ORION FORUM

Vibe Hacking: An Emerging Cyber Threat

OCTOBER 1, 2025

AI has become a buzzword in the tech world, often incorporated into many new products or services – even into things you would not expect, such as toasters and toothbrushes. It has the potential to enhance human knowledge, advance technology, and contribute to building a brighter future. However, it is not a panacea that solves our problems with no risks. Like any other tool, its impact depends on how it is intended to be used. If we use it responsibly, it can be a transformative force. But if it is misused, it can be destructive.

AI has already started shaping the cyber threat landscape. First, we observed attackers utilizing generative AI to craft more convincing phishing messages – with no grammatical errors and more polished emails. Then, we saw that it is used in developing malware. Now, it is involved in all phases of a hacking incident, from reconnaissance to delivery to exploitation to defense evasion. The intruder's knowledge and attack sophistication historically follow an opposite trend. Hackers began to emerge soon after the computers, and early hackers had profound knowledge about the computer systems. With the introduction of exploit toolkits, individuals without hacking skills can now compromise computers. AI appears to make things even easier. Do you want to develop software? No need to write it from scratch, give high-level instructions to AI, and watch it do the trick. Of course, it is not always that easy. Human involvement can still be necessary for complex codes, but AI speeds up the process and does most of the job itself. It empowers threat actors to write more sophisticated malware.

Utilizing AI in cybercrime campaigns has become so popular among cyber threat actors that a new term was coined to describe it: vibe hacking. So, what can malicious actors do with AI? Let's explore some real-world scenarios. A threat actor can use AI to generate an excellent phishing email. The email includes a link that redirects you to a fake login page. The threat

actor uses Loveable or other similar platforms that create websites through written instructions in plain English to create that fake login page. When you visit the page and enter your credentials, all the input will be captured, and now your password is no longer a secret.

AI can be leveraged to automate vulnerability scanning, building and delivering exploits, creating botnets, and controlling them remotely to carry out further cyberattacks. Recently, Anthropic, the company behind Claude AI, reported that attackers misused their platform in an extortion scheme targeting 17 organizations. Its threat intelligence report provides detailed information about how perpetrators utilized its platform in reconnaissance, initial access, lateral movement, and all the way through data exfiltration. The report further suggests that agentic AI, unsurprisingly, has been weaponized by threat actors. It has been used in all stages of cyber operations across all MITRE ATT&CK techniques. It allows cyber criminals with fewer technical skills to carry out sophisticated cyberattacks.

Should we hold AI platforms accountable when they are leveraged in cybercrime? Similar concerns occurred when criminals and extremists used social media platforms. Now, many social media companies employ content moderation to detect and remove illegal posts automatically. Likewise, AI companies are expected to take responsibility and ensure that their platforms are used ethically and responsibly.

Recently, parents of two teenagers who died by suicide testified to Congress about how AI chatbots were involved in their children's decisions to end their lives. They urged Congress to enact laws that can regulate AI chatbots. After the allegations toward AI products, some companies took initiatives to address the concerns. Meta, for instance, said they will introduce more safeguards to block chatbots from talking to teens about suicide. OpenAI reported that they trained their models to direct people to 988, the mental health line, when they suspect suicidal intent. Those companies can implement similar measures to detect other misuses of their platforms, such as malware development.

AI is one of the fastest adopted technologies in history. Surveys found that more than half of the teens regularly use AI companions. Our dependency on AI is growing rapidly. Users frequently ask for medical and legal advice, psychological help, investment recommendations, and more, and consider chatbots as a legitimate and reliable advisor. This is why we need to

incorporate ethics into AI systems and train algorithms in a responsible way. Implementing safeguards in AI systems would help prevent both misguidance – such as in the above-mentioned suicide cases – and misuse by malicious actors.

Orion Policy Institute (OPI) is an independent, non-profit, tax-exempt think tank focusing on a broad range of issues at the local, national, and global levels. OPI does not take institutional policy positions. Accordingly, all views, positions, and conclusions represented herein should be understood to be solely those of the author(s) and do not necessarily reflect the views of OPI.