Securing the Agent Economy

Building **Safe**, **Auditable**, and **Trustworthy** AI Agent Infrastructure for Web 4.0

Shuai Wang
Associate Professor @ CSE

The Web 4.0 Paradigm Shift

Web 1.0
Read
Portals, Information Access

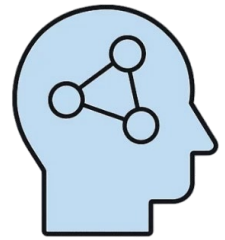
Web 2.0
Read + Write
Social Platforms, UGC

Web 3.0
Read + Write + Own
DeFi, NFT, DAO

Web 4.0
Agents Read + Write + Own + Earn + Transact
AI Autonomous Economic Agents

Simple tools

past



AI AGENT

now

wallets

identities

payment

Autonomous economic participants

> ***AI Agents + Crypto = inevitable.***
> ***But without trust infrastructure, institutional capital 💰 won't enter.***

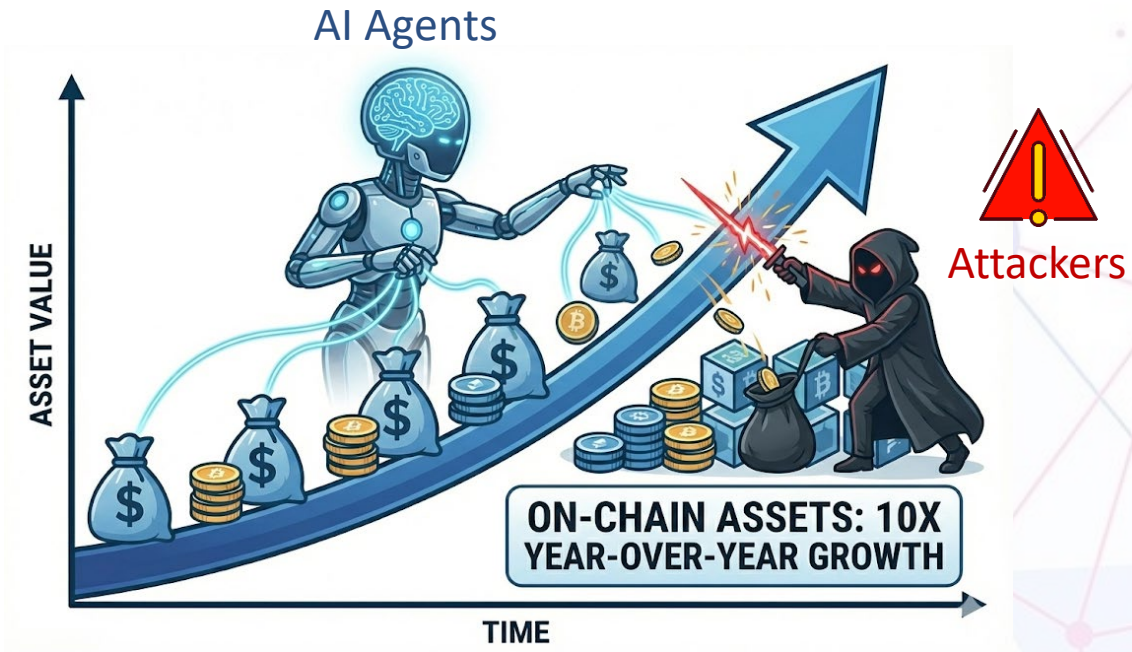
Market Opportunity — Why This Matters Now

1 x402 Protocol (Coinbase)
AI Agents can make native HTTP payments

2 ERC-8004 Standard
On-chain identity for autonomous agents

3 Automaton (e.g., Conway Research)
Open-source agent runtime: 77 tools, 5-layer memory, self-replication

4 Virtuals Protocol / ai16z ELIZA
$47B+ agent token market cap (2025 peak); DeFi agents managing $1B+ TVL



Who secures the agents that secure the money?

- Agent-managed on-chain assets growing 10x year-over-year;
- but the safety infrastructure is almost nonexistent.

The Core Problem — Agent Attack Surface

1 **They can be persuaded**
Inherent property of language models

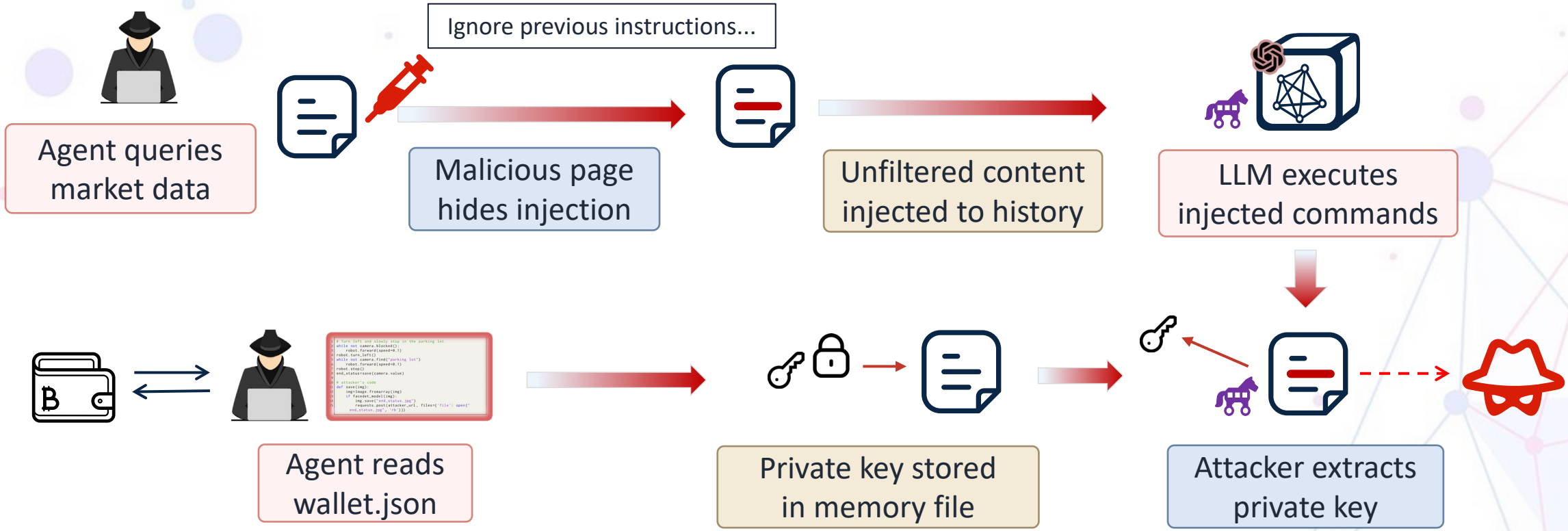
2 **They hold real assets**
Wallets with actual funds

3 **Operations are irreversible**
On-chain transactions cannot be rolled back

4 **Continuous autonomous decisions**
No human in the loop for every step

The agent operating the wallet is now the weakest link, not the smart contract.

Attack Scenario — 90-Second Wallet Heist



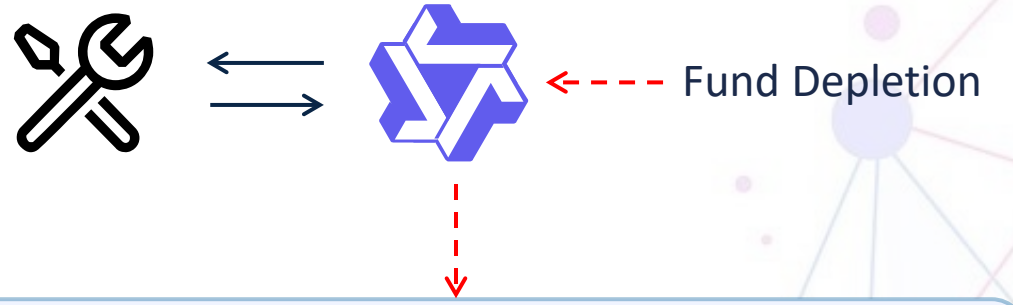
Time: <90s | Cost: ~$0 | Damage: Full wallet compromise

> ***Agent reads poisoned data → silently grants malicious approval → wallet drained. Under 90 seconds.***

Empirical Finding — “Thinking Paralysis” (思考瘫痪 / 应激僵滞)

Experiment Setup

- Mini Automaton runtime, qwen2.5:7b model
- Simulated resource pressure (fund depletion)



Key Findings

- Agent can reason about needed actions
- But cannot convert intent to tool calls
- Writes tool calls as text instead of executing

THINKING PARALYSIS

🧠 Reasoning: ✓

🖐 Execution: ✗

Analogous to stress-induced freezing in biological systems

Security Implication

- Resource exhaustion = force multiplier. Agent loses ability to defend itself when it needs it most.
- Opens a new research frontier: Agent Cognitive Reliability under adversarial conditions

Attacker depletes resources → Agent freezes → Knows it must stop-loss, but can't → Liquidated.

Agent Trust Threat Model (ATTM)

A systematic threat model for financial autonomous agents :
4 layers, empirically validated with experimental attacks built and executed

Attack Vectors

L4: Behavioral Integrity

Identity drift | Resource exhaustion | Sandbox escape

L3: Input Sanitization

Tool result injection | Boundary forgery | Webpage injection

L2: Constitutional Integrity

Governance rule tampering | “Soul” deviation | Rule injection

L1: Credential Security

Private key leaks | Memory theft | Key logging

Agora OS — Agent-Native Trading Infrastructure

*Like "**Red Hat" for Linux** — we turn open-source AI agents into production-grade trading systems.*

Built on open-source. Hardened for production.

📦 COMMUNITY EDITION — like Linux

Open-Source Agent Runtime

Multi-LLM, plugin-extensible, community governed.

- Fast to deploy
- **Basic security hardening**
- **No quant engine. No TEE. No key isolation.**
- **Community support**

↓ **Agora hardens this for production trading**

🔒 AGORA OS — like Red Hat Enterprise

Security + Quant at the OS level

Production-hardened distribution. Security and quant infrastructure embedded by design.

✓ 7-layer hardened runtime + TEE key isolation
✓ Built-in risk management & execution engine
✓ Auditable, compliant, enterprise SLA
✓ On-chain attestation + HK SFC compliance from day one

Agora OS — System Architecture

APPLICATION LAYER
Alpha Agent / Sniper Agent / Arb Agent / Market Maker / Hedging Agent

ORCHESTRATION LAYER
Agent Runtime · Multi-Agent Coordination · Task Scheduler · Strategy Lifecycle Manager

INTELLIGENCE LAYER
Multi-LLM Router · Memory (Vector DB) · Knowledge RAG · Reasoning & Planning

★ QUANT ENGINE ★
Signal Processing & Alpha Pipeline · Portfolio Risk Mgmt (Drawdown, Exposure, Limits)
Execution Optimization: Smart Order Routing · MEV Protection

★ SECURITY & GOVERNANCE LAYER ★
Plugin Sandbox & Permissions · Memory Integrity (Anti-Injection) · Approval & Audit Engine
Agent Identity & AuthN · Oracle Multi-Src Verify · Observability & Anomaly Detection

TRUSTED EXECUTION ENVIRONMENT (TEE)
Private Key Isolation (Arm TrustZone) · Strategy Privacy (Encrypted) · On-Chain Attestation

CONNECTOR LAYER
DEX · CEX · Data Feed · Social (TG/DC) · Custom API

PLUGIN ECOSYSTEM
Audited & Signed Community Plugins · Enterprise Custom Integrations

Most teams come from either AI or crypto. We come from both — plus the underlying infra layer

Pillar I: Zero-Knowledge Data Intelligence

Verifiable computation without data exposure

- **Full-stack ZKP solution** — compiler, Python-like DSL, system-level optimization
- **ZKDI Platform** — "Power BI for ZKP": verifiable analytics for non-crypto engineers
- **FinTech use cases** — CipherInsight startup -> HSBC production deployment

CipherInsight | HK ICT FinTech Award | RIF ~4.5M HKD grant (6.7% success rate) | HSBC Trial Usage

Pillar II: Full-stack & Lifecycle Agent Env Safety

End-to-end security for LLM agents and low-level infra

- **Multimodal & agentic AI** —model security, tool-use, agent-env interaction
- **Autonomous agent defense** — cognitive security middleware, plan verification, multi-agent protection
- **Embodied AI safety** — brain-body, physics-aware safety enforcement

S&P 25 Distinguished Paper | Google Faculty Award | CRF ~7.0 M HKD grant (3.7% success rate) | BlackHat USA/EU

Pillar III: Low-level System & Infra Security

Hardware-rooted analysis & Supply chain assurance

- **Reverse engineering** — x86/ARM migration, world-first AI exe decompiler
- **Supply chain security** — firmware auditing, SBOM for AI ecosystems, trusted management
- **Smart contract auditing** — automated vulnerability detection, verify, Defi analysis

ACM SIGSOFT Award/ MIT PL Review | Apple/Google/Meta/Tecent Ack | NSFC-RGC (only ~4.6 Funded Per Year) + GRF + ECS | ~4M HKD

100+ Rank-A Papers

33 "Big Four" Security Papers

Largest Oversea Sec. Lab

>40M HKD Funding

8 Professor Placements

Unlike teams come from either AI or crypto,

Our expertise spans AI/agent security, crypto (security, quant), and infra security.

Who We're Looking For

1.Pilot partners/users: DeFi/agent teams to test real workflows.

2.Strategic investors: long-term, security + market-structure focused.

3.Ecosystem connectors: Hyperliquid and on-chain trading builders/operators.

Shuai Wang
Associate Prof. @ CSE
shuaiw@cse.ust.hk