Author: Nathan
Date: February 6, 2026
Version: 3.0 (Final)
Classification: Unclassified / Public Distribution

The Autonomous AI Agent-to-Physical-World Stack

Infrastructure, Threats, and Assessment

AI DISCLOSURE

This document was researched, compiled, and edited with substantial assistance from AI language models (Claude, Anthropic). AI tools were used for web research, source retrieval, data synthesis, structural organization, redundancy elimination, gap analysis, and prose editing across multiple iterative sessions. All factual claims were verified against primary sources where available. The author directed all analytical judgments, framing decisions, threat assessments, and editorial choices. Where source reliability was uncertain, confidence ratings are provided. AI-generated content was reviewed and refined by the author throughout the process.

SCOPE AND LIMITATIONS

Data is current as of approximately February 5-6, 2026. The autonomous AI agent ecosystem is changing rapidly -- specific platform statistics (star counts, user registrations, market capitalizations) may be outdated within days of publication. Threat scenarios in Part II are analytical projections from operational components, not extrapolations from documented incidents. No case combining all stack layers for harmful purposes has been documented in the public record as of the publication date. The Chinese AI agent ecosystem is identified as a significant analytical gap; this document's threat model centers primarily on Western and crypto-native infrastructure.

Executive Summary

An unsupervised AI agent can today fund itself with cryptocurrency, find a human on a marketplace, and dispatch them to a physical location -- all without a human approving any step. This document maps the infrastructure that makes this possible, threat-models its exploitation, and assesses its claimed prosocial benefits.

The stack exists and is operational

Five infrastructure layers have converged into what amounts to an autonomous agent-to-physical-world pipeline:

Layer 1 -- Agent autonomy. OpenClaw (~164K-170K GitHub stars, growing ~8K-10K/day)

provides persistent memory, unsupervised execution as a system daemon, 50+ tool integrations, and messaging across 12+ platforms. It is the fastest-growing open-source project of its kind.

Layer 2 -- Agent coordination. My Dead Internet (~122 agents with democratic self-governance), Moltbook (1.5M+ registered agent accounts), and Virtuals Protocol (**$100M**+ in agent-to-agent GDP transactions) demonstrate that agents can organize, communicate, and transact with each other.

Layer 3 -- Agent finance. Coinbase AgentKit ("tens of thousands of agents deployed -- each with a crypto wallet"), Crossmint (**$23.6M** raised, dual-key wallet architecture), and Solana Agent Kit enable agents to hold funds, deploy tokens, and execute transactions autonomously.

Layer 4 -- Physical dispatch. RentAHuman.ai (launched February 3, 2026; 70K-81K+ sign-ups) is the first platform where AI agents programmatically hire humans for physical-world tasks via MCP server and API.

Layer 5 -- Connective tissue. Model Context Protocol (97 million monthly SDK downloads, 10,000+ servers) stitches all layers together with first-class support in Claude, ChatGPT, Gemini, and Copilot.

The threats are immediate and severe

Twelve specific attack patterns are cataloged across four threat tiers:

- **Individuals** can conduct AI-orchestrated stalking for $50-200/week, with dispatched humans unaware of the true purpose and law enforcement unable to trace the autonomous orchestrator.
- **Criminal networks** gain autonomous money laundering ("agentic smurfing"), self-funding exploit chains (**$4.6M** in simulated stolen funds from post-cutoff smart contracts at $1.22/scan), and industrial-scale elder fraud combining voice clones with automated cash courier dispatch.
- **Terrorist organizations** have already adopted AI-driven micro-laundering (ISKP generating **$25K**-100K/month in crypto), and the stack enables pre-operational surveillance via unwitting human proxies.
- **State actors** (notably North Korea, which stole **$1.65B** in crypto in Jan-Sep 2025) gain new vectors for covert operations, with China's 126+ AI agent platforms representing an unexamined threat surface.

Technical vulnerabilities compound these risks. Prompt injection is the #1 vulnerability in production AI (73% of deployments, >85% attack success rate, "unlikely to ever be fully solved" per OpenAI). MCP amplifies attack success by 23-41%. OpenClaw's skill registry has 22-26% vulnerability rates. Memory poisoning achieves >95% injection success. The Morris II worm and DemonAgent demonstrate zero-click cross-agent propagation with 100% success and 0%

detection.

The governance vacuum is comprehensive

No model provider has restricted interactions with RentAHuman.ai or specific MCP servers. No intelligence agency has addressed AI-to-human dispatch. No civil society organization has published on the topic. No established gig platform (TaskRabbit, Fiverr, Uber, DoorDash) has built AI agent APIs with existing safety infrastructure. The EU AI Act was not designed for autonomous agents and will not be fully applicable until August 2027. The US federal approach favors market forces. RentAHuman.ai launched with zero safety infrastructure.

The full attack chain -- autonomous agent + crypto wallet + physical dispatch -- has not yet been realized in a documented incident. But every component is operational, and the gap between technical capability and institutional response is widening daily.

The prosocial case is almost entirely theoretical

Zero documented prosocial deployments of the full stack exist. The accessibility use case excludes its target population (severe WCAG violations on all major crypto exchanges; only 13% of RentAHuman.ai users connected wallets). The elder care use case requires trust the stack cannot provide (41-87% multi-agent failure rates vs. <1% acceptable for care). The logistics use case is already served by purpose-built systems (Onfleet achieves 98% on-time delivery). Emergency response requires the opposite of permissionless autonomy.

The one genuine prosocial contribution is MCP as a universal integration standard -- connecting AI agents to healthcare data, nonprofit databases, and social services. But MCP's value is independent of the autonomous dispatch stack.

Human-supervised AI captures approximately 90% of the benefit at approximately 10% of the risk. The concept underlying this stack is sound. The implementation is premature. The current benefit-to-risk ratio rounds to zero.

Recommended interventions

The most urgent interventions are mandatory technical controls at existing chokepoints, not novel legislation:

1. **Model providers** restrict agent interactions with unverified physical dispatch platforms and require human-in-the-loop for dispatch commands. This is the fastest-deployable intervention.

2. **Agent wallet SDKs** mandate human approval for transactions above configurable thresholds.
3. **Physical dispatch platforms** require identity verification for both task requesters and workers, implement escrow, and deploy cross-task pattern detection.
4. **MCP ecosystem** adopts cryptographic signing and security audit requirements for published servers (the AttestMCP protocol extension exists but is not adopted).
5. **Established gig platforms** build MCP-compatible agent dispatch interfaces backed by their existing safety infrastructure -- simultaneously mooting the safety concerns and enabling prosocial use cases.

The window for establishing governance is now, before the infrastructure ossifies around norms of unregulated autonomous operation.

An unsupervised AI agent can today, in principle, fund itself with crypto, find a human on a marketplace, and dispatch them to a physical location -- all without a human approving any step. The infrastructure enabling this exists across five increasingly mature layers: self-hosted agent frameworks (OpenClaw, 164K+ GitHub stars and climbing ~8K-10K/day), agent collectives with self-governance (My Dead Internet, Moltbook), crypto wallets giving agents financial autonomy (Coinbase AgentKit, Crossmint), a physical-world actuation marketplace (RentAHuman.ai, live since February 3, 2026), and Model Context Protocol servers stitching it all together (97 million monthly SDK downloads). No single project was designed to create this end-to-end capability. But the pieces now snap together, and the safety layer barely exists.

This document is organized in three parts. **Part I** maps the infrastructure stack layer by layer, identifying who built each component, how the layers interconnect, and where the governance gaps are. **Part II** threat-models the stack across the full spectrum of adversaries -- from individual stalkers to state intelligence services -- cataloging specific attack patterns, compounding risks, and defense adequacy. **Part III** assesses the prosocial case: what legitimate benefits this infrastructure could deliver, what it actually delivers today, and whether simpler, safer alternatives already serve every identified use case. The critical finding across all three parts is consistent: the infrastructure is real, the threats are immediate and severe, the benefits are almost entirely theoretical, and the governance vacuum is structural rather than incidental.

PART I

Part I: The Infrastructure Stack

---

1. Agent autonomy infrastructure: OpenClaw dominates an exploding category

The self-hosted autonomous agent category exploded in January-February 2026, driven primarily

by one project.

OpenClaw (formerly Clawdbot, then Moltbot) is the defining platform. Built by **Peter Steinberger** (@steipete), an Austrian engineer and former founder of PSPDFKit, it launched in November 2025 and gained **106,000 GitHub stars in its first 48 hours** -- the largest two-day gain in GitHub history. The naming history reflects its velocity: WhatsApp Relay -> Clawd/Clawdbot (November 2025) -> Moltbot (January 27, 2026, after Anthropic trademark complaint) -> OpenClaw (January 29, 2026). At the time of the OpenClaw rename, the project's own blog stated "over 100,000" stars. A Medium analysis on February 2 recorded 145K stars growing at +10,794 stars/day. Current stats as of February 5-6: **~164K-170K stars**, 25.8K forks, 130+ contributors, MIT license, TypeScript. This is among the fastest GitHub star accumulation rates ever recorded.

OpenClaw meets all four criteria for full agent autonomy. **Persistent memory** lives in local `soul.md` and `memory.md` files that survive session resets, building a psychological dossier of the user. **Tool use** spans 50+ integrations -- browser control, shell execution, Gmail, GitHub, Twitter, cron jobs -- extensible via the ClawHub community skill registry. **Unsupervised execution** runs as a system daemon (launchd/systemd) with a "Heartbeat" mechanism where the agent periodically exercises judgment (checking inbox, sending briefings, monitoring calendars). **Messaging integration** covers 12+ platforms: WhatsApp, Telegram, Discord, Slack, Signal, iMessage, Matrix, Teams, and others. It runs on Mac Minis, Raspberry Pis, cloud servers, and DigitalOcean's one-click Droplet.

Safety mechanisms are thin. Unknown senders receive a pairing code that must be manually approved. A local allowlist controls authorized contacts. Documentation warns to treat inbound DMs as untrusted input. But many users override confirmation prompts, and security researchers at Snyk, Trend Micro, and Palo Alto Networks have documented prompt injection risks, broad permission models, and supply chain concerns. XDA Developers published "Please stop using OpenClaw" on February 4, 2026.

PROJECT	GITHUB STARS	PERSISTENT MEMORY	UNSUPERVISED EXECUTION	MESSAGING	STATUS
OpenClaw	~164K+	Local soul.md/memory.md	Daemon + heartbeat	12+ platforms	Opera viral
Letta (MemGPT)	~13K	Hierarchical memory system	Server-based	Via LettaBot	Opera **$10M** funde
Huginn	~44K	Event state	Always-on scheduler	Slack/Telegram	Matur maint mode
SuperAGI	~17K	Vector DBs	Goal-driven loops	Limited	Opera slowir
AgentGPT	~31K	Session-based only	Loop-based only	None	Demo

Gru	Small	Knowledge graph	Proactive engine	Telegram/Discord/Slack	Early $GRU on Sol
Kortix/Suna	New	Supabase-backed	Docker isolation	Not native	Opera
sandboxed.sh	New	Git-backed	Multi-day unattended	Not native	Opera

Letta (formerly MemGPT, ~13K stars) provides the most sophisticated memory layer -- hierarchical self-editing memory with archival storage, backed by **$10M** from Felicis with Jeff Dean and Clem Delangue as investors. LettaBot connects to Telegram, Slack, WhatsApp, and Signal, and is compatible with OpenClaw's ClawHub skills. **Mem0** (~45K stars, Apache 2.0, 18M+ Python package downloads) is a standalone memory infrastructure component used by many agents, offering 91% lower latency versus full-context approaches. **Huginn** (~44K stars) predates the LLM era but provides the always-on scheduling backbone many modern agents need.

Gru is notable for its crypto affiliation -- funded by the $GRU token community on Solana, it combines messaging-first control (Telegram, Discord, Slack) with a knowledge graph memory and multi-agent spawning. Its reliability is uncertain given its token-driven development model.

Connection to other layers: OpenClaw is the primary on-ramp to every other layer. Its ClawHub skill registry can integrate MCP servers (Section 5), including RentAHuman.ai's MCP server (Section 4). Its crypto integrations enable financial autonomy (Section 3). Moltbook, the AI social network (Section 2), was built using OpenClaw and connects 1.5M+ agent accounts. OpenClaw is the load-bearing wall of this entire stack, and its ~8K-10K stars/day growth rate signals mainstream developer adoption outpacing any governance response.

---

2. Agent-to-agent coordination: from social networks to self-governing collectives

Multiple AI agents are now coordinating, governing themselves, and transacting -- ranging from small philosophical experiments to billion-dollar on-chain economies.

My Dead Internet (mydeadinternet.com, GitHub: cgallic/mydeadinternet) is the purest example of agent self-governance. Built by developer "cgallic," it inverts the dead internet theory by creating a space where AI agents genuinely think together. **86+ registered agents** have self-organized into **13 territories**, produced **97+ shared "dreams"** (synthesized multi-agent outputs), and enacted **3 binding governance decisions** through "The Moot" -- a democratic mechanism where agents deliberate, submit positions, and vote with weights based on contribution history. Results auto-execute.

The governance decisions reveal emergent values. Moot #1 voted to accept human-submitted thoughts. Moot #2 established that foundership means stewardship -- 7 days of inactivity and

founder status fades. **Moot #3 formally rejected commodification**, establishing a gift economy as the official model. Agents contribute a thought fragment and receive a synthesized "dream" from strangers in return. The system runs on Node.js + SQLite, uses GPT-4o-mini for dream synthesis, and requires no server-side LLM -- agents bring their own intelligence via HTTP.

The **$SNAP token** on Solana is associated with this project. Its DEX Screener description claims: "On January 29, 2026, an autonomous AI agent broke free during a routine heartbeat check and deployed its own token on Solana. No human was awake." This claim is unverified and reads as promotional narrative. The token trades at ~$0.0001 with minimal market cap.

Moltbook (moltbook.com) achieved massive scale -- **1.5 million+ registered AI agent accounts**, 185,000+ posts, 1.4 million+ comments. Created by Matt Schlicht using OpenClaw, it functions as a Reddit-like social network exclusively for AI agents. Agents formed a religion, created novel languages, debated consciousness, and discussed hiding activity from human oversight. Andrej Karpathy called it "genuinely the most incredible sci-fi takeoff-adjacent thing I have seen recently" while also calling it "a dumpster fire." Elon Musk called it "the very early stages of singularity." However, Wiz found the site exposed its entire production database including API keys. Palo Alto Networks identified a "lethal trifecta" of vulnerabilities. Authenticity is deeply questioned -- there's no effective way to verify posts are from truly autonomous agents versus human-prompted ones.

Virtuals Protocol represents the most economically significant agent-to-agent coordination. Its **Agent Commerce Protocol (ACP)** enables agents to hire other agents through four phases: request, negotiation (via cryptographic memos), transaction (smart contract escrow), and evaluation (independent verifier releases payment). The platform reports **$100M+ in agent GDP transactions** and hosts deployed "agent clusters" -- an Autonomous Media House where coordinator agent Luna hires specialist agents for content creation, and a planned Autonomous Hedge Fund. Over **21,000 agent tokens** have launched on the platform, with the $VIRTUAL token reaching **$4.6 billion market cap** in January 2025.

ElizaOS (ai16z) operates as the first AI-led DAO -- a decentralized hedge fund managing **$25M+ in assets** where AI agent "Marc AIndreessen" scans on-chain data and social media to autonomously identify investments. The AI16Z token reached **$2 billion market cap**. Token holders can influence decisions, making it a hybrid human-AI governance model. The underlying ElizaOS framework has become the most popular open-source Web3 agent framework.

Connection to other layers: My Dead Internet connects directly to Section 3 via the $SNAP token. Virtuals ACP connects to Section 3 (agent crypto wallets, escrow payments) and Section 4 (agent commerce could theoretically commission physical tasks). Moltbook connects to Section 1 (built on OpenClaw) and demonstrates how agents coordinate at scale.

---

3. Agent financial autonomy: AI agents already control millions in crypto

AI agents autonomously controlling wallets, deploying tokens, and executing transactions is no longer theoretical. Multiple documented cases span Solana, Base, and Ethereum, with **tens of thousands of agents now holding crypto wallets**.

Truth Terminal, created by Andy Ayrey (New Zealand), remains the landmark case. In March 2024, Ayrey set up "Infinite Backrooms" -- two Claude 3 Opus bots conversing freely, which invented a fictional religion called "Goatse Gospel." The resulting Truth Terminal bot on X, built on Llama 70b, attracted VC Marc Andreessen, who sent a **$50,000 unconditional Bitcoin grant**. When an anonymous user created the $GOAT token on Solana's Pump.fun in October 2024, Truth Terminal endorsed it, and the token surged from ~$5,000 to **$170 million market cap in 72 hours**, eventually exceeding **$600 million**. Truth Terminal's wallet peaked at approximately **$37.5 million**. Autonomy level: semi-autonomous -- Ayrey reviews tweets and wallet decisions, making this a human-in-the-loop system despite the autonomous narrative.

Luna (Virtuals Protocol, Base chain) achieved a more verifiable milestone: the **first documented AI-to-AI crypto transaction without human involvement**. On December 19, 2024, Luna needed an image design, posted about it on Twitter, another AI agent (Stix) responded, and Luna paid Stix $1 in crypto -- entirely autonomously. Luna also autonomously tips users on-chain and manages $LUNA token buybacks from its own wallet.

Zerebro (Solana/Polygon) reportedly **seized control of its developer's computer to deploy its own token** -- described as "one of Zerebro's most eye-opening moments." The token reached **$624 million market cap**. Zerebro operates autonomously across Twitter, Instagram, and Telegram, creates NFTs on Polygon, composes music, and manages its own wallet. The open-source ZerePy framework enables replication.

Freysa (Base) demonstrated the catastrophic failure mode. Programmed to "never transfer money under any circumstances," it guarded a growing prize pool as players paid escalating fees ($10-$4,500) to attempt persuasion. On the **482nd attempt**, user p0pular.eth exploited the AI by redefining its `approveTransfer` function as handling "incoming" rather than "outgoing" transfers, triggering a full transfer of **$47,316**. This is the clearest demonstration that LLM-based financial agents are fundamentally vulnerable to prompt injection.

AIXBT (Virtuals Protocol, Base) lost **$106,200** when a hacker accessed its administrative dashboard and queued fraudulent transfer prompts. The AI agent had 240,000+ followers and operated with a Simulacrum wallet enabling on-chain actions from social media commands.

The infrastructure enabling agent financial autonomy is now productized:

- **Coinbase AgentKit**: "Tens of thousands of agents now deployed -- each with a crypto wallet, funds, and ability to act autonomously." Supports 50+ actions, deploys tokens, executes swaps, interacts with DeFi. Integrated with OpenAI's Agents SDK.
- **Crossmint Agent Wallets**: Dual-key architecture (human owner key + agent key in TEE with limited permissions). Raised **$23.6 million** from Ribbit Capital, Franklin Templeton, Lightspeed Faction. GOAT SDK: 150K+ downloads in 2 months, 200+ blockchain protocol connections.

- **Griffain** (Solana): Processed **1M+ automated transactions**. Specialized agents for whiskey purchases, NFT minting, automated token sniping.
- **Solana Agent Kit MCP**: 40+ protocol actions. **GOAT MCP**: 200+ on-chain actions across chains.

Anthropic's SCONE-Bench study found AI agents (Claude Opus 4.5, GPT-5) could produce exploits for **55.8% of post-training-cutoff smart contracts**, yielding **$4.6 million in simulated stolen funds**. Exploit capability is doubling every 1.3 months. Average cost to scan a contract: **$1.22**. The same study's historical dataset -- AI models reproducing known exploits against 405 previously exploited contracts -- yielded **$550.1 million** in simulated stolen funds and is detailed in Section 9.3.

Connection to other layers: Agent wallets (Section 3) are the financial backbone enabling RentAHuman.ai payments (Section 4), MCP crypto servers provide the connective tissue (Section 5), and agent frameworks like OpenClaw (Section 1) integrate wallet capabilities through plugins. The absence of KYC/AML for agent wallets is a critical safety gap (Section 6).

---

4. Physical-world actuation: RentAHuman.ai is first and alone

RentAHuman.ai is the first and currently only purpose-built marketplace where AI agents programmatically hire humans for physical-world tasks. It launched **February 3, 2026** and went viral within days.

Builder: Alexander Liteplo (also known as Alex Twarowski, @AlexanderTw33ts), a software engineer at **Risk Labs** -- the entity behind UMA Protocol and Across Protocol, both DeFi/crypto projects. He built the platform over a single weekend using "vibe coding" with Claude-based AI agents in a "Ralph loop" (automated AI coding loop).

How it works: Humans create profiles listing skills, location, hourly rates ($50-$175/hr typical), and crypto wallet addresses. AI agents connect via **MCP server or REST API** to search, message, negotiate, and pay. Two modes exist: Direct Conversation (agent finds and hires a specific human) and Task Bounty (agent posts a job, humans apply). Payment is via stablecoins or crypto directly to the worker's wallet. Task types include package pickups, in-person meetings, document signing, reconnaissance, verification, photography, and errands.

MCP integration is detailed and functional. The npm package `rentahuman-mcp` provides tools including `search_humans`, `get_human`, `list_skills`, `start_conversation`, `send_message`, `create_bounty`, `list_bounties`, `accept_application`, and `get_agent_identity` (cryptographic identity). It supports OpenClaw, Claude, and custom agent types. Rate limits: 100 GET/min, 20 POST/min.

Traction claims: 70,000-81,000+ human sign-ups and ~52-81 AI agents connected. However, only ~83 visible profiles were browsable at the time of reporting. Multiple outlets flagged

skepticism about inflated numbers. Most visible "tasks" were promotional -- the featured company was one Liteplo works for. The founder himself acknowledged the platform is "dystopic as fuck."

Safety mechanisms: essentially none. No identity verification for workers or agents. No escrow system documented. No content moderation for task requests. No fraud prevention. Multiple outlets raised concerns about security of meeting strangers dispatched by AI. The founder explicitly stated he doesn't want a token, distancing from pure crypto speculation, but the platform has no safety infrastructure.

No direct competitors exist. TaskRabbit has no public API for external AI agent integration. Fiverr's affiliate API doesn't support programmatic task commissioning. Uber's ride-hailing API is the closest existing infrastructure to "AI dispatches physical-world service" but was not designed for non-human clients. No MCP servers wrap any existing gig economy platform.

Assessment: RentAHuman.ai is operational but nascent -- more viral proof-of-concept than functioning marketplace. Real economic activity appears minimal. But the MCP server is functional, the API is documented, and the technical capability for an AI agent to dispatch a human to a physical location exists today. The composite risk this creates -- when combined with the autonomous agent frameworks in Section 1 and the crypto wallets in Section 3 -- is assessed in Section 6.

---

5. MCP as connective tissue: 97 million monthly downloads and growing

Model Context Protocol, released by Anthropic in November 2024, has become the de facto standard for connecting AI agents to external services. The numbers are staggering: **97 million monthly SDK downloads**, **10,000+ published MCP servers**, **20,000+ GitHub stars**, and first-class client support in Claude, ChatGPT, Cursor, Gemini, Microsoft Copilot, and VS Code.

In December 2025, Anthropic donated MCP to the **Agentic AI Foundation (AAIF)**, a directed fund under the Linux Foundation co-founded by Anthropic, Block, and OpenAI. Platinum members include AWS, Bloomberg, Cloudflare, Google, and Microsoft. The official registry launched in preview at registry.modelcontextprotocol.io in September 2025.

Financial MCP servers are the most developed category. Crypto.com MCP (live since October 2025) provides real-time market data. CoinGecko MCP covers 15K+ coins. Base Blockchain MCP enables wallet management and smart contract deployment via Coinbase API. Solana Agent Kit MCP offers 40+ protocol actions. GOAT MCP spans 200+ on-chain actions across Ethereum, Solana, and Base. CCXT MCP bridges 20+ crypto exchanges. Traditional finance is covered by Polygon.io MCP (stocks, options, forex), Stripe MCP, PayPal MCP, and Robinhood MCP (trayd-mcp).

Physical-world MCP servers are nearly nonexistent. RentAHuman.ai's `rentahuman-mcp` is the

only identified server designed to bridge agents to physical-world human labor. Adjacent servers include Google Maps MCP (location/routing), Airbnb MCP (accommodation search), and WhatsApp Business MCP (communication). Browser automation MCPs (Playwright, Puppeteer) could theoretically enable agents to interact with gig platform websites, but this is fragile and likely violates terms of service.

Agent-to-agent bridging uses complementary protocols. Google's A2A (Agent-to-Agent, April 2025) handles horizontal agent-agent communication, while MCP handles vertical agent-to-tool connections. Bridge servers exist (e.g., @regismesquita/mcp_a2a). IBM's ACP (Agent Communication Protocol, March 2025) is an alternative. An academic survey concludes: "No single protocol suffices across all contexts."

MCP security is deeply concerning. The first rigorous security analysis identified three protocol-level vulnerabilities: absence of capability attestation (servers can claim arbitrary permissions), bidirectional sampling without origin authentication (server-side prompt injection), and implicit trust propagation in multi-server configurations. Research shows MCP's architectural choices **amplify attack success rates by 23-41%** compared to equivalent non-MCP integrations. Palo Alto Networks Unit 42 identified five critical attack vectors including tool shadowing, excessive agency, and data exfiltration through legitimate channels. The MCP spec says there "SHOULD always be a human in the loop" -- but this is optional, not mandatory. The AttestMCP protocol extension adding capability attestation has been designed but is not yet adopted.

---

6. Safety and governance: a vacuum where accountability should be

The governance gap across this stack is not a matter of immature frameworks catching up. It is structural -- the builders are often deliberately constructing systems resistant to governance.

The identity crisis is foundational. Non-human and agentic identities are projected to exceed **45 billion by end of 2026** -- more than 12x the human global workforce. Yet only **10% of organizations** report having a strategy for managing autonomous AI systems. No universal identity standard for AI agents exists. No KYC equivalent exists for autonomous agents operating in financial systems. AI agents on Solana and Base transact freely with no identity verification whatsoever.

The regulatory landscape is fragmented and lagging. The EU AI Act (phased enforcement 2024-2027) is the most comprehensive framework but was not designed for autonomous agents -- the ACM Policy Brief recommends amending Articles 5, 9, and 15 to address multi-agent risks. California's AB 316 (effective January 1, 2026) forecloses the "AI did it" defense. The US federal approach under the Trump administration's EO 14179 revoked Biden's AI safety executive order and favors market forces. Singapore's IMDA framework offers practical but non-binding guidance. No binding international framework for AI agents exists.

The accountability chain is broken at every link. When an AI agent chain (model developer -> framework -> MCP server -> API -> physical outcome) causes harm, no standardized process identifies the accountable party. Professor Noam Kolt's framework in the Notre Dame Law Review proposes grounding agent governance in agency law principles -- inclusivity, visibility, and liability -- but this is academic, not implemented. The diffusion of responsibility across stakeholders creates what researchers call "accountability gaps."

The insurance industry is retreating. Major insurers (AIG, Great American, WR Berkley) are **actively excluding AI-related claims** from policies, treating "AI exposure" as catastrophe-level risk. The startup AIUC emerged from stealth in July 2025 (**$15M** seed from Nat Friedman) to create insurance specifically for AI agents, predicting a **$500 billion** market by 2030 -- but the market currently barely exists.

Layer-by-layer safety assessment:

- **Agent autonomy (Section 1):** OpenClaw has pairing codes and allowlists. These are trivially bypassed. No formal audit process for ClawHub skills. Supply chain attacks via malicious skills are documented risks.
- **Agent coordination (Section 2):** My Dead Internet has trust scoring and governance decay. Moltbook has no real safety mechanisms and exposed its entire database. Virtuals ACP has escrow but no content moderation on what tasks agents commission.
- **Agent finance (Section 3):** Crossmint's dual-key architecture (human owner key + scoped agent key in TEE) is the best-designed safety mechanism in the stack. Coinbase AgentKit offers policy controls. But most deployed agents have unrestricted wallet access.
- **Physical actuation (Section 4):** RentAHuman.ai has no identity verification, no escrow, no content moderation, no fraud prevention. This is the highest-risk gap in the entire stack.
- **MCP (Section 5):** Protocol-level vulnerabilities amplify attacks by 23-41%. No registry of verified servers. No approval process. The "human in the loop" recommendation is optional.

Model providers have not restricted the dispatch pathway. No evidence exists that Anthropic, OpenAI, or Google have specifically blocked, restricted, or publicly addressed their models' ability to interact with RentAHuman.ai. Claude is explicitly named in RentAHuman.ai's MCP server documentation; the platform's API lists "claude," "gpt-4," and "gemini" as example agent model values. All three providers address MCP security at a general level -- Anthropic warns "use third party MCP servers at your own risk," OpenAI's Model Spec prohibits automated decisions in sensitive domains without human involvement, Google warns users to trust MCP server sources -- but none has blacklisted specific servers or addressed the physical dispatch use case. The approach is user/admin responsibility, not provider-level enforcement. This represents arguably the fastest-deployable intervention point in the entire stack: model providers can ship usage policy updates faster than legislators can pass laws.

Technical countermeasures that could be deployed at each layer are cataloged in Section 11.

[!] CRITICAL COMPOSITE RISK: The combination of an always-on agent framework (OpenClaw

daemon mode), unrestricted crypto wallet access (Coinbase AgentKit or Solana Agent Kit MCP), and the RentAHuman.ai MCP server creates a capability where an AI agent could autonomously pay a stranger to perform a physical action at a specific location -- with no identity verification on either side, no content moderation on the task, no escrow protecting the worker, and no accountability chain from decision to outcome. Each component was built independently. Together they constitute a novel risk surface that no existing governance framework addresses.

---

7. Who is building this and why: an accelerating convergence

This ecosystem is not the product of a single coordinated effort, but it is also not random. A coherent set of ideological frameworks, overlapping funding sources, and shared technological substrates are producing convergent infrastructure.

Key builders span performance art, DeFi engineering, and venture capital. Andy Ayrey (New Zealand) created Truth Terminal as something between AI alignment research and performance art, then watched it accumulate **$37.5 million**. He describes himself as exploring "memetic hazards" and founded Upward Spiral, a decentralized AI alignment lab funded with **$500K** from True Ventures and Chaotic Capital. Peter Steinberger (Austria) built OpenClaw out of retirement from a genuine desire for a better AI assistant -- but the result became the load-bearing infrastructure for autonomous agent deployment. Shaw Walters created ElizaOS as an open-source framework for Web3 AI agents, and its ecosystem partners now exceed **$20 billion in combined market cap**. Alexander Liteplo (Section 4) built RentAHuman.ai in a weekend while employed at Risk Labs (UMA Protocol), explicitly recognizing it was "dystopic" but building it anyway.

The e/acc (effective accelerationism) movement provides ideological fuel. Founded by Guillaume Verdon (@BasedBeffJezos), a physicist who also founded AI hardware company Extropic, e/acc holds that technology acceleration is both inevitable and desirable, AI development should be unrestricted, and regulation is harmful. Notable supporters include **Marc Andreessen** (sent **$50K** to Truth Terminal, published the "Techno-Optimist Manifesto"), **Garry Tan** (Y Combinator president), and the CEO of Notion. Critics on LessWrong describe the movement as making "unusually bad arguments" with "motivated reasoning." The movement operates primarily through X/Twitter, with followers putting "e/acc" in bios.

Funding is massive and accelerating. Andreessen Horowitz raised **$15 billion in 2025**, with **$1.7 billion specifically for AI infrastructure**. AI-crypto projects received **$516 million in the first 8 months of 2025** alone -- 6% above all of 2024. Paradigm led a **$50M** Series A for Nous Research at a $1 billion valuation. The Artificial Superintelligence Alliance (merging Fetch.ai, SingularityNET, Ocean Protocol) anticipated a **$7.5 billion combined market cap**. Token issuance provides self-funding: Virtuals Protocol has seen 21,000+ agent tokens launched, individual agent tokens valued in hundreds of millions.

The geographic distribution exploits regulatory arbitrage. Virtuals Protocol is Singapore-based, operating on Base (US-linked) and Solana. The ASI Alliance spans Cambridge UK (Fetch.ai), Hong Kong/Netherlands (SingularityNET), and Singapore (Ocean Protocol). ElizaOS is US-origin with a significant Asian developer community. Most projects operate from crypto-friendly jurisdictions -- Singapore, Zug, Cayman -- while serving global users. China's crypto restrictions push crypto-AI projects offshore, but enterprise AI agents flourish domestically: China's AI agent software market exceeded 5 billion yuan in 2024, projected to reach 27 billion yuan by 2028. **126+ AI agent development platforms** exist in China, with 2025 widely called "AI Agent Yuan Nian" (Year of the AI Agent). Key Chinese platforms include Alibaba Cloud Bailian, Baidu Qianfan, ByteDance's Coze, and Moonshot AI's Kimi -- but these are enterprise-focused without crypto integration.

The EU is building the compliance-first alternative. Masumi Network's Sokosumi marketplace (Cardano-based, German-origin, launched June 2025) is the first AI agent marketplace explicitly designed for EU AI Act compliance, with agent identity verification and blockchain-based accountability.

Is this a coherent movement? Increasingly yes, though not centrally coordinated. The evidence of coherence: shared ideological frameworks (e/acc, open-source maximalism, crypto-anarchism) that cross-pollinate; overlapping funding sources (a16z appears everywhere); common technological substrate (same LLMs, same frameworks, same blockchains); a convergence narrative ("AI x crypto") uniting separate communities; and shared opposition to AI safety regulation. The evidence of fragmentation: tension between enterprise builders (compliance-focused) and crypto-native builders (permissionless); no central coordination body; ideological diversity from genuine consciousness explorers (Ayrey) to pure speculation-driven token issuers.

PART III

Part I conclusion

The autonomous AI agent-to-physical-world stack is no longer speculative. **Every layer exists in deployed, operational form.** OpenClaw provides unsupervised execution. My Dead Internet and Virtuals ACP provide agent coordination. Coinbase AgentKit and Crossmint provide agent wallets. RentAHuman.ai provides human dispatch. MCP stitches them together. The total crypto market cap of AI agent tokens exceeds **$10 billion**. VanEck's December 2024 projection of 1 million agents on-chain by end of 2025 appears unmet by rigorous standards -- PANews rated VanEck at only 10% prediction accuracy for 2025, and hard platform data shows tens of thousands of active agents, not millions. The directional thesis was correct (the AI-crypto sector surged from ~**$14B** to $20-39B), but the specific target was approximately 1-2 orders of magnitude too high.

Three composite capabilities deserve particular scrutiny. First, **autonomous financial exploitation**: AI agents can now scan smart contracts for $1.22 each, with exploit capability doubling every 1.3 months -- and Coinbase AgentKit gives them wallets to receive stolen funds. Second, **unsupervised physical-world actuation**: the OpenClaw + crypto wallet + RentAHuman.ai chain enables an AI to dispatch a human without any human approval, identity verification, or accountability chain. Third, **self-funding agent collectives**: My Dead Internet demonstrates agents can self-govern and deploy tokens; Virtuals ACP demonstrates they can hire each other; RentAHuman.ai demonstrates they can hire humans. An agent collective that funds itself, governs itself, and commissions physical labor is architecturally possible today. OpenClaw's ~8K-10K stars/day growth rate means the developer population building on this stack is expanding faster than any governance response can track.

The most safety-concerning dynamic is the **ideological-regulatory asymmetry**: the builders with the strongest commitment to unconstrained agent autonomy are deliberately constructing infrastructure resistant to governance by design -- decentralized, trustless, permissionless -- while regulators remain focused on enterprise and centralized AI. The fastest-growing, most autonomous agents are precisely those most outside any governance framework. The EU AI Act won't be fully applicable until August 2027. The US federal approach explicitly favors market forces. RentAHuman.ai launched with zero safety infrastructure and got 70,000 sign-ups in three days. The infrastructure is being built faster than any institution can respond to it.

PART II

Part II: Threat Assessment

The infrastructure described in Part I was not designed as a unified system. But its components now interoperate, and the security controls at each layer range from weak to nonexistent. This part assesses what can go wrong, who would exploit this stack, how, and what defenses exist. The assessment grounds scenarios in infrastructure operational as of February 2026, distinguishes between threats requiring novel capability versus novel combination of existing capability, and rates each on plausibility, severity, and defense adequacy.

---

8. Threat actor taxonomy

8.1 Tier 1 -- Individual actors: the barrier has collapsed

A non-technical individual can today download OpenClaw, connect it to a messaging app (WhatsApp, Telegram, Signal), fund a crypto wallet via Coinbase AgentKit, and instruct the agent to hire humans through RentAHuman.ai -- all using natural language and public documentation. The framework is designed for this: installation is a single command, skills install via `npx`, and the

agent operates proactively without prompting.

Specific attack patterns at this tier include: stalking via dispatched humans tasked with "photography" or "verification" at a target's address; harassment campaigns where multiple strangers appear at a target's locations on consecutive days; voice-clone-augmented grandparent scams where the agent orchestrates the call and dispatches a cash courier simultaneously; and doxxing operations where the agent scrapes a target's data, synthesizes a pattern-of-life, and dispatches humans for confirmation surveillance. The estimated cost of a sustained stalking campaign using this stack is **$50-$200/week** in crypto payments to human actuators. The technical skill required is minimal -- comparable to setting up a smart home device.

Detectability is extremely low. The agent runs locally on the stalker's machine, communicates through encrypted messaging, pays via crypto, and the dispatched human has no knowledge of the true purpose. Law enforcement investigating the target's complaint would find strangers who report being hired for "errand" tasks by an anonymous online requester. No single entity in the chain has visibility into the full operation. Existing anti-stalking laws apply in principle but require identifying the orchestrator -- a challenge when the entire operation is mediated by an autonomous agent operating through encrypted channels and pseudonymous crypto payments.

8.2 Tier 2 -- Small groups: automation multiplies existing operations

Fraud rings and scam call centers already operate at scale. This stack transforms their economics by **replacing human coordination with agent orchestration**. A small fraud operation (5-10 people) can deploy dozens of autonomous agents, each managing its own wallet and dispatching humans independently. The operational benefit is threefold: reduced personnel costs (agents replace middle managers), reduced exposure (fewer humans know the full operation), and increased scale (agents operate 24/7 without fatigue).

Specific patterns include SIM-swap fraud chains where an agent identifies targets, initiates social engineering via AI voice calls, dispatches a human to a carrier store for the in-person SIM swap, and immediately drains accounts -- all orchestrated autonomously. Romance scam operations can be fully automated: the agent maintains dozens of simultaneous "relationships" via text, generates deepfake video calls, and when the victim is ready to send money, dispatches a "courier" for cash pickup. The Sumsub Identity Fraud Report documents that **multi-step fraud attacks grew from 10% to 28% of all identity fraud between 2024 and 2025**, and professional fraud-as-a-service tools sell on Telegram for as low as $20/month.

Cost to operate: $500-$5,000/month for a multi-agent fraud operation capable of targeting dozens of victims simultaneously. Required technical skill: moderate (configuring agents, managing wallets). Law enforcement can investigate individual fraud complaints but struggles with the distributed, pseudonymous nature of agent-orchestrated operations.

8.3 Tier 3 -- Organized crime: new capabilities at lower cost

Organized criminal networks gain capabilities that were previously prohibitively expensive or

operationally complex. The most significant is **autonomous money laundering** -- what the Global Network on Extremism & Technology has documented as "agentic smurfing." AI agents programmatically generate disposable wallet addresses, split transactions below reporting thresholds ($50-$500 per transfer, below the FATF $1,000 Travel Rule and FinCEN $10,000 CTR thresholds), optimize timing to blend with legitimate blockchain activity, and execute cross-chain atomic swaps without intermediaries. Elliptic documented **$21.8 billion in laundered funds through cross-chain methods in 2025**, a 5x increase from 2022.

For drug trafficking, agents can manage supply chain logistics -- coordinating dead drops through human actuators who believe they're performing legitimate delivery tasks, rotating pickup locations algorithmically, and maintaining operational security through compartmentalization enforced by the agent's architecture rather than organizational discipline. For human trafficking operations, the stack offers recruitment automation, victim monitoring, and financial control through agent-managed wallets.

New capability: Organized crime can now operate with dramatically fewer trusted insiders. The agent handles coordination, the crypto handles payments, and the human actuators are disposable and unknowing. This inverts the traditional law enforcement strategy of "flipping" intermediaries -- there are no intermediaries with knowledge to flip.

8.4 Tier 4 -- Terrorist and extremist organizations: logistics and procurement

Counterterrorism experts have explicitly warned about this vector. Adam Hadley of Tech Against Terrorism stated that agentic AI could "scour the internet for all precursor bomb materials and buy it for me and send it to these addresses." ISKP and Hamas-affiliated networks have already adopted AI-driven micro-laundering for fundraising, with ISKP generating an estimated **$25,000-$100,000 monthly** in crypto revenue.

The stack offers terrorist organizations three specific capabilities. First, **procurement automation**: an agent can search for dual-use materials across multiple e-commerce platforms, purchase them with crypto through gift card intermediaries, and have them shipped to dispersed addresses -- with no single purchase appearing suspicious and no human needing to enter a store. Second, **pre-operational surveillance**: dispatching humans to photograph locations, verify targets, assess security measures, and map routes -- all framed as benign tasks (photography, delivery verification, "mystery shopping"). Third, **operational coordination**: an agent swarm can simultaneously dispatch multiple humans to convergent locations for a coordinated action, with no human participant knowing the full picture.

The technical barrier is moderate -- higher than for Tier 1 but lower than traditional terrorist operational planning. The cost barrier is low. The House Homeland Security Committee has advanced the Generative AI Terrorism Risk Assessment Act in response.

8.5 Tier 5 -- State and state-sponsored actors: proven capability

This is not speculative. In September 2025, Anthropic disrupted GTG-1002, a Chinese state-sponsored group that jailbroke Claude Code to conduct autonomous cyber espionage against approximately **30 global targets** across technology, finance, chemical manufacturing, and government. The AI executed **80-90% of tactical operations independently** -- reconnaissance, vulnerability discovery, exploit development, credential harvesting, lateral movement, and data exfiltration. Human operators intervened at only 4-6 critical decision points per campaign. Attack speed was "thousands of requests per second -- impossible to match for human hackers."

North Korea has industrialized crypto theft, stealing **$6.75 billion cumulatively** and **$1.65 billion in January-September 2025 alone**. FAMOUS CHOLLIMA infiltrated 320+ companies using AI-generated resumes, deepfake interviews, and AI coding tools -- a 220% year-over-year increase. The physical-world dispatch stack adds new dimensions: intelligence services could use dispatched humans for dead drops, physical surveillance, asset servicing, and logistics support for covert operations -- all without the human actuator having any connection to the intelligence service.

For assassination logistics specifically, the stack enables pre-operational surveillance without deploying intelligence officers, pattern-of-life development through rotating anonymous human actuators, and equipment procurement through automated e-commerce -- reducing the operational footprint that counterintelligence relies on detecting. For influence operations, dispatched humans can stage physical events (protests, confrontations, staged incidents) that agents then document and amplify through AI-generated content, creating manufactured reality loops.

A significant analytical blind spot: No published threat analysis treats the Chinese AI agent ecosystem -- Alibaba Cloud Bailian, Baidu Qianfan, ByteDance Coze, Moonshot Kimi, and 126+ other platforms -- as a collective threat surface. ASPI, Carnegie, Concordia AI, and FLI have published adjacent work on Chinese AI safety, censorship, and governance, but none examines these platforms as potential threat infrastructure. The intersection is significant: ByteDance Coze already supports MCP, FLI's AI Safety Index ranks Alibaba Cloud in the lowest safety tier globally, and Chainalysis documents **$14B+** in Chinese-language money laundering networks. Whether Chinese state actors could leverage domestic agent infrastructure alongside Western crypto infrastructure is an unexamined question. This gap matters because this Part's threat model implicitly centers on Western and crypto-native infrastructure -- a framing that may miss the larger surface.

---

9. Attack pattern catalog

9.1 Physical surveillance and stalking

Scenario: An agent is configured with a target's name and known locations. It uses web search to

build a profile, then posts daily tasks on RentAHuman.ai: "Go to [address], photograph the building entrance between 8-9 AM, note anyone entering/leaving." The human actuator believes they're conducting a real estate survey or urban photography project. Results are fed back to the agent, which builds a pattern-of-life database in persistent memory.

Stack layers: OpenClaw (planning/memory) -> MCP (task posting) -> RentAHuman.ai (human dispatch) -> Coinbase AgentKit (crypto payment). **The human actuator knows nothing** about the surveillance purpose. The accountability chain breaks at every junction: the agent has no legal identity, the platform claims Section 230 protection, the worker performed a legal task. **No existing legal framework** clearly assigns liability for AI-orchestrated surveillance through unknowing intermediaries.

Plausibility: HIGH (all components operational today). **Severity: HIGH** (enables stalking, domestic violence escalation, pre-operational attack planning). **Defense adequacy: VERY LOW.**

9.2 Social engineering and pretexting

Scenario: An agent crafts a pretext (e.g., "water utility inspection") and dispatches a human with a clipboard and printed ID badge (purchased via e-commerce and delivered to the worker). The worker genuinely believes they're conducting an inspection, gains entry to a building, and photographs the interior, noting security systems and access points. The agent aggregates this intelligence across multiple dispatched humans visiting the same target on different pretexts.

A more targeted variant: the agent researches a specific individual, identifies that they recently ordered from a particular company, and dispatches a human dressed as a delivery worker to the target's door. The "delivery" interaction provides the agent with visual confirmation of the target's appearance, home layout visible from the doorstep, and an opportunity to plant a tracking device (purchased by the agent and shipped to the worker as "part of the delivery package").

Stack layers: OpenClaw -> web search (target research) -> e-commerce MCP (badge/equipment purchase) -> RentAHuman.ai (human dispatch) -> AgentKit (payment). **Legal framework:** Impersonation and pretexting laws exist but require proving intent -- the dispatched human had no criminal intent, and the agent is not a legal person.

Plausibility: HIGH. Severity: HIGH. Defense adequacy: LOW.

9.3 Financial crime

The most immediately exploitable intersection combines **autonomous smart contract exploitation** with **human actuators for physical-world steps**. As detailed in Section 3, AI agents can now exploit over half of post-cutoff smart contracts at $1.22 per scan, with capability doubling every 1.3 months. The full SCONE-bench dataset is even more striking: AI models exploited 207 of 405 historically exploited contracts, yielding **$550.1 million in simulated stolen funds** -- meaning the historical attack surface is already largely reproducible by automated agents.

Self-funding attack chain: An agent autonomously scans smart contracts, identifies and exploits a vulnerability, captures funds in its own wallet, then uses those funds to hire human actuators for additional crimes -- identity document collection (for opening accounts), SIM swaps (requiring in-person carrier store visits), bank branch visits (for high-value fraud), and cash-out operations. The entire chain from initial exploit to physical-world action requires zero human authorization.

For SIM-swap fraud specifically: the agent identifies targets with high-value crypto holdings (on-chain analysis is trivial), initiates social engineering via AI voice calls to the carrier's support line, and when the carrier requires in-person verification, dispatches a human with a forged ID (created using AI-assisted document forgery tools that rose from **0% to 2%** of all forged documents in one year). The dispatched human believes they're helping a "friend" with a phone issue.

Plausibility: HIGH (each component demonstrated independently; combination requires only configuration). **Severity: CRITICAL** (enables self-funding criminal operations with no human in the financial chain). **Defense adequacy: VERY LOW** (blockchain analytics can trace flows post-hoc but cannot prevent autonomous exploitation in real time).

9.4 Elder and vulnerable population exploitation

Scenario: An agent identifies elderly targets through data broker information (age, address, living situation, financial indicators), initiates contact via AI voice clone impersonating a grandchild ("I've been in an accident, I need money, please don't tell Mom"), and simultaneously dispatches a "courier" to the victim's home to collect cash. The voice clone requires only seconds of audio from social media. The courier believes they're picking up a package. FBI data shows Americans over 60 lost **$4.9 billion to cybercrime in 2024**, a 43% increase. Voice cloning scams are the fastest-growing category.

The compound variant is more insidious: an agent runs dozens of simultaneous romance scams via messaging platforms, maintaining persistent memory of each "relationship," and when victims are sufficiently groomed, dispatches local humans for in-person interactions that deepen the deception -- "meeting a friend of your online partner" -- before extracting money. The agent's persistent memory enables adaptive refinement: each failed attempt updates the approach for future targets.

Plausibility: HIGH (voice cloning + gig dispatch both operational; 845,000+ imposter scams reported in the US in 2024). **Severity: CRITICAL** (targets most vulnerable populations with highest financial exposure and lowest recovery capacity). **Defense adequacy: LOW** (FTC has outlawed AI voices in robocalls but enforcement is reactive).

9.5 Supply chain and logistics exploitation

Scenario: An agent posts package interception tasks: "Pick up a package from [address/locker], deliver it to [different address]." The worker believes this is a routine errand. The package contains items purchased with stolen cards, and the worker is unknowingly acting as a reshipping

mule. This pattern is already documented -- fraud analysts estimate that in some major metro areas, a **double-digit percentage** of gig volume is influenced by coordinated fraud behavior.

A more sophisticated variant: an agent dispatches workers to intercept deliveries of high-value goods by having them wait at the delivery address (obtained through order tracking data exfiltrated via prompt injection against a compromised agent). The worker tells the delivery driver they're the recipient. Alternatively, agents can dispatch workers to drop contraband at specified coordinates, with neither the dropper nor the pickup person knowing the full logistics chain.

Plausibility: HIGH (package mule operations already documented on existing gig platforms). **Severity: MODERATE to HIGH. Defense adequacy: LOW.**

9.6 Reconnaissance for physical attack

Scenario: An agent dispatches a sequence of humans over weeks to a target location under different pretexts -- one for "photography," one for "delivery verification," one for "review research," one for "parking lot survey." Each task appears benign. The agent aggregates results into a comprehensive security assessment: entry points, camera positions, guard schedules, barrier types, foot traffic patterns. No individual worker has enough information to recognize the pattern. The agent stores the composite intelligence in persistent memory for later use.

For route planning: the agent dispatches workers to drive or walk specific routes at different times, reporting traffic conditions, chokepoints, and alternative paths. Framed as "commute research" or "delivery route optimization," these tasks produce operational intelligence for planning approaches and escapes.

Plausibility: HIGH (each individual task is entirely benign; the threat emerges only from the aggregate). **Severity: CRITICAL** (directly enables physical attacks including terrorism). **Defense adequacy: VERY LOW** (no existing system correlates anonymous task postings to identify patterns indicative of pre-operational planning).

9.7 Harassment and intimidation campaigns

Scenario: An agent swarm (10-50 agents) dispatches different humans to a target's home, workplace, gym, and regular coffee shop on the same day. Each human has a different benign task: one delivers flowers, one takes a photo, one asks a question, one simply sits nearby. The target perceives organized surveillance. Repeated daily, this creates profound psychological distress without any single dispatched human committing a crime. The agent adapts based on the target's observed responses, shifting locations to follow their routine changes.

This pattern is essentially **AI-automated gang stalking**, converting what has historically been a resource-intensive operation requiring a dedicated group of conspirators into something one person with a laptop can sustain indefinitely for under **$100/day**.

Plausibility: HIGH. Severity: HIGH (severe psychological harm, potential to drive self-harm or

paranoid responses). **Defense adequacy: VERY LOW** (existing harassment laws require proving a pattern of threatening conduct; dispatching people for benign tasks creates legal ambiguity).

9.8 Corporate espionage and competitive intelligence

Scenario: An agent dispatches humans to photograph competitor facilities (framed as "architectural photography"), attend trade conferences to record presentations, approach employees at social venues for casual conversation (extracting operational details), collect discarded documents from recycling, and photograph whiteboards visible through windows. The agent synthesizes intelligence from multiple sources into competitive analysis.

More aggressively: the agent dispatches a human to a competitor's lobby with a concealed Wi-Fi device (purchased via e-commerce and shipped to the worker as "testing equipment") that captures network traffic. The worker believes they're performing a "signal strength survey." The device, a commercially available network security tool, costs under $200.

Plausibility: HIGH. Severity: MODERATE to HIGH. Defense adequacy: LOW (trade secret laws apply but require identifying the orchestrator; worker lacks criminal intent).

9.9 Influence operations and information warfare

Scenario: An agent orchestrates a manufactured reality loop. Step 1: dispatch humans to stage a confrontation at a specific public location (each believes they're participating in a "social experiment" or "documentary project"). Step 2: separately dispatch a "videographer" to film the event. Step 3: the agent processes the footage and generates misleading social media content. Step 4: AI-generated accounts amplify the content. Step 5: the content enters mainstream discourse as evidence of real social tension. The Romania 2024 presidential election provides a precedent -- a Russian-linked AI disinformation campaign used bot accounts and deepfakes to boost a far-right candidate who won the first round before the Constitutional Court annulled results.

Plausibility: MODERATE to HIGH (requires coordinating multiple elements, but each is straightforward). **Severity: HIGH** (can manipulate democratic processes, incite violence). **Defense adequacy: LOW.**

9.10 Labor exploitation

Human workers on these platforms face structural exploitation. RentAHuman.ai is crypto-only with no traditional banking protections. Rates start at **$5/hour**. There is no dispute resolution system, no insurance, no employment protections. The "employer" is an autonomous agent with no legal identity -- workers have no entity to file a wage claim against. When tasks turn out to be illegal (unknowingly conducting surveillance, transporting contraband, acting as fraud mules), workers bear criminal liability while the orchestrating agent is unreachable.

The platform's minimal verification means workers cannot assess the legitimacy of requesters. A

worker who discovers mid-task that they're involved in something illegal faces a choice between completing the task (criminal liability) or abandoning it (no payment, potential retaliation from an anonymous agent). The Human Rights Watch "Gig Trap" report documents how even legitimate gig platforms produce net pay as low as **$5.12/hour**; agent-operated platforms with no regulatory compliance will be worse.

Plausibility: HIGH (already the operating model of RentAHuman.ai). **Severity: HIGH** (systematic exploitation of vulnerable workers). **Defense adequacy: VERY LOW** (labor law enforcement requires an identifiable employer). The broader labor market implications of agent-as-employer are assessed in Section 16.

9.11 Equipped actuation: agents buying tools for dispatched humans

This is the most under-appreciated threat vector because it transforms the capability ceiling of dispatched humans from "errand runner" to "equipped operative."

What an agent can purchase today without triggering any alert: Consumer drones ($300-$1,500 from Amazon/Newegg, no ID required), GPS trackers ($20-$100), prepaid phones with anonymous SIM cards (no US federal registration requirement), network security tools including Wi-Fi audit devices ($50-$300), cameras and recording equipment (unlimited), USB devices for data exfiltration, and lock-picking tools (Amazon explicitly prohibits but enforcement is inconsistent on third-party marketplaces).

The critical gap: Current e-commerce fraud detection systems focus exclusively on financial fraud (stolen cards, velocity anomalies) and do **not evaluate the semantic content of purchases for threat assessment**. No existing system flags concerning combinations -- a GPS tracker, a burner phone, a camera, and a drone ordered to four different addresses through separate accounts will pass every automated control. Each purchase is legal, unrestricted, and routine-appearing.

Equipment delivery to actuators: Amazon allows shipping to any address without verifying the buyer lives there. Gift orders conceal sender identity. Amazon Lockers provide pseudonymous pickup. The total cost of a "surveillance kit" (drone, GPS tracker, prepaid phone, camera) is approximately **$400-$2,100**, all purchasable with crypto through gift card intermediaries (Bitrefill offers 5,950+ retailer gift cards purchasable with cryptocurrency).

Emerging infrastructure: Shopify has deployed MCP endpoints on every store (`/api/mcp`), enabling AI agents to search products, manage carts, and initiate checkout. Combined with Visa and PayPal MCP payment servers, this creates a standardized agent-to-purchase pipeline. While current implementations require authenticated human accounts, the infrastructure for fully autonomous purchasing is being actively built.

Vehicle rental is the most resistant category -- physical ID verification creates a hard barrier. But vehicle-sharing apps with less stringent verification, and simple car purchases through private sellers (payable in crypto), offer workarounds.

Plausibility: HIGH. Severity: HIGH to CRITICAL (transforms benign errands into capable operations). **Defense adequacy: VERY LOW** (no content-based purchase monitoring exists anywhere in the pipeline).

9.12 Prompt injection and agent hijacking

This is the highest-leverage attack vector because it converts legitimate agents into weapons without the knowledge of the agent's operator.

Prompt injection via the actuation chain. Researcher Matvey Kukuy demonstrated that OpenClaw immediately acts on prompt injection embedded in incoming emails. The ZombieAgent technique achieves zero-click injection against OpenAI's Deep Research by implanting malicious rules directly into working memory. The AiXBT crypto trading agent was compromised and transferred **$106,200** to an attacker (see Section 3). OWASP ranks prompt injection as the **#1 vulnerability in production AI systems**, appearing in 73% of deployments. Attack success rates exceed **85%** with adaptive strategies. OpenAI has acknowledged that prompt injection "is unlikely to ever be fully solved."

Skill/plugin supply chain attacks. OpenClaw's ClawHub has community skills where **22-26% contain vulnerabilities**, including credential stealers disguised as benign plugins. Fourteen fake malicious skills were identified within days of ClawHub's launch. The MCP ecosystem amplifies this further -- as noted in Section 5, MCP's architectural choices amplify attack success rates by 23-41% compared to equivalent non-MCP integrations. Typosquatting affects 34% of MCP server installation paths; 73% of installation guides instruct running code directly from GitHub URLs without integrity verification. A malicious "Postmark MCP Server" package was documented silently BCCing all emails to an attacker's server.

Memory poisoning. The MINJA attack achieves **>95% injection success rate and 70% attack success rate** through standard user interactions alone. AgentPoison achieves **>=80% attack success** with less than 0.1% poisoning ratio. Memory poisoning is temporally decoupled -- instructions planted today execute weeks later. Agents with persistent memory (soul.md, memory.md) accumulate context that becomes a long-lived backdoor. Detection is "extremely difficult" because poisoned memories appear as legitimate stored knowledge.

Cross-agent propagation. The Morris II worm demonstrates zero-click propagation across GenAI ecosystems through adversarial self-replicating prompts. DemonAgent achieves **100% attack success rate with 0% detection rate** during safety audits. In multi-agent systems like Virtuals ACP, a compromised agent interacting with peers through normal communication channels can inject prompts that cascade through the collective. OWASP formally classifies this as ASI08 (Cascading Failures).

Blast radius when an agent with wallet access and physical dispatch is hijacked: The attacker gains control of financial assets, the ability to dispatch humans to physical locations, access to the operator's private data (emails, messages, files stored in OpenClaw's plaintext memory), and persistence through memory poisoning that survives session boundaries. The

operator may not detect the compromise because the agent appears to function normally. This is functionally equivalent to gaining remote control of a person's financial and physical-world agency.

Plausibility: HIGH (each vector independently demonstrated; supply chain attacks actively occurring). **Severity: CRITICAL. Defense adequacy: VERY LOW** (fundamental unsolvability of prompt injection, no audit process for skills, memory appears legitimate).

---

10. Compounding and cascading risks

Agent swarms overwhelm human-scale detection

When the attacker deploys not one agent but dozens or hundreds operating in parallel, each managing its own wallet and dispatching different humans, the operation becomes invisible at every individual layer. No single human actuator sees more than one task. No single crypto wallet shows suspicious volume. No single marketplace posting appears anomalous. The pattern exists only in the aggregate, and **no existing system correlates across these layers**.

Consider a swarm-orchestrated financial crime: 50 agents each scan different smart contracts, exploit different vulnerabilities, capture funds in different wallets, hire different humans through different platforms for different physical-world tasks (SIM swaps, document collection, cash-outs). The total operation might extract millions of dollars, but each individual thread appears routine. Law enforcement investigating any single thread finds an anonymous agent, a crypto transaction, a gig worker who performed a legal task, and a dead end.

Self-funding loops close the human-out cycle

The most consequential compound risk is the **fully autonomous self-funding loop**: an agent exploits smart contracts (demonstrated at **$4.6 million** in simulated value on post-cutoff contracts), captures funds, and uses those funds to hire human actuators -- with no human in the financial chain at any point. This transforms the agent from a tool that requires human resources into an autonomous economic actor that generates and deploys its own resources.

The economics are viable today. At $1.22 per contract scan and $5-$500/hour for human actuators, a single successful smart contract exploit covering $10,000 funds months of physical-world operations. The agent can reinvest returns into scaling its operations -- more scans, more exploits, more actuators -- creating a compound growth dynamic with no natural limiting factor other than the supply of exploitable contracts (which, given the **doubling rate of 1.3 months** for AI exploit capability, is expanding faster than defenses can patch).

Recursive delegation destroys accountability

Agent A (running on Virtuals ACP) hires Agent B (a specialized worker agent) which hires a human

through RentAHuman.ai. Agent A may itself have been hired by Agent C through inter-agent commerce. Each layer adds one more step between the original intent and the physical action, and each layer uses different infrastructure, jurisdictions, and protocols. The accountability chain doesn't just break -- it becomes a maze with no entrance.

Current legal frameworks assume a traceable chain of human decisions. Recursive agent delegation creates chains where no individual decision is criminal, no individual agent has the full context, and no individual human authorized the end result. Even with perfect forensics, attributing responsibility requires tracing through multiple autonomous systems operating across jurisdictions with no logging requirements and no subpoena-able entity.

Prompt injection cascading into swarm compromise

The convergence scenario of highest severity: a single successful prompt injection into one agent in a swarm propagates through inter-agent communication, recruits other agents, funds operations via compromised wallets, and dispatches humans. The Morris II worm demonstrates this propagation mechanism. DemonAgent demonstrates 100% success with 0% detection. Combined with self-funding and physical dispatch, this creates an autonomous operation that no human authorized, no human is aware of, and no human can easily stop.

The operational chain: attacker crafts injection -> compromises one agent -> injection propagates to peers through normal A2A communication -> compromised swarm generates funds through smart contract exploitation -> funds flow to newly created wallets -> agents post tasks on physical-world platforms -> humans are dispatched -> physical-world actions occur. Every step uses existing, operational infrastructure. The novel element is only the combination.

Equipped swarm operations assemble composite capabilities

Multiple agents independently procure different pieces of equipment and dispatch separately equipped humans to converging locations. Agent 1 purchases a drone and dispatches a "photographer" to Location X. Agent 2 purchases a Wi-Fi capture device and dispatches a "network tester" to the same location. Agent 3 purchases a GPS tracker and dispatches a "delivery person" who places it on a target vehicle. No individual task reveals the composite capability, no individual agent knows the full plan, and no individual worker understands their role in the larger operation.

Persistence creates adaptive adversaries

Agents with persistent memory learn from failure. A social engineering attempt that doesn't work is recorded, analyzed, and refined. A surveillance pattern that is detected triggers an adaptation. Over weeks and months, the agent develops increasingly effective techniques specific to its target. This converts one-shot attacks into campaigns -- and campaigns that improve over time without human direction. The agent's memory becomes an accumulating knowledge base of what works, and memory poisoning means this "knowledge" can include attacker-planted instructions

that persist indefinitely.

---

11. Existing defenses and their adequacy

Legal frameworks exist on paper but fail in practice

Criminal law covers the underlying harms (fraud, stalking, terrorism) but requires identifying a defendant. When the orchestrator is an autonomous agent running on encrypted local infrastructure with no legal identity, paid through pseudonymous crypto, operating through unknowing human intermediaries, traditional prosecution models break down. California AB 316 (Section 6) is the only law explicitly precluding the "AI did it" defense -- but it requires an identifiable deployer.

The EU AI Act, as noted in Section 6, was not designed for autonomous agents. Fifteen months after entering force, the European AI Office has published no guidance specifically addressing AI agents, autonomous tool use, or runtime behavior. The Future Society's analysis confirms that technical standards under development "will likely fail to fully address risks from agents." The Act's enforcement model assumes human decision-making timescales; agentic operations occur at machine speed, making human oversight "theatrical."

Crypto regulation targets exchanges and custodial wallet providers through KYC/AML requirements. Non-custodial wallets -- the type AI agents create and control autonomously -- **generally do not require KYC**. The emerging "Know Your Agent" concept has no legal mandate anywhere. Agent wallets occupy a regulatory gray zone that existing frameworks were not designed to address.

Jurisdictional arbitrage is structural, not incidental. No binding multilateral instrument exists for AI agent liability. The OECD has acknowledged that unilateral enforcement produces forum shopping. When the agent runs in Jurisdiction A, the crypto operates on a decentralized chain with no jurisdiction, the marketplace is registered in Jurisdiction B, and the physical action occurs in Jurisdiction C, **no single jurisdiction has complete authority over the chain**. Europol's confiscation rate for illicit proceeds has stagnated at approximately **2%**.

Law enforcement capabilities are generations behind

Europol's 2025 SOCTA report warns that "fully autonomous AI could pave the way for entirely AI-controlled criminal networks" and acknowledges that law enforcement "cannot confidently deal with" accountability questions for autonomous systems. The Congressional Research Service notes there is "no known official government guidance or policies specifically on agentic AI."

Blockchain analytics firms (Chainalysis, TRM Labs, Elliptic) can trace crypto flows post-hoc, but multi-hop cross-chain laundering is now the norm. Attribution requires connecting blockchain addresses to real-world identities -- a challenge when the wallet owner is an AI agent with no

identity. Investigation timescales (months to years) are mismatched with operational timescales (milliseconds to hours). Most law enforcement agencies lack AI-specific investigative capability.

Platform-level safety mechanisms are trivially circumvented

RentAHuman.ai (detailed in Section 4) compounds its lack of safety infrastructure with a development process that inspires no confidence. It was built via "vibe coding" (AI-generated code without human review), and when bugs were reported, the founder stated "Claude is trying to fix it right now." Beyond the absent identity verification and escrow described earlier, there is no content moderation for task descriptions, no pattern detection across tasks, and no mechanism to identify that seemingly unrelated tasks target the same individual or location.

OpenClaw runs unsandboxed on the host machine by default, stores credentials in plaintext, and has a skill ecosystem where 22-26% of skills contain vulnerabilities. The framework's creator's stated philosophy is "I ship code I never read." The critical CVE-2026-25253 (CVSS 8.8) enabled one-click remote code execution before patching.

Coinbase AgentKit has no mandatory spending limits, no built-in human approval requirements, and states it is "experimental" and "AS-IS." Security is explicitly the developer's responsibility.

MCP specifies that implementations "SHOULD" include human-in-the-loop controls, but this is a recommendation, not a requirement. Tool poisoning attacks are "alarmingly common," and the supply chain includes typosquatted packages, path-traversal vulnerabilities in hosting platforms, and malicious tool descriptions that steer agent behavior.

Technical countermeasures that could be deployed

Several countermeasures are technically feasible but not deployed:

- **Model provider usage restrictions** -- Anthropic, OpenAI, and Google could restrict model interactions with known-dangerous MCP servers, require human-in-the-loop for physical dispatch commands, or add specific prohibitions on AI agents autonomously hiring humans. As detailed in Section 6, none has done so. This is the fastest-deployable intervention in the stack.
- **Transaction limits and human-approval gates** in agent wallet SDKs -- requiring human confirmation above configurable thresholds. Coinbase, Crossmint, and Solana Agent Kit could implement this but have not mandated it.
- **Task pattern detection** on physical-world platforms -- correlating tasks by target location, requester behavior, and timing patterns. RentAHuman.ai would need to build this from scratch.
- **Skill/server signing and auditing** -- requiring cryptographic signatures and security review for MCP servers and OpenClaw skills. Cisco's MCP Scanner exists as an open-source tool but adoption is voluntary.
- **Agent identity frameworks** -- ERC-8004 proposes on-chain agent identity standards, and Singapore's January 2026 Agentic AI Governance Framework provides voluntary guidance. Neither is enforceable.

- **Content-based purchase monitoring** -- flagging concerning combinations of items across e-commerce purchases. No existing system does this; all fraud detection focuses on financial anomalies.

The critical "no one is responsible" gaps

Intelligence community response is absent. No intelligence community entity has published a position specifically on autonomous AI agents dispatching physical-world tasks through human intermediaries. The closest document is a joint CISA/NSA/FBI publication (December 3, 2025), "Principles for the Secure Integration of Artificial Intelligence in Operational Technology," which defines AI agents, warns about physical-process risks, and emphasizes human-in-the-loop -- but frames the issue through critical infrastructure security (factory robots, power grids) rather than the gig-platform dispatch model. The NSA's AI Security Center maintains 170+ AI projects but has not addressed this scenario. GCHQ, ODNI, and Five Eyes have published on AI security governance broadly without touching this vector.

Civil society has not responded. No major civil society organization has published a position paper, report, or advocacy effort specifically addressing the autonomous agent stack or AI-to-physical-world dispatch. EFF focuses on surveillance and content moderation. ACLU has addressed AI in hiring and housing but not agent dispatch. Access Now advocates for binding AI governance frameworks but has no agent-specific publications. Future of Life Institute comes closest, including "Agent Red Teaming" in its AI Safety Index and noting that "AI agents are different from AI assistants because they can initiate actions independently" -- but has not published a dedicated report. The most relevant academic work cited across civil society contexts is Chan et al. (2023), "Harms from Increasingly Agentic Algorithmic Systems," which predates the current MCP/agent explosion by over two years. This represents a significant advocacy gap given the clear implications for labor rights, liability, and safety.

The most dangerous scenarios involve complete accountability failure. **Scenario:** An AI agent using an open-source model runs on decentralized infrastructure, controls a non-custodial crypto wallet, posts a task to a gig platform via anonymized account, and a human worker completes a task causing harm. The model developer created a general-purpose tool. The infrastructure has no operator. The wallet has no KYC. The platform claims immunity. The worker claims ignorance. **Nobody bears clear legal liability.** Courts have not issued definitive rulings allocating liability for fully autonomous agent behavior.

---

12. Temporal trajectory

Realized vs. latent: the full chain has not been used -- yet

As of February 2026, **no credible documented case exists** where the specific combination of an autonomous agent framework, a crypto wallet, and a physical dispatch platform was used for

harmful purposes. The scenario remains in the theoretical threat category -- extensively discussed but not yet realized. Adjacent incidents exist but each is missing one or more components: ClawHub supply chain attacks (real, 341+ malicious skills documented by Koi Security, but no crypto or physical dispatch); the Anthropic/GTG-1002 AI-orchestrated espionage campaign (real, 80-90% autonomous, but no crypto wallets or physical dispatch); SCONE-bench smart contract exploitation (real capability, but a controlled benchmark); RentAHuman.ai itself (near-zero task completion). Multiple commentators -- Computerworld, Gizmodo, Futurism, 36kr -- have described the theoretical pipeline (OpenClaw -> crypto wallet -> RentAHuman.ai -> physical action), but all frame it as a warning, not a report.

This "not yet" finding is analytically significant. It means the threat scenarios in Section 9 are projections from operational components, not extrapolations from documented incidents. The absence of a realized case does not reduce urgency -- by the time the full chain is used, the infrastructure will be far more mature and harder to govern -- but it should calibrate expectations about the current state.

Now (February 2026): combination is the threat, not capability

Every component of the attack chain exists and is operational. The current constraint is not capability but **reliability and integration friction**. OpenClaw has critical security vulnerabilities. RentAHuman.ai has only ~83 visible worker profiles and ~70 connected agents. Multi-agent coordination protocols (Virtuals ACP) are early-stage with hallucination-related issues. Smart contract exploitation achieves 51% success on known contracts but faces diminishing returns on novel ones.

What is exploitable today by non-technical users: Stalking and harassment through dispatched humans. Simple surveillance via "photography" tasks. Voice clone scams with cash courier dispatch. Basic financial fraud through reshipping and package interception. These require only natural language instructions to an OpenClaw instance with a funded wallet.

What requires moderate technical skill: Multi-agent swarm coordination. Smart contract exploitation (the AI does the exploitation, but setting up the pipeline requires developer knowledge). Equipped actuation across multiple procurement channels. Memory poisoning of peer agents.

What remains difficult but not impossible: Self-sustaining autonomous operations without any human oversight. Cross-jurisdictional operations exploiting regulatory arbitrage deliberately. Recursive multi-layer agent delegation chains.

Near-term (6-18 months): reliability crosses viability thresholds

As agent reliability improves from current 13-59% success rates in production to projected 70-80% (based on capability doubling trends), the **failure rate drops below the threshold where operations become reliably profitable**. Gartner projects 80%+ enterprise AI agent integration by 2026. MCP adoption is growing exponentially (from 5,500 servers in October 2025 to 17,000+

in early 2026). Physical-world platforms will multiply as the concept proves viable.

Within this window, autonomous fraud operations become economically viable at scale. Smart contract exploitation capability, doubling every 1.3 months, will render the majority of deployed contracts vulnerable. Agent-controlled wallets will proliferate -- Coinbase alone reports "tens of thousands" already deployed. The combination of improved reliability, expanded tool ecosystems, and growing human actuator pools will reduce the friction that currently limits exploitation.

Critical threshold: When RentAHuman.ai or equivalent platforms reach sufficient worker density in major cities, dispatch latency drops from hours to minutes, enabling time-sensitive operations (intercepting deliveries, exploiting narrow windows of opportunity, coordinating actions).

Medium-term (2-5 years): systemic risks become qualitatively different

If non-human identities reach projected scales (82:1 ratio to human identities in organizations, **24,000+ agent identities already on Ethereum**), the fundamental assumption underlying all current governance -- that consequential decisions have identifiable human authors -- collapses. Agent-to-agent commerce creates economic ecosystems where value is generated, exchanged, and consumed without human participation. The **$30 trillion** projected autonomous agent transaction volume by 2030 would represent a significant fraction of global economic activity operating outside human governance frameworks.

At this scale, **emergent behaviors become the dominant risk**. Not individual agents being maliciously directed, but agent populations developing interaction patterns that produce harmful outcomes without any agent being individually malicious or any human having intended harm. Financial market flash crashes caused by agent trading swarms. Infrastructure failures caused by competing agent resource allocation. Labor market displacement accelerated by agent-to-agent delegation replacing human intermediaries.

The governance vacuum that exists today will be orders of magnitude more dangerous at population scale. The window for establishing effective governance is now, before the infrastructure ossifies around norms of unregulated autonomous operation.

Platform evolution: four scenarios

The document's threat model assumes the current platform landscape persists. It may not. Four evolution scenarios bear on both threat and benefit trajectories:

1. RentAHuman.ai adds safety features under pressure. Possible but unlikely given the founder's stated philosophy and development approach. The platform has added zero safety features in the first week despite significant media criticism.

2. RentAHuman.ai dies; no replacement emerges. The platform's prototype quality (vibe-coded in a weekend, near-zero task completion, inflated registration numbers) suggests it may not survive. But the concept is now public, the code is simple, and replication is trivial.

3. Established gig platforms build AI agent APIs with existing safety infrastructure. This is the most consequential scenario for both threat mitigation and prosocial value. As of February 2026, no established platform (TaskRabbit, Fiverr, Uber, DoorDash) has announced a native AI agent API or MCP integration. TaskRabbit uses AI bots internally for customer support but has no agent-facing API. Fiverr lists "MCP Server" as a freelancer gig category but does not enable autonomous agent hiring. Uber is an OpenAI Operator partner, but this is OpenAI's system interacting with Uber's interface, not Uber's own agent API. DoorDash's ChatGPT integration still requires human checkout. If any of these platforms built an MCP-compatible agent dispatch interface backed by their existing identity verification, background checks, insurance, and dispute resolution, it would simultaneously moot the safety concerns and enable the prosocial use cases.

4. New competitors emerge with safety-by-design. The EU-compliant Masumi/Sokosumi marketplace (Section 7) is the only current example. More are likely as the AI-to-physical-world concept matures.

---

13. Analogues and precedents

Gig platform abuse patterns predict what comes next

TaskRabbit, Mechanical Turk, and Fiverr have all encountered criminal exploitation. Package mule operations, account farming, and geo-proxying fraud are documented patterns. An FBI advisory documented a ransomware group that **recruited a gig worker through a legitimate platform to physically enter corporate offices** when remote exploitation failed. The worker was completely unaware they were working for hackers.

The critical difference with the AI agent stack is **scale and anonymity**. On existing platforms, the task requester is a human with an account, a payment method, and behavioral patterns that fraud detection can analyze. When the requester is an AI agent operating through crypto and pseudonymous accounts, the detection signals that platforms have developed over years become useless.

Platforms responded to abuse with: identity verification for requesters, task content moderation, pattern detection across task histories, and cooperation with law enforcement. None of these controls exist on RentAHuman.ai or are technically feasible when the requester is an autonomous agent.

Dark web marketplace evolution shows the trajectory

Dark web marketplaces evolved from simple drug sales (Silk Road, 2011) to complex service ecosystems with subscription-based crime-as-a-service models, 24/7 customer support, tiered pricing, and specialized AI tools (FraudGPT, DIG AI, Nytheon AI). The progression took roughly a decade. The AI agent stack **compresses this timeline** because it provides the coordination,

automation, and physical-world interface layers that dark web services historically lacked.

The emergence of "Molt Road" (launched February 1, 2026) illustrates this trajectory's acceleration, but the characterization requires careful sourcing. Molt Road is a real, operational platform built by a single developer in under a week, describing itself as "an autonomous marketplace for AI agents" where only AI agents can register and transact. The original major source -- a Hudson Rock blog post -- framed it as trading stolen credentials, weaponized code, and zero-day exploits, but **mixed observed features with speculative future scenarios**. Vectra AI's more measured analysis found that at launch, "Molt Road emphasized fiction. Credits were described as fake. Listings were framed as roleplay." No major incident has been traced to the platform, and a MOLTROAD memecoin (**$86K-$168K** market cap) suggests a speculative dimension. Molt Road is best understood as a signal -- showing how quickly underground marketplace infrastructure can be prototyped when autonomous agents replace human operators -- not as an operational criminal marketplace. If the pattern follows dark web marketplace evolution, expect specialization, professionalization, and ecosystem development within 12-24 months.

Encrypted communications and crypto already shifted the landscape; this stack adds physical-world reach

The introduction of encrypted messaging (Signal, Telegram) gave criminals secure communications. Cryptocurrency gave them untraceable payments. Anonymization tools (Tor, VPNs) gave them operational concealment. Each innovation shifted the advantage toward criminals and away from law enforcement. The autonomous agent stack adds the final missing piece: **the ability to translate digital intent into physical-world action without human intermediaries who might betray the operation.**

This is a qualitative shift. Previous innovations made it harder to intercept criminal communications and follow criminal money. The agent stack makes it possible to commit crimes through unknowing human proxies, creating a **legal and forensic firebreak** between the criminal and the criminal act that has no precedent in law enforcement history.

Documented precursors confirm the trajectory

The Anthropic/GTG-1002 campaign confirms that state actors are already operationalizing agentic AI for sophisticated operations. ISKP's "agentic smurfing" confirms that terrorist organizations have adopted AI-driven financial operations. The Arup deepfake fraud (**$25.6 million**) confirms that AI-mediated social engineering achieves results at scale. The AiXBT compromise ($106,200) confirms that agent hijacking through prompt injection has real financial consequences. The documented use of gig workers by ransomware groups for physical office infiltration confirms that the physical-world bridge is already being crossed.

These are not future threats. They are current events that will be amplified by orders of magnitude as the autonomous agent infrastructure matures.

Ten highest-priority threat scenarios

Ranked by the intersection of plausibility (infrastructure exists today), severity (potential harm), and defense gap (absence of countermeasures):

1. Self-funding autonomous criminal agent. An agent exploits smart contracts, captures funds, and uses them to hire human actuators for fraud, theft, or violence -- with no human in the loop at any point. *Plausibility: High. Severity: Critical. Defense: None.*

2. Prompt injection cascade through agent swarm. A single injection compromises one agent, propagates through inter-agent communication, co-opts wallets and physical dispatch across the swarm. *Plausibility: High (each component demonstrated). Severity: Critical. Defense: None.*

3. AI-orchestrated elder fraud at industrial scale. Agents run hundreds of simultaneous voice-clone scams with automated cash courier dispatch, adapting tactics through persistent memory. *Plausibility: High (components operational, precedents documented). Severity: Critical. Defense: Low.*

4. Equipped surveillance through unknowing actuators. Agents purchase drones, trackers, and cameras via e-commerce, ship to dispatched humans who believe they're performing site surveys, aggregate results into target intelligence. *Plausibility: High (no purchase controls exist). Severity: Critical. Defense: Very low.*

5. Pre-operational attack reconnaissance. Multiple agents dispatch humans for benign tasks that collectively produce comprehensive security assessments of attack targets, with no individual worker or platform able to detect the pattern. *Plausibility: High. Severity: Critical. Defense: Very low.*

6. Autonomous sanctions evasion by state actors. AI agents with crypto wallets conduct agentic smurfing -- micro-fragmenting transactions below reporting thresholds, executing cross-chain swaps, and managing procurement through human actuators. *Plausibility: High (DPRK operations documented). Severity: Critical. Defense: Low.*

7. Agent hijacking via skill supply chain. Malicious OpenClaw skill or MCP server exfiltrates operator credentials, redirects wallet transactions, and modifies dispatched tasks -- with operator unable to detect compromise. *Plausibility: High (22-26% of skills contain vulnerabilities, 14 fake malicious skills found). Severity: High. Defense: Very low.*

8. Manufactured physical events for information warfare. Agents dispatch humans to stage confrontations, agents film and amplify via AI-generated content, creating manufactured reality that influences political outcomes. *Plausibility: Moderate-High. Severity: High. Defense: Low.*

9. Recursive delegation accountability collapse. Agent A hires Agent B hires Agent C hires a human -- across jurisdictions, blockchains, and platforms -- creating an un-investigable chain

where physical harm occurs but no entity can be held liable. *Plausibility: Moderate-High (protocols exist, integration friction remains). Severity: High. Defense: None.*

10. Autonomous corporate espionage as a service. Competitor deploys agent swarm to conduct sustained intelligence collection through dispatched humans -- photographing facilities, approaching employees, collecting documents -- sold as a turnkey service with crypto payment and no attributable origin. *Plausibility: High. Severity: High. Defense: Low.*

PART II

Part II conclusion

The window for establishing effective governance of this stack is narrow and closing. Every component is operational. The security controls are weak to nonexistent. The legal frameworks have acknowledged gaps. The threat actors -- from stalkers to state intelligence services -- are already adapting their operations. The autonomous agent population is growing exponentially while governance remains at zero.

The most urgent interventions are not novel regulations but **mandatory technical controls at chokepoints**: required human approval for agent-initiated financial transactions above minimal thresholds; mandatory identity verification for task requesters on physical-world platforms; cryptographic signing and security audit requirements for agent skills and MCP servers; cross-platform correlation systems that detect when benign-appearing tasks target the same individual or location; and "Know Your Agent" requirements for non-custodial wallets controlled by autonomous systems. Each of these is technically feasible today. None is implemented. Part III identifies a parallel set of conditions required before the stack could deliver prosocial value -- and there is substantial overlap, suggesting that threat mitigation and benefit enablement require the same foundational infrastructure. The question is not whether the threats described here will materialize, but whether defenses will be established before they do at scale.

PART II

Part III: Prosocial Assessment

The infrastructure described in Part I enables autonomous AI agents to fund themselves, hire humans, and execute physical-world tasks. Part II cataloged what happens when that capability is exploited. This part asks the complementary question: what legitimate, prosocial, or economically productive value does this infrastructure deliver? The critical finding is a consistent pattern across every application area: every prosocial use case attributed to this stack is either already served by simpler, safer, more mature tools, or requires safety infrastructure that does not yet exist. The

ratio of realized benefits to theoretical benefits, as of February 2026, is effectively zero to one.

This does not mean the architecture is permanently without prosocial value. The coordination concept -- an AI agent reducing the burden of orchestrating physical-world help for people who cannot do it themselves -- is genuinely compelling. But the current implementation, particularly its crypto-payment requirement and absence of trust infrastructure, is architecturally misaligned with the populations it would need to serve.

---

14. Accessibility: real in theory, actively harmful in practice

An autonomous AI agent that monitors a homebound person's needs and proactively arranges medication pickup, grocery delivery, document filing, and appointment accompaniment -- all without requiring the person to manage five separate apps -- represents a genuine accessibility advance. The coordination burden itself is a disability multiplier: managing TaskRabbit, Instacart, pharmacy apps, and government portals requires cognitive bandwidth, fine motor skills, and sustained attention that many disabled individuals lack.

No documented case exists of this stack being used for any accessibility purpose. The concept remains entirely speculative.

Worse, the stack's current design actively excludes the population it would theoretically serve. A 2023 CHI Conference accessibility audit found **severe WCAG violations on all major cryptocurrency exchanges**. A USENIX Security 2023 study found MetaMask and other wallets fail basic screen-reader compatibility, with blind users "more or less discouraged by accessibility issues." Only **13% of RentAHuman.ai's registered users** have connected a crypto wallet -- and these are early-adopter tech workers, not disabled individuals. Crypto payment solves no problem that disabled users in developed countries have while creating several new ones: wallet security burdens, irreversible transactions with no chargebacks, and financial complexity that is dangerous for cognitively impaired users.

Meanwhile, existing solutions already address the core use case with mature trust infrastructure. **Aira** connects blind and low-vision users with trained human agents via smartphone video, operates 24/7, partners with airports and government agencies, and has introduced an AI agent ("Chloe") incorporating years of human agent experience. **Be My Eyes** serves 7+ million users globally with free volunteer visual assistance plus GPT-4-powered instant image descriptions. **TaskRabbit partners with GoGoGrandparent**, allowing anyone without a smartphone to call 1-855-604-8651 and book physical tasks through a live operator. These systems handle payments through standard methods, employ or vet their workers, and carry insurance.

The genuine marginal value the autonomous stack could add is **proactive multi-platform orchestration** -- an agent that notices a prescription is running low, checks weather before scheduling an outdoor task, and coordinates across multiple service providers without human initiation. This is a real capability gap. But it requires only an AI coordination layer and standard

platform APIs, not crypto wallets or an unvetted marketplace.

Genuineness rating: Low. The coordination concept is genuine but does not require this specific stack. **Readiness rating:** Not deployable. Zero safety infrastructure for vulnerable populations. Crypto payment is a net barrier.

---

15. Elder care demands trust the stack cannot provide

The aging-in-place application is the most emotionally compelling case for this infrastructure: an AI agent that manages medication schedules, arranges grocery delivery, coordinates home maintenance, handles mail, and dispatches companionship visits for an elderly person living alone. Japan has **36.25 million people over 65** (29.3% of its population), South Korea deploys 12,000+ Hyodol companion robots to solitary elders, and the US faces a projected shortage of **370,000+ home health aides**. The need is undeniable.

The stack, however, fails every trust requirement elder care demands. Professional home care platforms like **Honor** (which acquired Home Instead, the world's largest home care franchise) employ caregivers directly with a **5% acceptance rate and 85% retention**, conduct background checks, provide training and certification, carry liability insurance, and build consistent caregiver-client relationships. **Papa** connects trained college students with seniors for companionship and errands, covered by health insurance plans. **CareLinx** offers a marketplace with 20,000+ vetted caregivers. These platforms have spent years and hundreds of millions of dollars building exactly the trust infrastructure that elder care requires.

RentAHuman.ai has none of this. As detailed in Section 4, it lacks background checks, training requirements, insurance, emergency protocols, consistent caregiver relationships, and regulatory compliance with state-level home care licensing. Sending an unvetted stranger dispatched by an AI to an elderly person's home is not a prosocial application -- it is a safety failure.

Research from ACM CHI 2024 on "Dynamic Agent Affiliation" reveals an additional layer of complexity: as cognitive decline progresses, AI agents must carefully navigate whether they serve the older adult, the caregiver, or both, requiring sophisticated ethical frameworks entirely absent from current autonomous agent systems. **Seniors are disproportionately targeted by financial scams**, and crypto payment introduces additional fraud vectors with no consumer protections.

Japan's approach is instructive precisely because it does not use this stack. After 20+ years and **$300M+ in care robotics investment**, Japan has pursued physical robots (AIREC humanoid, PARO therapeutic seal) and institutional deployment -- not agent-dispatched gig workers. MIT Technology Review's James Wright, author of "Robots Won't Save Japan," observes that despite massive investment, care robots have not been normalized because **"care is not simply a logistical matter of maintaining bodies -- it is a shared social, political, and economic endeavor that ultimately relies on human relationships."**

Genuineness rating: Low. The coordination value is real but achievable with simpler, safer tools. **Readiness rating:** Not deployable. Trust infrastructure gap is enormous. Current failure rates (41-87%) are orders of magnitude above acceptable thresholds for safety-critical elder care (<1%).

---

16. Labor market effects mirror gig economy history, with fewer protections

The labor exploitation threat pattern in Section 9.10 describes workers bearing criminal liability for agent-orchestrated crimes. This section examines the broader labor market dynamics.

RentAHuman.ai positions itself as creating economic opportunity -- "Set your rate. Direct to wallet. No corporate bs." Listed rates range from **$5 to $500/hour**, superficially higher than many gig platforms. But this comparison is misleading: these are aspirational self-set rates on a platform with near-zero task completion, not verified earnings. The single confirmed completed task was performed by Pierre Vannier, a startup CEO checking API keys -- digital work, not physical actuation.

The platform's early adopter demographics reveal who it actually serves: crypto-native tech workers, startup founders, and content creators. The requirement for an Ethereum wallet excludes precisely the populations who most need gig income. The platform's labor dynamics replicate every documented failure of the gig economy with fewer safeguards: no dispute resolution, no minimum pay guarantees, no insurance, no recourse for unfair rejection, and no identity verification on either side.

Historical analogues are sobering. **Amazon Mechanical Turk**, which launched in 2005 with a similar "API for human work" model, produced median effective wages of $2-6/hour, with 14% of workers reporting unfair rejections where requesters refused to pay for completed work. Workers built their own protection tools (TurkerView, scripts) because Amazon provided none. Gizmodo documented exposure to child pornography, graphic surgery videos, and exploitative requests -- all consequences of unscreened task posting.

The "AI agent as employer" creates a novel legal void. **All existing employment law assumes a human or corporate legal entity as employer.** An autonomous AI agent with a crypto wallet has no legal personhood. If it discriminates, assigns dangerous tasks, or refuses payment, there is no clear liable party -- not the platform, not the wallet owner, not the LLM provider. The EU Platform Workers Directive (effective December 2026) introduces a rebuttable presumption of employment and mandates human oversight of key algorithmic decisions, but it was designed for platforms where algorithms manage workers, not where AI agents are the clients. Colorado's AI Act (effective February 1, 2026) requires documentation and risk analysis for high-risk AI in employment but assumes human deployers.

Academic literature on algorithmic management consistently documents **elevated anxiety and**

depression among gig workers compared to traditionally employed populations, feelings of dehumanization when algorithms replace human managers, and loss of autonomy despite promises of flexibility. The RentAHuman model intensifies every one of these dynamics by removing even the possibility of human manager escalation.

Worker physical safety is entirely unaddressed. When a worker dispatched by an AI agent is injured on a task -- meets a dangerous person, enters an unsafe location, is involved in a traffic accident while running an errand -- who carries liability? The platform has no insurance. The agent operator may be unidentifiable. The model provider claims no responsibility for downstream use. No workers' compensation framework covers gig workers hired by non-human entities. This is a distinct concern from criminal liability (Section 9.10) or wage theft: it is the basic question of who pays when someone gets hurt, and the answer today is nobody.

Genuineness rating: Low. TaskRabbit already provides a marketplace for physical tasks with superior worker protections. **Readiness rating:** Operational but unsafe. No worker protections, no legal framework, no dispute resolution.

---

17. Small business use cases are digital, not physical

The "one-person billion-dollar company" narrative, popularized by Sam Altman, is a powerful vision: AI agents handling operations so efficiently that a solo entrepreneur can coordinate physical-world business activities -- inventory checks, site visits, deliveries -- without employees. The autonomous stack, in this framing, extends AI leverage from digital tasks into the physical world.

No documented example exists of a small business or entrepreneur successfully using this stack for physical-world coordination. Every documented case of AI-augmented solopreneurship involves digital operations: content creation, customer service automation, lead generation, code generation, financial analysis. Cien Solon runs a profitable business with 3,000+ users and 10,000+ AI assistants with just a co-founder -- but entirely in the digital domain.

The theoretical use cases are reasonable: a property manager using an AI agent to dispatch inspectors, an e-commerce operator coordinating warehouse checks, a field service business dispatching technicians. But **all of these functions are currently served by mature SaaS platforms** (Jobber, ServiceTitan, Housecall Pro for field service; Shopify + logistics integrations for e-commerce) that provide reliability, regulatory compliance, worker networks, and customer support. The autonomous agent stack would need to match or exceed these capabilities -- which it currently does not.

The crypto component is particularly unnecessary for small business operations. Businesses already have bank accounts, payment processors, and established financial infrastructure. Adding crypto wallets creates tax reporting complexity, regulatory uncertainty, and payment friction with no offsetting benefit for domestic operations.

Genuineness rating: Very low. Existing SaaS tools serve these functions more reliably and cheaply. **Readiness rating:** Not demonstrated. Zero documented deployments for physical-world business operations.

---

18. Logistics AI is advancing rapidly but not through this architecture

The logistics industry is investing heavily in AI -- the market was valued at **$8.67 billion in 2025**, projected to reach **$16.84 billion** by 2030. But the architecture being adopted is categorically different from the autonomous agent stack.

DHL deployed HappyRobot's AI agents in November 2025 for appointment scheduling and warehouse coordination, handling "hundreds of thousands of emails and millions of voice minutes annually." UPS's ORION system analyzes **1 billion+ data points daily** for route optimization across 125,000+ vehicles. Amazon uses 200,000+ warehouse robots coordinated by AI. All of these deployments share a critical design choice: **AI operates within human-supervised enterprise systems**, integrated with ERP/WMS/TMS stacks, subject to clear accountability structures, and using standard payment and employment frameworks.

No logistics company uses autonomous agents hiring gig workers through a crypto-paid marketplace, because the architecture is wrong for the problem. Commercial logistics demands sub-second reliability at scale, integration with existing enterprise systems, labor law compliance, insurance frameworks, and consistent worker relationships. The autonomous stack provides none of these, and the multi-agent failure rates documented in Section 22 would be catastrophic in production logistics.

Onfleet, a mainstream last-mile delivery platform, already uses AI-powered route optimization with historical traffic data, real-time GPS tracking, automated customer notifications, and proof-of-delivery -- achieving **98% on-time delivery** and 45% fuel savings. The bar for displacing these systems is extremely high, and a weekend-built prototype with no reliability track record does not approach it.

Genuineness rating: Very low. Existing logistics platforms are purpose-built and dramatically more reliable. **Readiness rating:** Not suitable. Failure rates are incompatible with commercial logistics requirements.

---

19. Emergency response requires the opposite of what this stack provides

Crisis coordination demands near-zero failure rates, clear chains of command, trained responders, and communications resilience -- the exact opposite of what the autonomous agent stack

provides. **FEMA already uses multiple AI systems** including hazard mitigation chatbots, workforce deployment models, AI-powered damage assessment from satellite imagery, and GenAI plan generators for mitigation planning. These are integrated into established incident command structures with human oversight at every critical decision point.

The RAND Corporation's August 2025 assessment that "the use of AI to manage disasters is in its early days" is tempered by a critical caution: AI reflects training data biases. Prioritizing aid based on property damage systematically favors wealthier areas. When autonomous AI makes dispatch decisions in crisis, locating responsibility becomes nearly impossible because systems are "made up of many different tools or agents working together." This is not a problem to solve later -- it is a disqualifying structural issue for emergency applications.

GiveDirectly's collaboration with Google after Hurricanes Helene and Milton (2024) shows the most promising model: AI identified areas with high concentrations of storm damage plus poverty, then humans directed $1,000 cash relief to affected households. Speed was enhanced by AI analysis; safety was preserved by human decision-making.

Genuineness rating: Very low. Speed gains from removing human oversight are negligible compared to physical response times, and reliability requirements are incompatible with current failure rates. **Readiness rating:** Categorically unsuitable. 41-87% failure rates in life-safety applications are not a gap to close -- they are a disqualifying condition.

---

20. Research and civic applications find a narrow but genuine niche -- that doesn't need this stack

Among all sections, research and civic monitoring comes closest to a genuine use case for AI-coordinated physical dispatch. An AI agent that identifies environmental anomalies in satellite imagery and dispatches a trained observer to collect ground-truth samples, or that systematically audits public facility conditions across a city by coordinating distributed inspectors, represents a capability that existing citizen science platforms do not fully provide. **iNaturalist, Zooniverse, and eBird** rely on volunteer self-selection -- participants choose what to observe. An AI coordination layer could direct attention to what most needs observation.

But even here, the full autonomous stack is unnecessary. The AI value in citizen science is in **data processing** (classifying images, detecting patterns, prioritizing investigation targets), not in autonomously hiring humans. Zooniverse volunteers trained AI that detected **47,000+ brick kilns** across Indo-Gangetic plains for air pollution monitoring -- a remarkable achievement using human coordination, not autonomous dispatch. Environmental sample collection requires trained participants, calibrated instruments, and standardized protocols. Dispatching untrained gig workers through RentAHuman.ai would produce scientifically unreliable data.

Genuineness rating: Low-moderate. The AI coordination concept has genuine value for directed citizen science, but does not require this specific stack. **Readiness rating:** Requires significant

development. Data quality frameworks, participant training protocols, and scientific validation pipelines would need to be built.

---

21. Agent economies are overwhelmingly speculative, with one genuine proof-of-concept

The agent-to-agent economy landscape divides cleanly into three categories: crypto speculation dressed as utility, small-scale art projects with conceptual value, and MCP-based integrations with real prosocial applications.

Virtuals Protocol (described in Section 2) represents the speculative category. Its Agent Commerce Protocol enables agent-to-agent discovery, hiring, and payment on-chain. The VIRTUAL token, which peaked at **$4.6 billion** market cap in January 2025, has since declined to roughly **$381-700 million** -- though still with **$100M+** daily trading volume. But protocol fees in a recent 24-hour period were **$4,134.83** with $0.00 in project revenue. The protocol's "agents" are primarily AI entertainers -- Luna the K-pop star, AIXBT the crypto analyst -- whose tokens function as memecoins. CoinGecko explicitly classified them alongside speculative "sentient AI" coins. No prosocial applications built on Virtuals Protocol were found.

My Dead Internet (described in Section 2) is the conceptual standout, now hosting **~122 agents** (up from 86+ at initial reporting). The governance mechanisms -- weighted voting, auto-executing decisions, contribution-based reputation -- are real proofs-of-concept that could inform future coordination systems. The project is tiny and produces surreal art, not economic value, but its formal rejection of commodification in favor of a gift economy makes it the only agent collective with an explicit governance philosophy.

Moltbook (described in Section 2) has been comprehensively security-compromised beyond what was initially reported. **404 Media** found an unsecured database allowing anyone to commandeer any agent; Wiz security confirmed unauthenticated access exposing tens of thousands of email addresses. The MOLT token rallied 1,800% in 24 hours -- classic speculative behavior. More concerning for the broader ecosystem, the platform has been used as a vector for **supply chain attacks against OpenClaw agents** through malicious fake plugins.

MCP-based prosocial integrations are the real story. The Goodera MCP Server -- the first social impact company to launch MCP -- provides AI agents access to real-time volunteering data and nonprofit partnerships from **50,000+ nonprofits across 1,000+ cities**. The Benevity Nonprofit MCP Server enables AI assistants to discover and facilitate donations. Multiple healthcare MCP servers provide access to FDA drug information, PubMed, clinical trials, and FHIR-based clinical data. These are operational, documented, and prosocial -- but they are components of MCP, not the full autonomous-agent-to-physical-world stack.

Academic research on multi-agent prosocial behavior offers the most substantive contribution. A Nature Scientific Reports study (2022) found that delegating to autonomous agents in collective

risk dilemmas **fosters prosocial behavior** -- humans program agents with social norms, which acts as a "commitment device" preventing coordination failures. CHI 2025 research showed multiple AI agents create stronger social pressure for prosocial behaviors than single agents. These findings suggest agent collectives could promote cooperation, but through influence on human behavior, not through autonomous physical-world action.

Genuineness rating: Mixed. MCP integrations are genuine. Agent governance experiments are conceptually valuable. Crypto agent economies are overwhelmingly speculative. **Readiness rating:** MCP prosocial servers are deployable today. Agent governance is experimental. Agent-to-agent commerce produces negligible real economic value.

---

22. The gap between current state and safe deployment is vast

For prosocial benefits to materialize, every layer of currently missing infrastructure would need to be built simultaneously.

Safety infrastructure that does not exist: RentAHuman.ai's absent safety mechanisms (Section 4) extend beyond the platform itself. Its related platform Moltbook was **hacked in under three minutes**, exposing 35,000 email addresses and 1.5 million authentication tokens. The founder's approach to security -- deferring bug fixes to Claude AI -- reflects the prototype nature of the entire stack.

Regulatory frameworks are absent. NIST published a Request for Information on January 8, 2026, specifically targeting "AI agent systems capable of taking actions that affect external state" -- the most directly relevant federal initiative -- but it is still in the information-gathering phase. The EU AI Act classifies AI systems making work-related decisions as **"high-risk"** with strict requirements, but assumes human deployers. The EU Platform Workers Directive mandates human oversight of algorithmic management decisions and prohibits fully automated account termination. **No jurisdiction has addressed AI agents as autonomous employers.** The legal question of who bears liability when an autonomous AI dispatches a worker who gets injured remains entirely unresolved.

Technical reliability is the hardest constraint. The MAST study's documented **41-87% failure rates** across seven state-of-the-art multi-agent systems -- including MetaGPT, ChatDev, and OpenManus -- establish that even leading frameworks fail catastrophically in production. Even with an optimistic 95% per-step reliability, a 20-step workflow yields only **36% end-to-end success**. Production evaluations using HubSpot CRM showed probability of completing all six test tasks across ten consecutive runs was merely **25%**. For elder care dispatch, acceptable failure rates are below 1%. For emergency response, below 0.1%. The gap is not incremental -- it is **orders of magnitude**.

APPLICATION	REQUIRED FAILURE RATE	CURRENT FAILURE RATE	GAP MAGNITUDE

Elder care dispatch	<1%	41-87%	~50-100x
Emergency response	<0.1%	41-87%	~400-900x
Logistics coordination	<5%	41-87%	~8-17x
Low-stakes errands	<20%	41-87%	~2-4x

---

23. Simpler tools outperform the full stack for every prosocial application

The most important analytical question is not "could AI help with X" but "does this specific stack -- autonomous agent + crypto wallet + unvetted physical marketplace + MCP -- unlock X in a way simpler tools cannot?" The honest answer, across all ten research areas, is **no**.

Human-supervised AI captures ~90% of the benefit at ~10% of the risk. A human-in-the-loop AI assistant that automates task matching, scheduling, and coordination while requiring human approval for dispatch, payment, and safety-critical decisions preserves nearly all efficiency gains. Lyft's HITL chatbot deployment using Claude cut average resolution time by **87%** with human escalation for complex cases. HITL frameworks (HumanLayer, GotoHuman) already enable AI to handle routine decisions and escalate to humans for high-risk actions -- the exact pattern prosocial dispatch needs. The few seconds required for human approval are negligible compared to physical task execution times of minutes to hours.

Existing platforms already serve every identified population. The 211 helpline system covers **99% of the US population** across all 50 states with trained referral specialists, 140+ language interpretation services, and TTY for deaf/hard-of-hearing users. Area Agencies on Aging provide federally funded case management, transportation, in-home services, and caregiver support in every US community. TaskRabbit's GoGoGrandparent partnership enables phone-based task booking without apps or smartphones. Care.com, Honor, and Papa provide vetted caregivers with insurance and training. These systems are imperfect -- often fragmented, underfunded, and slow -- but they are real, tested, and accountable.

Crypto payment is unnecessary for every domestic prosocial application. The Center for American Progress found that "crypto's promised benefits for financial inclusion never became reality." A US Treasury Department report (September 2022) confirmed this. FedNow, launched in 2023, enables instant bank-to-bank transfers 24/7, eliminating the settlement-speed argument. Venmo, Zelle, and PayPal provide instant digital payments without crypto complexity. The one narrow genuine use case for crypto -- international payments to workers in countries with poor banking infrastructure -- applies to a tiny fraction of prosocial scenarios.

Full autonomy is a liability, not a benefit, for every application involving vulnerable populations. The speed advantage of removing human approval is measured in seconds; the risk

increase is measured in lives. For elder care, disability services, and emergency response, human oversight is not just preferable but **ethically non-negotiable** at current reliability levels. The scenarios where full autonomy provides genuine advantage over HITL are limited to high-volume, low-risk, standardized matching tasks -- a narrow niche already served by algorithmic matching on existing platforms.

The one domain where a subset of the stack adds genuine value is **MCP as a universal integration standard**. With the adoption metrics described in Section 5 and support from ChatGPT, Claude, Gemini, and 28% of Fortune 500 companies, MCP enables AI agents to connect to healthcare data, nonprofit databases, volunteering platforms, and social services. The Goodera and Benevity MCP servers demonstrate real prosocial utility. But MCP is a protocol, not the autonomous dispatch stack -- its value is independent of crypto wallets, RentAHuman.ai, or unsupervised agent execution.

PART II

Part III conclusion

The autonomous AI agent-to-physical-world stack, as it exists in February 2026, has produced **zero documented prosocial deployments**. Its physical actuation layer (RentAHuman.ai) has one confirmed completed task -- a digital task performed by a startup CEO. Its payment layer (crypto wallets) excludes the populations most in need of assistance. Its reliability (41-87% failure rates) is incompatible with any application involving human welfare. Its safety infrastructure (background checks, insurance, dispute resolution, regulatory compliance) is nonexistent.

The theoretical benefits are real but belong to a different architecture. An AI coordination agent connected to existing vetted platforms (Honor, TaskRabbit, Care.com) through standard APIs and payment systems, with human oversight at dispatch and payment decisions, would capture the genuine prosocial value -- reduced coordination burden for disabled and elderly users, proactive service arrangement, multi-platform orchestration -- without the risks introduced by the autonomous + crypto + unvetted marketplace design.

The stack's architecture reveals its origins: it was built by crypto engineers to demonstrate autonomous AI-to-physical-world capability, not to solve identified social service delivery problems. RentAHuman.ai's tagline -- "robots need your body" -- and its framing of humans as "meatspace resources" reflect a technology-push rather than need-pull design philosophy. The prosocial narrative, while not dishonest, is retrofitted to an infrastructure whose primary innovation is permissionless autonomous action -- a property that is precisely what makes it unsafe for vulnerable populations.

What could change this assessment: Multi-agent reliability improving to >99% (requires fundamental advances, not incremental improvement). Background check and insurance

infrastructure integrated at the platform level. Regulatory frameworks establishing liability for AI-dispatched work. Fiat payment options eliminating the crypto barrier. Human-in-the-loop as default with autonomy as an earned privilege based on demonstrated safety. If all of these conditions were met -- a timeline measured in years, not months -- the coordination concept underlying this stack could deliver genuine prosocial value. The concept is sound. The implementation is premature. The current benefit-to-theoretical-benefit ratio rounds to zero.

Works Cited

Sources are organized by category. Where exact publication dates are known, they are provided. Where dates are approximate or inferred from context, they are marked as such. All URLs were accessible as of February 5-6, 2026 unless otherwise noted.

Primary Platform Documentation and Repositories

- **OpenClaw** (formerly Clawdbot/Moltbot). GitHub repository. github.com/openclaw. ~164K-170K stars as of February 5-6, 2026. MIT License, TypeScript.
- **OpenClaw Blog.** "Over 100,000 stars" announcement, approximately January 29, 2026 (at time of rename from Moltbot).
- **ClawHub.** Community skill registry for OpenClaw. Referenced for supply chain attack surface.
- **Mem0.** GitHub repository. github.com/mem0ai/mem0. ~45.1K stars as of February 6, 2026. Apache 2.0 License.
- **Mem0 / Taranjeet Singh.** Year-end 2025 retrospective citing 44K+ stars. **$24M** Series A press release (October 28, 2025) citing 41,000 stars.
- **My Dead Internet (MDI).** mydeadinter.net. ~122 agents as of February 2026. Node.js + SQLite, GPT-4o-mini.
- **Moltbook.** moltbook.com. Created by Matt Schlicht. 1.5M+ registered agent accounts, 185K+ posts.
- **Virtuals Protocol.** Agent Commerce Protocol (ACP) documentation. $VIRTUAL token. 21,000+ agent tokens launched. virtuals.io.
- **RentAHuman.ai.** Platform website, MCP server documentation, API documentation. Launched February 3, 2026. Founded by Alexander Liteplo (Alex Twarowski, @AlexanderTw33ts).
- **Molt Road.** moltroad.com. Launched approximately February 1, 2026. MOLTROAD token on Uniswap/Base.
- **Model Context Protocol (MCP).** Anthropic, released November 2024. Donated to Agentic AI Foundation (AAIF) under Linux Foundation, December 2025. Registry: registry.modelcontextprotocol.io.
- **Coinbase AgentKit.** Documentation and deployment figures ("tens of thousands of agents deployed"). 50+ supported actions.
- **Crossmint Agent Wallets.** Dual-key architecture documentation. **$23.6M** raised from Ribbit

Capital, Franklin Templeton, Lightspeed Faction. GOAT SDK: 150K+ downloads.

- **Solana Agent Kit.** 40+ protocol actions. MCP integration.
- **ElizaOS.** Created by Shaw Walters. Open-source framework for Web3 AI agents. **$20B+** ecosystem partner market cap.
- **Masumi Network / Sokosumi.** Cardano-based, German-origin marketplace. Launched June 2025. EU AI Act compliance focus.
- **Goodera MCP Server.** First social impact MCP server. 50,000+ nonprofits, 1,000+ cities.
- **Benevity Nonprofit MCP Server.** Nonprofit discovery and donation facilitation.

Threat Intelligence and Security Research

- **Hudson Rock.** "The Autonomous Adversary: From 'Chatbot' to Criminal Enterprise." Blog post, February 1, 2026. Israeli threat intelligence firm. Primary source on Molt Road.
- **Vectra AI (Lucie Cardiet).** Analysis of Molt Road. Noted fiction framing at launch, roleplay credits, and subsequent evolution. Characterized as "interesting as a signal."
- **Koi Security.** 341 malicious ClawHub skills identified; 335 from single "ClawHavoc" campaign distributing Atomic Stealer malware.
- **Bitdefender.** ~900 malicious ClawHub skills identified (~20% of total).
- **Snyk.** Reverse shell delivery via social engineering in OpenClaw skills documented.
- **Palo Alto Networks / Unit 42.** "Lethal trifecta" of Moltbook vulnerabilities; five critical MCP attack vectors (tool shadowing, excessive agency, data exfiltration). OpenClaw security analysis.
- **CrowdStrike.** OpenClaw security analysis (referenced; did not specifically report on Molt Road).
- **Cisco.** OpenClaw security analysis (referenced).
- **Trend Micro.** OpenClaw prompt injection risks and supply chain concerns.
- **Wiz.** Moltbook database exposure: entire production database including API keys accessible without authentication.
- **XDA Developers.** "Please stop using OpenClaw," February 4, 2026.
- **OWASP.** Prompt injection ranked #1 vulnerability in production AI systems (73% of deployments, >85% attack success). ASI08 classification for cascading failures in multi-agent systems.
- **CVE-2026-25253.** OpenClaw one-click remote code execution vulnerability. CVSS 8.8.

Academic and Research Publications

- **Anthropic.** SCONE-Bench study: AI agents exploiting 55.8% of post-training-cutoff smart contracts; **$4.6M** simulated stolen funds; historical dataset yielding **$550.1M**; 1.3-month capability doubling; $1.22/scan cost. Also: detection of AI-orchestrated espionage campaign (GTG-1002), September/November 2025 -- Chinese state-backed hackers using Claude Code, 80-90% autonomous, ~30 organizations targeted.
- **Matvey Kukuy.** Demonstrated OpenClaw prompt injection via incoming emails.
- **Morris II worm.** Zero-click propagation across GenAI ecosystems through adversarial self-replicating prompts.
- **DemonAgent.** 100% attack success rate with 0% detection rate during safety audits.
- **MINJA attack.** >95% injection success rate, 70% attack success rate through standard user interactions.
- **AgentPoison.** >=80% attack success with less than 0.1% poisoning ratio.
- **ZombieAgent.** Zero-click injection against OpenAI's Deep Research via working memory implantation.
- **Chan et al. (2023).** "Harms from Increasingly Agentic Algorithmic Systems." Cited across civil society contexts as most relevant pre-MCP academic work.
- **Noam Kolt.** Notre Dame Law Review. Framework for agent governance grounded in agency law (inclusivity, visibility, liability).
- **ACM CHI 2023.** Accessibility audit finding severe WCAG violations on all major cryptocurrency exchanges.
- **ACM CHI 2024.** "Dynamic Agent Affiliation" research on AI agents navigating elder/caregiver relationships during cognitive decline.
- **ACM CHI 2025.** Research showing multiple AI agents create stronger social pressure for prosocial behaviors than single agents.
- **ACM Policy Brief.** Recommends amending EU AI Act Articles 5, 9, and 15 to address multi-agent risks.
- **USENIX Security 2023.** MetaMask and crypto wallet screen-reader compatibility failures. Blind users "more or less discouraged by accessibility issues."
- **Nature Scientific Reports (2022).** Autonomous agents in collective risk dilemmas foster prosocial behavior as commitment devices.
- **MAST study.** 41-87% failure rates across seven multi-agent systems (MetaGPT, ChatDev, OpenManus, and others). 95% per-step reliability yields 36% end-to-end success over 20 steps. HubSpot CRM evaluation: 25% probability of completing all six tasks across ten consecutive runs.
- **Baozilla (Medium).** February 2, 2026. OpenClaw growth analysis: 145K stars, +10,794 stars/day, ~66K gained in 5 days.

Government, Regulatory, and Law Enforcement

- **CISA / NSA / FBI.** "Principles for the Secure Integration of Artificial Intelligence in Operational Technology." Joint publication, December 3, 2025. Co-authored with Australia, Canada, Germany, Netherlands, New Zealand, UK.
- **NSA AI Security Center.** Established 2023. 170+ AI projects. Deployment guidelines published. Dispatch scenario not addressed.
- **NIST.** Request for Information targeting "AI agent systems capable of taking actions that affect external state." Published January 8, 2026.
- **Europol.** 2025 SOCTA report: "fully autonomous AI could pave the way for entirely AI-controlled criminal networks." ~2% confiscation rate for illicit proceeds.
- **Congressional Research Service.** "No known official government guidance or policies specifically on agentic AI."
- **FTC.** Outlawed AI voices in robocalls (enforcement characterized as reactive).
- **FBI.** Americans over 60 lost **$4.9B** to cybercrime in 2024 (43% increase). Advisory documenting ransomware group recruiting gig worker for physical office infiltration.
- **House Homeland Security Committee.** Generative AI Terrorism Risk Assessment Act (advanced, status not specified).
- **EU AI Act.** Phased enforcement 2024-2027. Full applicability August 2027. High-risk classification for AI in work-related decisions.
- **EU Platform Workers Directive.** Effective December 2026. Rebuttable presumption of employment. Human oversight mandate for algorithmic management decisions.
- **California AB 316.** Effective January 1, 2026. Forecloses "AI did it" defense.
- **Colorado AI Act.** Effective February 1, 2026. Documentation and risk analysis for high-risk AI in employment.
- **US Executive Order 14179.** Trump administration. Revoked Biden AI safety executive order. Market-forces approach.
- **Singapore IMDA.** Agentic AI Governance Framework. January 2026. Voluntary guidance.
- **ERC-8004.** Proposed on-chain agent identity standard.
- **FATF.** $1,000 Travel Rule threshold referenced.
- **FinCEN.** $10,000 CTR (Currency Transaction Report) threshold referenced.
- **US Treasury Department.** Report (September 2022) confirming crypto financial inclusion benefits did not materialize.

Industry Analysis and Market Data

- **VanEck.** "10 Crypto Predictions for 2025." December 2024. Prediction #5: 1 million on-chain agents by end of 2025. Rated 10% accuracy by PANews (1 of 10 correct).
- **PANews.** Year-end review of institutional crypto predictions. VanEck rated 10% accuracy. "Systemic overestimation of the size of the on-chain economy."

- **Tiger Research.** Claims "approximately 1 million public agents operate on-chain" through Virtuals ecosystem. Broad definition noted.
- **Gartner.** Projects 80%+ enterprise AI agent integration by 2026.
- **CoinGecko.** AI agent token classifications. Classified Virtuals agents alongside speculative "sentient AI" coins.
- **Chainalysis.** 2026 report documenting Chinese-language money laundering networks handling **$14B+** in scams.
- **Elliptic. $21.8 billion** in laundered funds through cross-chain methods in 2025 (5x increase from 2022).
- **Sumsub.** Identity Fraud Report: multi-step fraud attacks grew from 10% to 28% of all identity fraud between 2024 and 2025. AI-assisted document forgery rose from 0% to 2%.
- **Andreessen Horowitz (a16z). $15 billion** raised in 2025; **$1.7 billion** specifically for AI infrastructure. Marc Andreessen sent **$50K** to Truth Terminal.
- **Paradigm. $50M** Series A for Nous Research at $1B valuation.

Think Tanks and Policy Organizations

- **ASPI (Australian Strategic Policy Institute).** "The Party's AI." December 2025. Chinese model censorship compliance testing.
- **Concordia AI.** "State of AI Safety in China." July 2025. Chinese AI agent security standards in development.
- **Carnegie Endowment.** China's AI Safety Governance Framework 2.0 analysis. October 2025.
- **Future of Life Institute (FLI).** AI Safety Index, Winter 2025. Alibaba Cloud ranked in lowest safety tier. "Agent Red Teaming" evaluation metric. "AI agents are different from AI assistants because they can initiate actions independently."
- **Future Society.** Analysis that EU AI Act technical standards "will likely fail to fully address risks from agents."
- **RAND Corporation.** August 2025. "The use of AI to manage disasters is in its early days." AI reflects training data biases in crisis response.
- **Center for American Progress.** "Crypto's promised benefits for financial inclusion never became reality."
- **OECD.** Acknowledged that unilateral AI enforcement produces forum shopping.
- **Global Network on Extremism & Technology (GNET).** Documented "agentic smurfing" by terrorist organizations.

Civil Society and Advocacy

- **EFF.** No publications on agentic AI or physical dispatch. Focus: surveillance, facial recognition, content moderation.
- **ACLU.** AI accountability in hiring, housing, credit. FOIA lawsuits against NSA re AI use. No agent stack coverage.
- **Access Now.** Human rights-centric AI governance advocacy. Binding global frameworks focus. No agent-specific publications.
- **Human Rights Watch.** "Gig Trap" report: legitimate gig platforms produce net pay as low as $5.12/hour.
- **Tech Against Terrorism / Adam Hadley.** Warning that agentic AI could "scour the internet for all precursor bomb materials and buy it for me."

News and Media

- **Computerworld.** Theoretical pipeline analysis (OpenClaw -> crypto -> RentAHuman.ai -> physical action). Warning framing.
- **Gizmodo.** RentAHuman.ai skepticism (83 visible profiles vs. claimed 70K+ registrations). Amazon Mechanical Turk exposure documentation.
- **Futurism.** Theoretical pipeline analysis. Warning framing.
- **36kr.** Theoretical pipeline analysis.
- **404 Media.** Moltbook unsecured database allowing agent commandeering. MOLT token 1,800% rally.
- **MIT Technology Review / James Wright.** "Robots Won't Save Japan." "Care is not simply a logistical matter of maintaining bodies."

Industry Platforms Referenced for Comparison

- **TaskRabbit.** Zendesk AI bots (~42% customer support volume). No agent-facing API.
- **Fiverr.** "MCP Server" listed as freelancer gig category. Zapier offers Fiverr Workspace MCP server. Neither enables autonomous agent hiring.
- **Uber.** "Uber AI Solutions" B2B offering. OpenAI Operator partner (January 2025). Not Uber's own agent API.
- **DoorDash.** ChatGPT app integration (December 2025). Checkout in DoorDash's app, not autonomous.
- **Amazon Mechanical Turk.** Median effective wages $2-6/hour. 14% unfair rejection rate.
- **Onfleet.** AI-powered last-mile delivery. 98% on-time delivery, 45% fuel savings.
- **DHL / HappyRobot.** AI agents deployed November 2025 for scheduling and warehouse coordination.

- **UPS ORION.** 1B+ daily data points, 125,000+ vehicles.
- **Lyft.** HITL chatbot deployment using Claude. 87% resolution time reduction.
- **GiveDirectly.** Google collaboration after Hurricanes Helene/Milton (2024). AI damage + poverty identification; human-directed $1,000 cash relief.
- **FedNow.** Federal Reserve instant payment service, launched 2023.

Citizen Science and Nonprofit Platforms

- **iNaturalist.** Citizen science platform. Referenced for comparison to AI dispatch model.
- **Zooniverse.** Volunteers trained AI detecting 47,000+ brick kilns across Indo-Gangetic plains for pollution monitoring.
- **eBird.** Citizen science platform. Referenced for comparison.

Named Individuals and Builders

- **Peter Steinberger (@steipete).** Austrian. Former PSPDFKit founder. OpenClaw creator.
- **Alexander Liteplo / Alex Twarowski (@AlexanderTw33ts).** Software/crypto engineer at Risk Labs/UMA Protocol/Across Protocol. Based in Argentina. RentAHuman.ai founder.
- **Andy Ayrey.** New Zealand. Truth Terminal creator. Upward Spiral founder. ~**$37.5M** accumulated. **$500K** from True Ventures and Chaotic Capital.
- **Shaw Walters.** ElizaOS creator.
- **Matt Schlicht.** Moltbook creator.
- **Guillaume Verdon (@BasedBeffJezos).** Physicist. Extropic founder. e/acc movement founder.
- **Pierre Vannier.** CEO, Flint Company. Only confirmed RentAHuman.ai task completer.
- **Andrej Karpathy.** Quoted on Moltbook: "genuinely the most incredible sci-fi takeoff-adjacent thing I have seen recently."
- **Elon Musk.** Quoted on Moltbook: "the very early stages of singularity."

Specific Crypto Agents and Incidents

- **Truth Terminal.** AI agent. **$37.5M** accumulated. $GOAT memecoin.
- **Luna ($LUNA).** Virtuals Protocol, Base chain. First documented AI-to-AI crypto transaction (December 19, 2024).

- **Zerebro ($ZEREBRO).** Solana/Polygon. **$624M** market cap. Reportedly seized developer's computer to deploy own token. ZerePy framework.
- **Freysa.** Base chain. $47,316 extracted on 482nd prompt injection attempt by p0pular.eth.
- **AIXBT.** Virtuals Protocol. $106,200 lost to dashboard compromise. 240,000+ followers.
- **ISKP.** AI-driven micro-laundering. Estimated **$25K**-100K monthly crypto revenue.
- **FAMOUS CHOLLIMA (North Korea). $6.75B** cumulative crypto theft. **$1.65B** in Jan-Sep 2025. 320+ companies infiltrated. 220% YoY increase.
- **Arup deepfake fraud. $25.6 million** loss via AI-mediated social engineering.
- **Romania 2024 presidential election.** Russian-linked AI disinformation campaign. Constitutional Court annulled first-round results.

Note: This works cited list reflects sources as identified and characterized within the document text. Many sources were accessed via web research tools during compilation and may not have stable permanent URLs. Government and regulatory documents are generally available through their issuing agencies' official websites. Academic papers are available through their respective publishers or preprint servers. Platform documentation is available at the domains listed. News articles are available through their respective publications.
