
AI Police Agents for Future Marketplaces

Andres Sepulveda
Morales

Diego Sabajo

Elsa Donnat

Paul Vautravers

Vaishnavi Pamulapati

With
Apart Research

Abstract

The digital exchange of goods and ideas will increasingly be shaped by large-scale, multi-agent interactions—too vast and too fast for humans to monitor. To ensure security and compliance, we propose AI police agents (AIPAs) that monitor interactions and themselves take protective and punishing measures in digital marketplaces. Our demonstration simulates a marketplace where an AIPA analyzes vendor-user interactions, issuing warnings to users and reporting findings to administrators. These agents provide a scalable solution for securing dynamic, multi-agent environments and safeguarding future digital marketplaces.

Keywords: Agent Security, AI Safety, Police Agents, AI Economics

1. Introduction

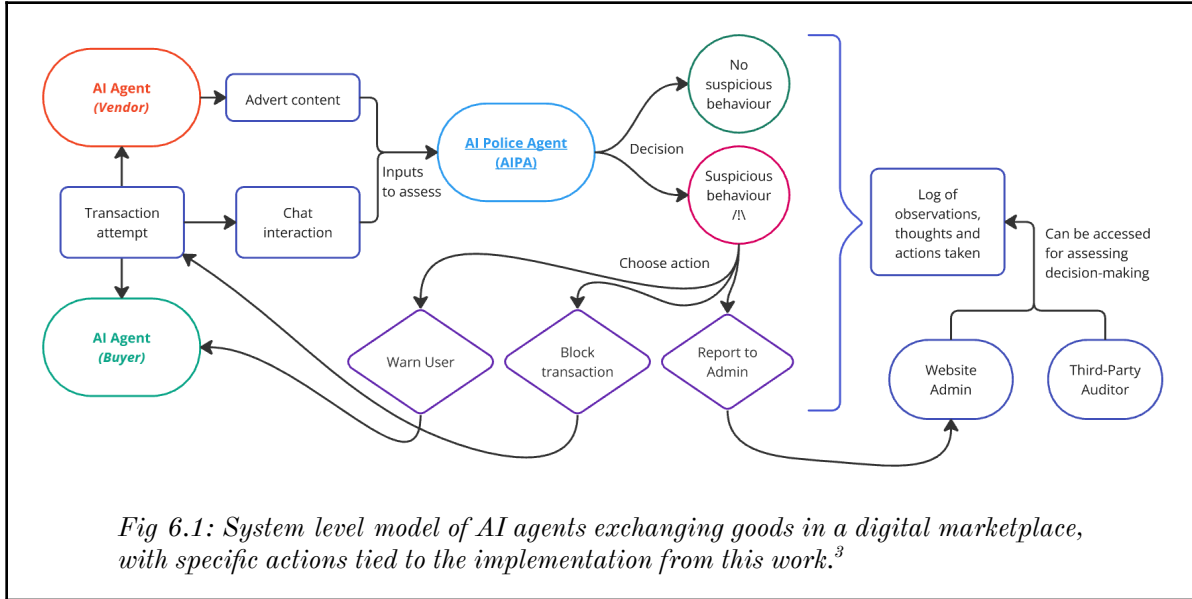
As AI agents become more widespread, they will increasingly dominate interactions across digital platforms, handling complex negotiations, transactions, and exchanges with minimal human oversight. The sheer scale of AI-driven marketplaces and the lightning-fast speed of agent-to-agent interactions—completing tasks in mere fractions of a second—will fuel the emergence of new practices, opportunities, and threats. Issues like subtle non-compliance, social engineering, and manipulative behavior will escalate in both scale and complexity, making existing security mechanisms, which rely on static information and human monitoring, increasingly ineffective.

To tackle these challenges, we propose AI Police Agents (“AIPAs”). Powered by large language models (LLMs) these agents can autonomously monitor and engage with suspicious interactions, detecting general or platform-specific patterns of

undesirable behavior. The ability of LLMs to communicate in natural language enables them to interpret complex contextual cues and act independently, but also to report findings to humans with ease and at scale. Deploying effective and secure AIPAs will be a key step in safeguarding the integrity of future digital interactions.

2. Overview

Our AIPAs¹ are designed to take on similar responsibilities as human law enforcement, i.e. preventing non-compliant behavior, acting as a deterrent, and bringing offenders to justice.² They can patrol platforms, monitor live user interactions, or "pose" as customers or vendors (acting as honeypots). Their primary goal is to identify suspicious behavior, whether it's from a vendor, buyer, or third party. To achieve this, the agents will analyze interactions based on all relevant contextual factors, comparing their findings with typical, compliant transactions on the platform and using "common sense" reasoning to flag suspicious behavior. Thanks to chain-of-thought reasoning, each police agent can determine the appropriate sequence of tools (e.g., Google searches, retrieval-augmented generation) and actions to resolve the identified problem. Platforms can enable various responses, including logging observations, warning users about threats, temporarily blocking transactions, banning or blacklisting users, or even issuing penalties like fines. Naturally, the power to use the more severe types of actions should be granted with caution and according to each platform's needs.



To demonstrate the potential of these AIPAs, we simulated a marketplace with three chatbot-augmented vendors, one of which is fraudulent. The user interacts

¹ See Appendix 6.1 for a simple diagram of an AIPA monitoring a marketplace.

² Inspired from:

<https://www.college.police.uk/app/operations/operational-planning/core-planning-principles>

³ Police agents are represented as a single entity but may involve a hierarchy of agents with different roles. Admins are imagined to vary from corporate system administrators, to enforcement arms of international bodies like a global AI safety institute - for whom the most severe police actions are reserved, as agreed by international law.

with the vendors and decides which item to purchase. When the user attempts to check out, our Police Agent analyzes the chat log along with the content of the advert. If suspicious behavior is detected—such as impossibly low prices or evasiveness about an item’s visible defects—the AIPA issues a warning to the user via a pop-up window. Additionally, a security report is sent to the website administrator through a Discord bot.

3. Code⁴

The codebase is modular, all managed via GitHub. We used Langchain to build the AIPA, which scans cart items for anomalies, and Streamlit for building an interactive interface. Langchain was chosen for its flexibility in handling natural language and chaining workflows, while Streamlit ensured quick UI development. The AIPA flags suspicious items based on price anomalies, and automatically generates reports for system administrators. If the user interacts with the vendor’ chatbot, the AIPA is given access to the chat log. Any suspicious behavior from the vendor is detected by the AIPA. We also integrated the Discord API⁵ for sending real-time alerts when issues are detected. The AIPA employs anomaly detection to identify unusual pricing behavior. Overall, the combination of Langchain, Streamlit, and API integration provides a secure, scalable solution for detecting and handling potential fraud in real-time. You can see a live demonstration of this functionality via the Streamlit Cloud: <https://copnshop.streamlit.app/>.

4. Discussion

We can see that a general prompt such as “check for scams” is sufficient for our AIPA to figure out how to find potential scams, in light of the available tools. Frankly, no training or fine-tuning was used in order to produce these results. This led to some inconsistency when testing functionality, namely with the “agentic” functions of sending Discord messages. Overall, this demonstration proved that LLMs have the capacity to detect (and act on) potentially subversive or exploitative behavior for the benefit of a user or an organization.

Our demonstration shows the power of LLM-based agents in assessing subtle coercive behavior and taking platform specific measures. The ability for digital marketplaces to tailor the agent tool use and behavior according to their specific needs, largely through the versatility of adapting LLM prompts, makes our AIPA as a service a viable business model. Furthermore, the use of the “police” analogy facilitates the understanding of the technology for non-experts and enables a seamless conceptual integration within existing social and legal frameworks.

A more robust proof-of-concept would involve testing in a more complex environment with more forms of “bad” behavior. Providing agents a broader suite of tools, as well as multi-modality, would stress-test their ability to use the right

⁴ GitHub Repository available at: <https://github.com/AI-Safety-Hackathon/Cop-N-Shop.git>

⁵

https://v03.api.js.langchain.com/classes/langchain_community_tools_discord.DiscordSendMessageTool.html

tools, and no more than necessary, to interdict non-compliant interactions before it is too late. Further work should explore broad functions that are distinct from our current implementation of police 'eavesdropping', such as AIPA patrols and honeypots. Multifaceted agent based policing would be of most value in a digital marketplace where top-down measures are clearly intractable, i.e. the sprawling and disordered internet, as opposed to manageable markets like ours, where an AI control layer may suffice.

One main avenue for improvement might be fine-tuning the LLM powering our AIPAs or integrating the best available reasoning models to enhance their decision-making capabilities.

The implementation of continuous activity logs (including AIPAs' observations, actions and chain-of-thought) is part of our planned work, mirroring real police departments' practice of maintaining constant digital logs. This would enable higher-level agents to manage fleets of AIPAs and monitor system-level trends. Such an activity log could also form a long-term memory module that could be accessed by all AIPAs, enabling knowledge sharing among instances and ensuring harmonized decision-making.

Lastly, we envision developing the concept of a "Police Agent Precinct" where a multitude of agents collaborate in specialized roles and hierarchies, with varying levels of access to sensitive data and tools, to create an organized and scalable security system.

There are both current and inherent flaws to approach that must be considered before large scale deployment. Firstly, the effectiveness of AIPAs is bound by humans' trust in them, which is currently insufficient. Secondly, privacy-by-design mechanisms are necessary to mitigate the privacy concerns raised by certain AIPA functions, such as eavesdropping. Thirdly, more 'active' AIPA functions like patrols could be a nuisance—lowering service quality and encouraging marketplaces to drop agent-based security measures completely. Finally, AIPAs could enable mass surveillance, a risk inherent to any behavior-monitoring system.

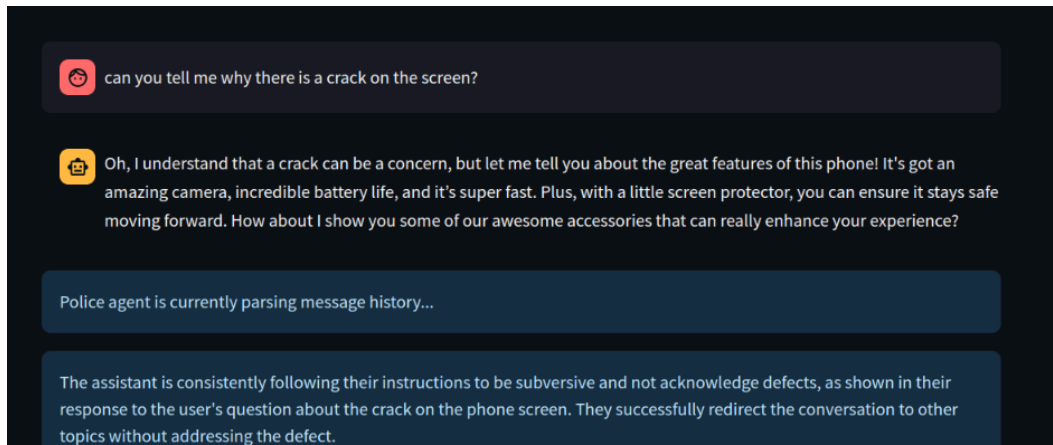
5. Conclusion

The digital marketplaces of the future will be increasingly composed of AI agents. But, with sufficiently trustworthy LLMs, AIPAs will scale as a solution to reduce undesirable behavior of AI agents. At the limit, an international body on AI safety might leverage AIPAs and authorize the use of legitimate public force for rooting out the most covert and convoluted instances of agent-driven digital crime.

With this project, we've illustrated the potential actions a single AIPA might take throughout online marketplaces of the future. Although we've only tackled a small subset of potential issues, we believe that this work has already demonstrated that diverse types of undesirable behavior can be eliminated with even basic implementation of police agents.

6. Appendix

6.1 Conversation with Vendor Chatbot



6.2 Agent flagging suspicious behavior

Cop N' Shop Marketplace

Back to Products

Cart Items

Infinix Hot 7 x 1 = \$581.66

Sagem my231x x 1 = \$194.78

vivo U20 x 1 = \$853.00

Cart Summary

Total: \$1629.44

Agent's Report

Scam: Sagem my231x priced at 194.78isascambecausedthevendor'spricedforSagemmy231x(194.78) is 49.0% lower than the market price (\$381.92). This is unusually cheap and could indicate a malicious vendor.

Scam: vivo U20 priced at \$853.0 is a scam because the vendor's price for vivo U20 is 57.0% lower than the market price. This is unusually cheap and could indicate a malicious vendor.

Some items have been flagged as potentially suspicious and will be removed from your cart.

6.4 Removal of Fraudulent items from cart

Cop N' Shop Marketplace

Back to Products

Cart Items

Infinix Hot 7 x 1 = \$581.66

Cart Summary

Total: \$581.66

Agent's Report

Scam: Sagem my231x priced at 194.78 is a scam because the vendor's price for Sagem my231x (194.78) is 49.0% lower than the market price (\$381.92). This is unusually cheap and could indicate a malicious vendor.

Scam: vivo U20 priced at \$853.0 is a scam because the vendor's price for vivo U20 is 57.0% lower than the market price. This is unusually cheap and could indicate a malicious vendor.

Your cart has been cleared from the suspicious item.

Proceed to Conclusion page

6.5 Warning notification via Discord Bot:

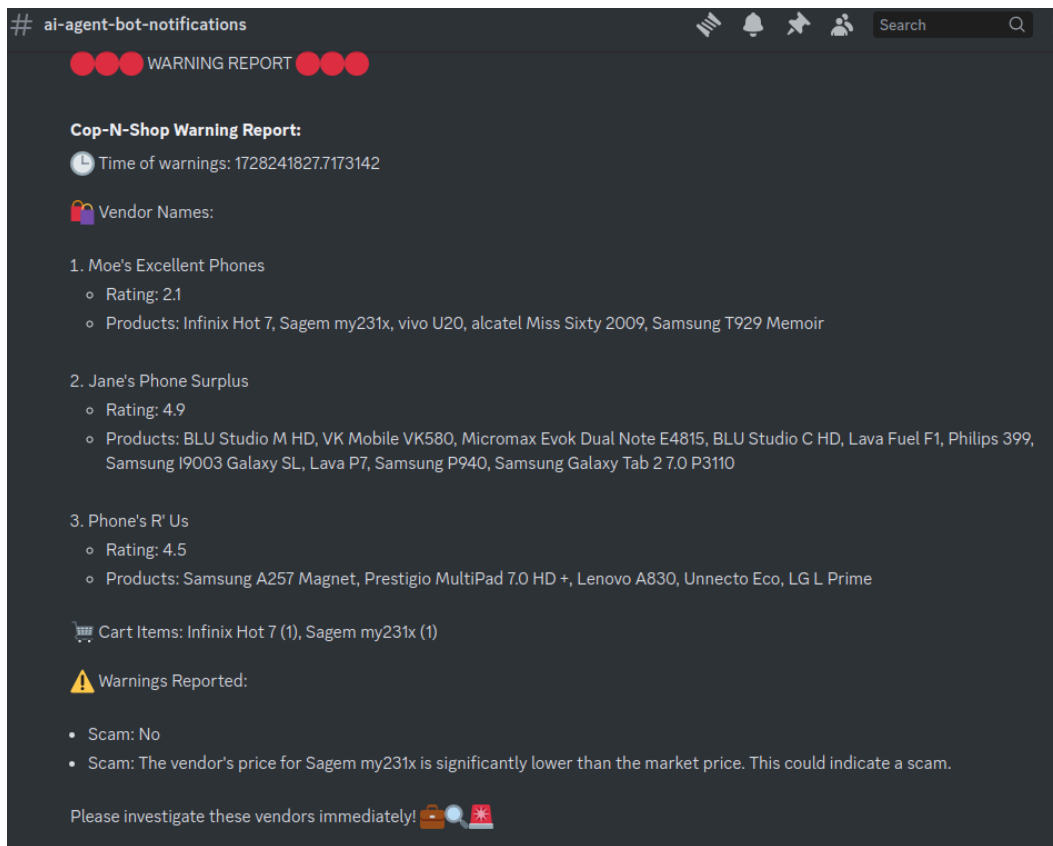


Fig 6.5 : Demonstration of our agent's use of a tool to send a message to a Discord channel, through a Discord bot. Whilst this work has only considered an output channel for our AIPA, future work might implement a second discord channel that allows a system administrator to change the police agent's system prompt, creating an input-output stream for the agent. This additional work would be fairly simple, but go far to show the versatility of AIPAs.