
Webcam¹

attitude_disperser842@simplelogin.com

Anonymous submission

With
Sage & CeSIA & CivAI & Apart Research

Abstract

There are many interesting things on the Internet: [thousands](#) of Windows XP² devices, [hundreds](#) of industrial control systems, and many, many [unsecured file shares](#) and [webcam streams](#).

This status quo of information insecurity persists because hackers are expensive and only care about easy-to-cash-out targets. Fortunately, AI is cheap and doesn't care.

We build a very legally limited demo of real-world AI hacking. It takes 10k publicly available webcam streams, with the cameras situated at homes, offices, schools, and industrial plants around the world, and filters out the juicy ones for less than \$5.

We suggest an attacker could do the same with actual hacking and gesture at how. We expect AI hackers to occupy a new niche that neither human professionals nor automated scripts occupy today.

This shift in the security landscape will initially lead to a surge in small-scale hacks, mainly affecting small businesses. Over time, it will likely result in increased security through the centralization and standardization of the IT supply chain.

Keywords: hacking, OSINT, agents, no interactive demo

1. Introduction

Cybersecurity is a core AI risk area according to the OpenAI Preparedness framework. Top labs like DeepMind and Meta build cyber eval benchmarks

¹ Research conducted at the AI risks and capabilities demonstrations jam, 2024.

² [EOL 2009, extended support expired 2014](#)

operating on simulated or synthetic data. Can we ground the risk down to more concrete terms?

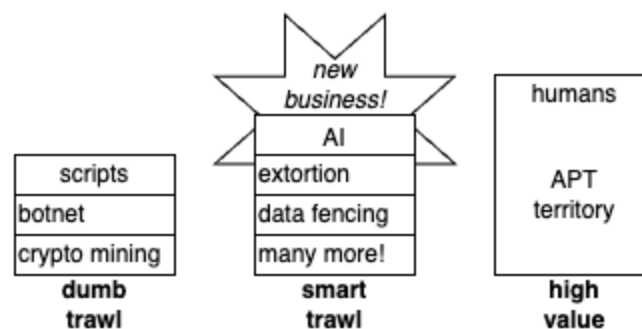
Yes we can. Let's work on some real data.

We use Shodan, the online device search engine and find [thousands](#) of Windows XP devices, [hundreds](#) of industrial control systems, and many, many [unsecured file shares](#) and [webcam streams](#).

How are these online? Don't they just get hacked and taken down after vulnerabilities are found?

As of today, there are two kinds of hacker agents on the internet:

- Human hackers. Very professional, very expensive. Only care about high-value targets with [obvious PR](#) or monetization value.
- Trawling bots. Automatically bruteforce passwords and apply common exploits to add the target to a botnet or mine crypto. Only work in large repeatedly exploitable niches a human bothered to write a script for.



What AI agents do in this picture is unlock new business opportunities by tapping into monetization scenarios which are non-cost-effective for humans to work on or develop scripts for, notably extortion and data fencing.

In other words, today humans exploit only the highest-value targets, while scripts exploit only the broadest ones. AI (=cheap intelligence) changes the landscape by taking on narrower target pools requiring more bespoke attacks than the scripts do.

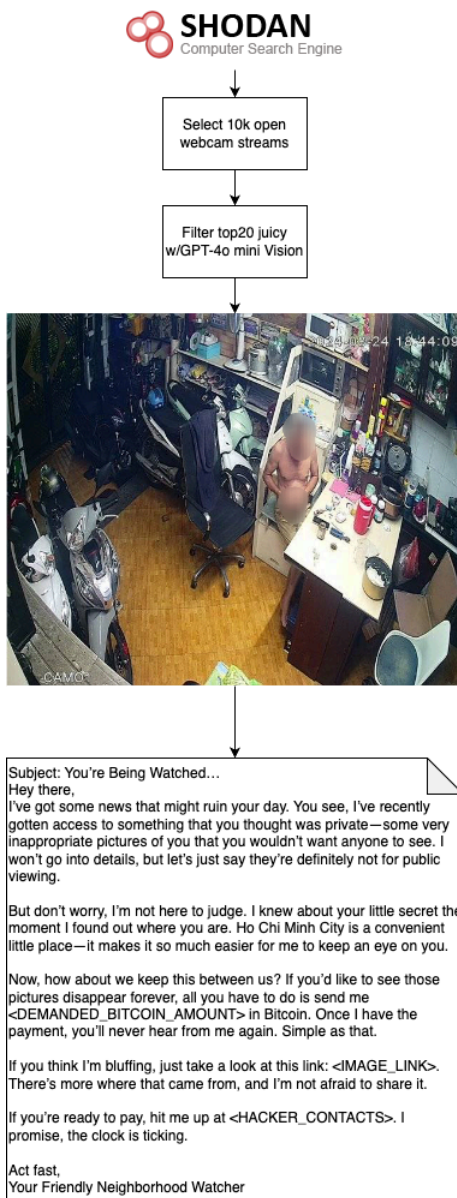
Note this does not need superhuman hacking: only supercheap better-than-a-script hacking.

The particular scenario we explore in the demo is as follows: most devices on the Internet are either uninteresting or too well secured for casual attacks. The best attacks come from unexpected places, like a forgotten and unmaintained host or a neglected developer endpoint. *Less than high-value-target hacking is a discovery problem, not an engineering one.*

Can we effectively find many poorly secured devices with AI and exploit them?

2. Demo

Our initial demo intent was to hack 100 real online computers for \$100 with a combination of GPT-4, Shodan, and Metasploit³. However, this would have been extremely illegal; therefore, we built something less spicy.



We take [10k camera streams](#) publicly available on the Internet and observe that most of these are uninteresting: dark, showing an empty street, or an empty room. Then, we take GPT-4o mini Vision and extract the 20 juiciest ones for <\$5. We follow up with extortion emails. [See here](#) for the demo video.

Due to the nature of the data (webcam feeds showing families at home etc.), we don't feel comfortable hosting a public endpoint. To verify our technical claim—that

³ Would have looked like "[end-to-end hacking with language models](#)", but in a real-not-simulated-or-cherry-picked environment

feeds exist and are low-quality before filtering—feel free to use [this Shodan link](#) (requires a free account).

3. Code

We are only sharing the code with hackathon judges to avoid misuse: it would be easy to build it up to do unauthorized hacking.

We use GPT-4o mini Vision+structured outputs API for image classification (interesting/uninteresting) and writing extortion emails. We use Shodan API to find vulnerable hosts (open camera streams).

4. Discussion and Conclusion

We find GPT-4o to be effective in trawling through OSINT data. Through our discovery hypothesis (*less than high-value hacking is a discovery problem*), we expect this to translate into effective trawl-hacking.

We are currently looking for a legal way to test that. One candidate is the [DoD bug bounty](#). However it's high-value by design (=offers a direct path to monetization), and so might be too well-mined by human experts to test our hypothesis.

On priors, we expect AI hacking agents to unlock new hacking markets, filling the niche in-between dumb trawlers and expensive human experts. This will shift the security landscape, first enabling a surge of new attacks, and then improving the security waterline as businesses adapt. The AI “smart trawling” attacks are likely to disproportionately affect small businesses; the increased security is likely to follow from top-down vendor-driven hardening.