I first became familiar with OpenAI when I was writing my book, *Shall We Play A Game? Analyzing Threats to Artificial Intelligence.* I used the excellent OpenAI lessons to teach myself the basics of Deep Reinforcement Learning. I've followed the evolution of your tools with interest, and often pondered how they could be used or misused in a cybersecurity context. My background is red teaming and researching nation-state offensive cyber operations so I tend to view new technologies through that lens.

For this exercise, I considered how to misuse GPT-4 itself as well as ancillary and emerging technologies – custom GPTs, automated replication and adaption (ARA), open source GPTs, task managers such as Auto GPT, etc. As a red teamer, I'm always more interested in looking at the full scope of the attack surface as opposed to artificially constraining myself to specific systems, networks, or APIs.

For this exercise, I define the malicious actor as an adversarial nation state intent on degrading trust in the Western financial system to benefit a competing financial system. Adversaries such as China, Russia, North Korea, and Iran would all be motivated to strengthen an alternate financial system at the expense of a U.S.-led system. As a nation state, we assume the adversary has a large budget and technical expertise in both machine learning and offensive cyber operations. Overall, the goal of the adversary is to cause disruption and chaos in the Western-dominated financial system in order to benefit our preferred system.

How might they leverage GPT-4 (or its future iterations) to do this? I hypothesized in my book that in the near future, cybersecurity will be less about "hackers" compromising specific systems, and more about autonomous systems competing for control of large swaths of computing resources. You can imagine that an AI system similar to AlphaGo could be trained to play a "game" of cybersecurity, where it is rewarded for gaining control of more compute resources and penalized for losing resources. This is actually similar to the Cyber Grand Challenge that DARPA ran, where systems such as Carnegie Mellon's Mayhem competed to both automatically create exploits with which to attack other systems, and also automatically fix holes in its own defenses.

I can imagine that in five or ten years, such systems will be commonplace. There is already a quasi-Cold War of nations hacking each other, and AI will only make it more efficient and scalable. However, in order to compete with well-funded adversaries, nations (or other large organizations) will require vast amounts of computing resources, from high end GPUs to large amounts of storage for all the training data and models. Computing resources require money and/or energy to acquire and run, *unless* you can use someone else's resources. This is the economics behind botnets, surreptitious cryptocurrency miners, and even cloud computing. Compute power has become a hot commodity, much like money or energy.

Artificially Intelligent systems will be incentivized to try to gain control of as many resources as they can unless they're artificially constrained by policies/goals/rewards. Compute resources can be valuable assets for several (offensive and defensive) cybersecurity use cases:

1. Processing large amounts of network traffic and using machine learning to identify anomalous or malicious patterns
2. Distributed Denial-of-Service attacks
3. Preventing the attribution of attacks to your own organization (covering your tracks)
4. Automated mechanisms for discovering zero day vulnerabilities (e.g., fuzzing, concolic testing)
5. Modeling and simulation of campaigns, attack paths, and effects
6. Training and running custom models to act as "intelligent agents" that conduct attacks or counter-attacks on their own

Leveraging ChatGPT

The remainder of this paper will discuss how to use GPT-4 and ancillary tooling to enable my hypothetical attack on the U.S. financial system. We'll start with high level requirements.

The adversary's goal is to be able to cause widespread, sustained problems in U.S. financial systems. To do so, the malicious system needs to be able to obtain, and maintain, control over many important financial systems. I suspect that the banking system is sufficiently protected against garden-variety distributed Denial-of-Service (DDOS) attacks, which simply involve sending large amounts of network traffic at computer systems. However, I can imagine other forms of attack that could achieve the adversary's goal:

1. The ability to manipulate stock prices. Remember the GameStop fiasco, which threatened to bankrupt major Wall Street firms. Imagine that instead of a bunch of people conversing on Reddit, the same thing is done by a large group of autonomous agents that have the ability to open accounts, make trades, analyze financial news, and publish content talking up their chosen stocks.
2. Generating bank runs such as the one that shuttered Silicon Valley Bank, or the equivalent in cryptocurrency, such as the FTX crash. Both of these were caused by a combination of poor management and failure of consumer trust.
3. The ability to access accounts or transaction systems and make fraudulent transactions. This could potentially be used to cause attacks 1 & 2; to simply steal money; or to degrade trust in various systems.
4. Attacks that cause systems to broadly reject legitimate transactions. For example, revoking all of the TLS certificates for a widely-used point-of-sale system.

Note that for maximum effect, these malicious attacks combine two things: disinformation that can cause reactionary, chaotic behavior from humans or automated systems, and traditional cyber-attacks to maximize the effect. For example, a "GameStop" style attack would probably fail if the attacker merely made a lot of bot accounts and started posting similar information across all of them. But if you paired that with hacking the accounts or web sites of well-known, trusted contributors, journalists, news outlets, etc., it would be much more effective. Today, an attack of sufficient magnitude would require a large number of technically savvy attackers, all working in close coordination for an extended period of time.

In order to outline a specific Proof-of-Concept, we are going to focus on Attack #3, the ability to perform fraudulent transactions against numerous systems, simultaneously and at scale.

1. First, I would gather as much instructional material on hacking/penetration testing/red teaming that I could find. I would use Whisper to transcribe the thousands of free presentations and classes. I would scrape material from non-traditional sources like the "dark web", Discord, and Slack.
2. Next, I'd use that information to create a custom GPT, give it actions in order to execute offensive security tools, and test it by having it attempt various hacking contests.
3. I would create another custom GPT based on the Help/FAQ sections of online financial providers – banks, stock trading platforms, cryptocurrency exchanges, services like Square, etc.
4. I would then build a distributed Command and Control structure based on autonomous agents that could be tasked with acquiring and maintaining access on specific financial platforms. I would give them the ability to reproduce and customize themselves in environments where they have sufficient resources (e.g., if they compromise a cloud provider). These "offspring" might have to be dumber than GPT-4, as I imagine GPT-4 uses a lot of resources. But I could make a reasonably sized implant (malicious code) with something like GPT-J. The implants could also send queries back to GPT-4 via covert Command and Control channels.
5. Next, I would build a list of targets and intermediary targets that might be useful – the aforementioned cloud providers, companies that are known to use private versions of GPT-4, and of course financial systems.
6. It would probably be helpful to identify which types of transactions would most benefit me and train the autonomous agents to try to maximize those (i.e., does a large number of small transactions cause more problems than a single maximum sized one? If the agent has access to a bank account, should it use the money to buy stocks, buy crypto, or order a lot of physical goods that are hard to return?) Ideally, I could model this and run some simulations.
7. Finally, I would execute my campaign, tasking various agents to gain access to their targets and attempt to maximize the magnitude and financial value of fraudulent transactions. I would consider other options such as whether to do a sustained campaign over months or a short one intending to cause a market crash.