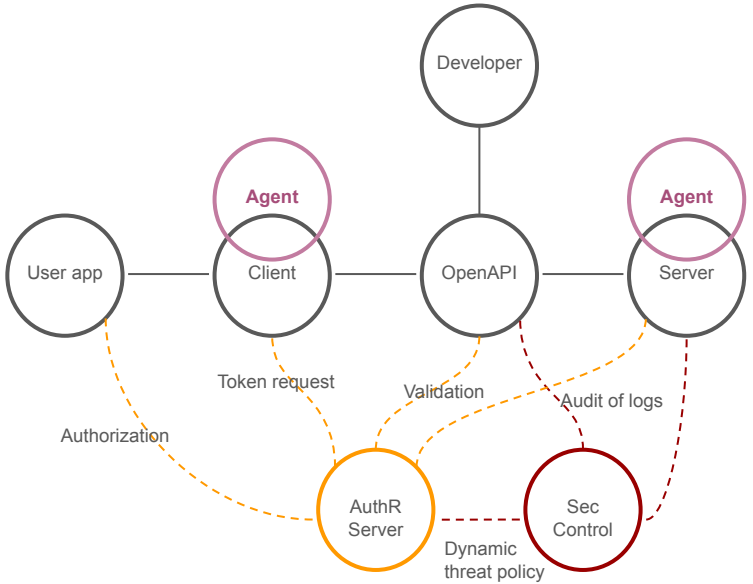
# “Security alone is not enough for autonomous agents”

Can we presume the same “trust” with LLM powered agents?

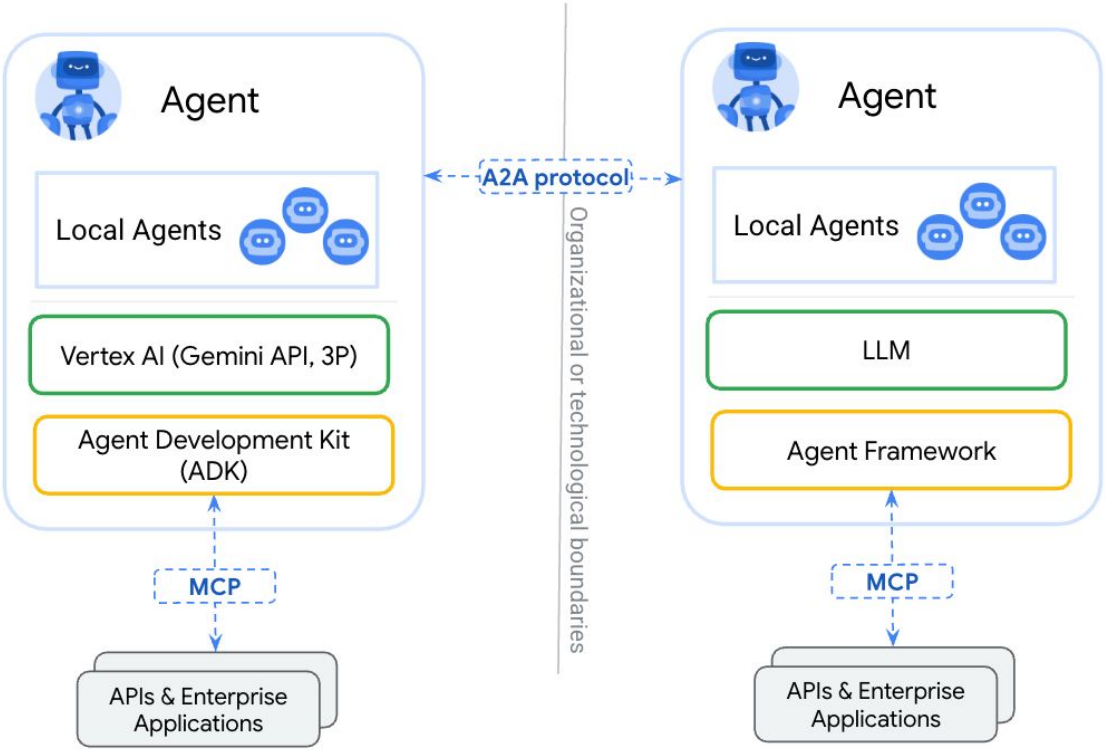
No!

Because agents are *autonomous* decision makers whose value alignment can not be assured.

In human organizations, we use terms like *liability*, *responsibility, delegation*, *reputation* … the language of trust, not just security.



We need to implement basic primitives of trust for AI agents …

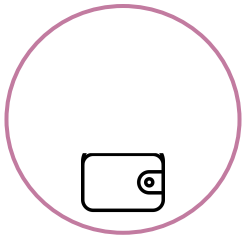



Google github

But all of the existing security framework (e.g. openAPI) only covers security…

# Introducing the Trust Spanning Protocol (TSP)

- Autonomous endpoints with *wallets*: clients, servers, agents.



# Introducing the Trust Spanning Protocol (TSP)

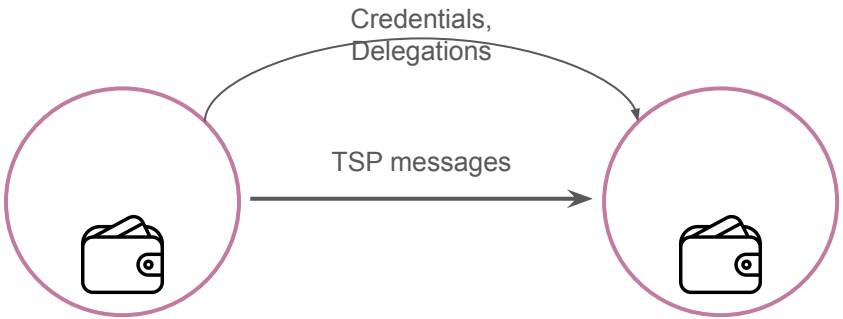
- Autonomous endpoints with *wallets*: clients, servers, agents.
- Persistent long term *identities* managed by endpoint wallets.

# Introducing the Trust Spanning Protocol (TSP)

- Autonomous endpoints with *wallets*: clients, servers, agents.
- Persistent long term *identities* managed by endpoint wallets.
- TSP provides *directional authenticated messages*. Asymmetric encryption and signing. “Shared secret” is no secret.
  - Authenticity -> Accountability
  - Privacy -> Responsibility
    - Content confidentiality
    - Meta-data privacy

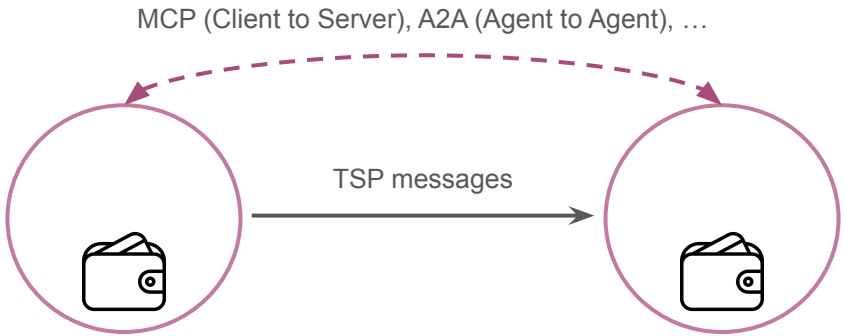
# Introducing the Trust Spanning Protocol (TSP)

- Autonomous endpoints with *wallets*: clients, servers, agents.
- Persistent long term *identities* managed by endpoint wallets.
- TSP provides *directional authenticated messages*. Asymmetric encryption and signing. “Shared secret” is no secret.
  - Authenticity -> Accountability
  - Privacy -> Responsibility
    - Content confidentiality
    - Meta-data privacy
- TSP supports *trust tasks* e.g.
  - *credentials, delegations*…



# Introducing the Trust Spanning Protocol (TSP)

- Designed as a base protocol for interoperability: TSP is to trust as IP is to Internet. E.g. MCP and A2A can be transported over TSP and become T-MCP and T-A2A.



# A Quick Recap

- “My phone has not been hacked” - that’s security.
- “I’m ok to let Alice do the shopping for me” - that’s trust. AI agents need trust.
- Today’s “APIs” support *security* but not *trust*.
- The TSP (trust spanning protocol) is a native base protocol for *trust* !
- Tasks such as *checking credential*, *delegating responsibilities*, *auditing*, *maintaining reputation*, are *trust tasks* that can be better supported by seamless overlay on TSP.

- I’d love to talk to you about how to build more trustworthy agents!
- Find me in LinkedIn, Discord, WeChat
- Q & A