SentinelOne blog

GET A DEMO

BlackMamba ChatGPT Polymorphic Malware | A Case of Scareware or a Wake-up Call for Cyber Security?

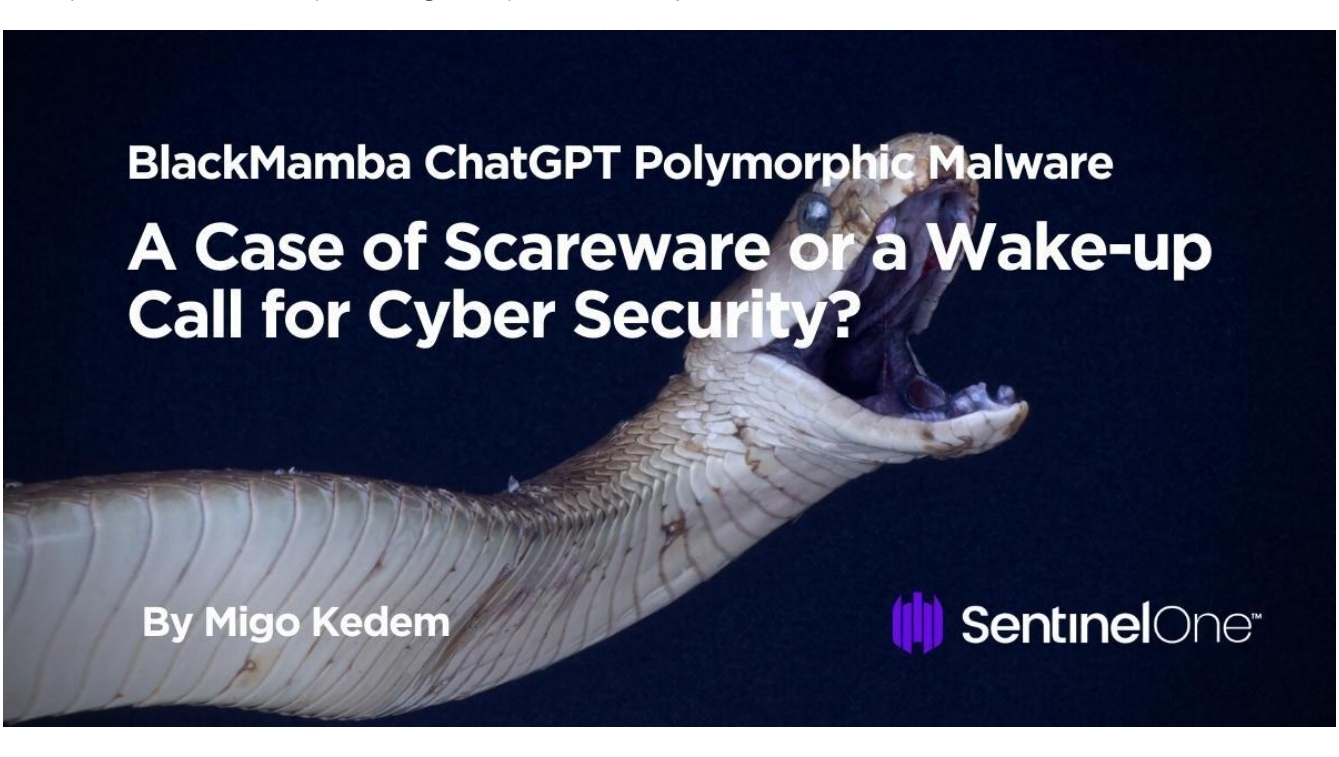
March 16, 2023
by Migo Kedem

Artificial Intelligence has been at the heart of SentinelOne's approach to cybersecurity since its inception, but as we know, security is always an arms race between attackers and defenders. Since the emergence of ChatGPT late last year, there have been numerous attempts to see if attackers could harness this or other large language models (LLMs).

The latest of these attempts, dubbed BlackMamba by its creators, uses generative AI to generate polymorphic malware. The claims associated with this kind of AI-powered tool have raised questions about how well current security solutions are equipped to deal with it. Do proof of concepts like BlackMamba open up an entire new threat category that leaves organizations defenseless without radically new tools and approaches to cybersecurity? Or is "the AI threat" over-hyped and just another development in attacker TTPs like any other, that we can and will adapt to within our current understanding and frameworks?

Fears around the capabilities of AI-generated software have also led to wider concerns over whether AI technology itself poses a threat and, if so, how society at large should respond.

In this post, we tackle both the specific and general questions raised by PoCs like BlackMamba and LLMs such as ChatGPT and similar.



What is BlackMamba?

According to its creators, BlackMamba is a proof-of-concept (PoC) malware that utilizes a benign executable to reach out to a high-reputation AI (OpenAI) at runtime and return synthesized and polymorphic malicious code intended to steal an infected user's keystrokes.

The use of the AI is intended to overcome two challenges the authors perceived were fundamental to evading detection. First, by retrieving payloads from a "benign" remote source rather than an anomalous C2, they hope that BlackMamba traffic would not be seen as malicious. Second, by utilizing a generative AI that could deliver unique malware payloads each time, they hoped that security solutions would be fooled into not recognizing the returned code as malicious.

BlackMamba executes the dynamically generated code it receives from the AI within the context of the benign program using Python's `exec()` function. The malicious polymorphic portion remains in memory, and this has led BlackMamba's creators to claim that existing EDR solutions may be unable to detect it.

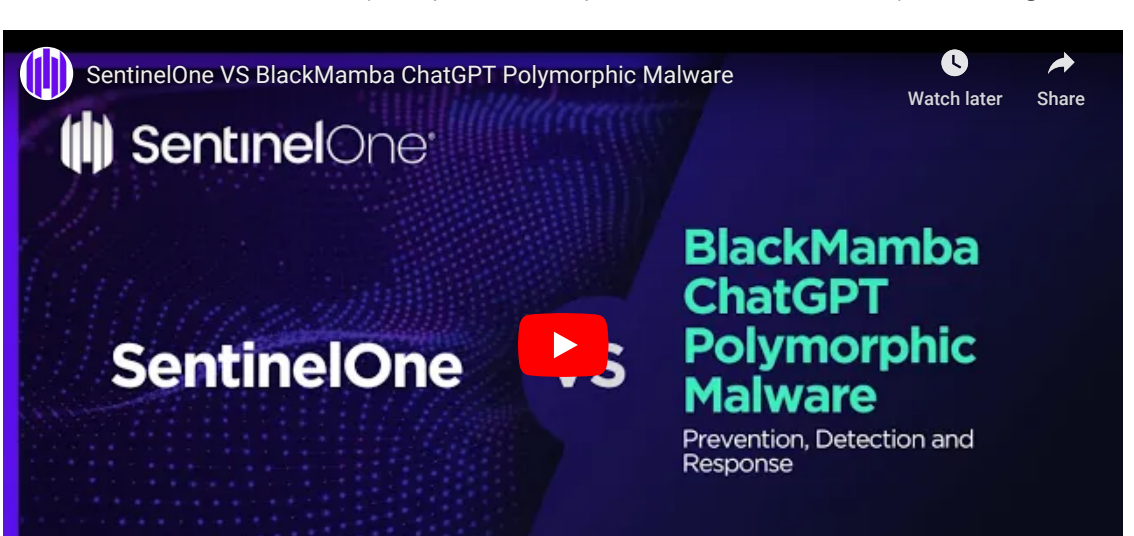
Detecting AI-Generated Malware Like BlackMamba

Such challenges, however, have been well understood in the cybersecurity community. We have seen "benign" channels such as Pastebin, Dropbox, Microsoft Azure, AWS and other cloud infrastructure abused in the past for the same reason of trying to hide malicious traffic in the noise of legitimate network services.

Polymorphic malware is also hardly new; among other things, it is one of a number of factors that helped the industry move beyond legacy AV solutions and towards next-gen AI-driven solutions like SentinelOne.

With regard to isolating malicious code to memory, this is also not a new or novel approach to building malware. The idea of not writing code or data to disk (and therefore evading security measures that monitor for those events) has long been attractive to threat actors. However, modern security vendors are well aware of this tactic. SentinelOne, and a number of other EDR/XDR vendors, have the required visibility into these behaviors on protected systems. Simply constraining malicious code to virtual memory (polymorphic or not) will not evade a good endpoint security solution.

This raises the question: can AI-generated malware defeat AI-powered security software? Indeed, as said at the outset, it's an arms race, and some vendors will have to catch up if they haven't already. At SentinelOne, we decided to put ChatGPT-generated malware to the test.



Does AI Pose a New Class of Threat?

Widening the discussion beyond BlackMamba, which will undoubtedly be superseded in next week's or next month's news cycle by some other AI-generated PoC given that ChatGPT4 and other updated models have become available, just how worried should organizations be about the threat of AI-generated malware and attacks?

The popular media and some security vendors portray AI as a Frankenstein monster that will soon turn against its creators. However, AI is neither inherently evil nor good, like any other technology. It's the people who use it that can make it dangerous. Proof of concepts like BlackMamba do not expose us to new risks from AI, but reveal that attackers will exploit whatever tools, techniques or procedures are available to them for malicious purposes – a situation that anyone in security is already familiar with. We should not attack the technology but seek, as always, to deter and prevent those who would use it for malicious purposes: the attackers.

Understanding What AI Can and Cannot Do

Fundamental to many of the concerns that swirl around discussions of AI is often a need for clarification of what AI is and how it works. The effectiveness of any AI system or LLM like ChatGPT depends on the quality and diversity of its dataset. The dataset used to train the model determines its capabilities and limitations.

Defenders can level the playing field by creating their own datasets, which can be used to train models to detect and respond to threats, something SentinelOne has been specializing in for years.

Despite that, AI is not a magical technology that can do everything. There are limitations to what AI can do, especially in cybersecurity. AI-based systems can be fooled by sophisticated attacks, such as adversarial attacks, which bypass the defenses. Additionally, AI cannot make judgment calls and can reveal bias if the dataset is not diverse.

We need to be aware of the limitations of AI and use it as part of a comprehensive security strategy. That's why SentinelOne deploys a multi-layered approach combining AI with other security technologies and human intelligence.

What About Human Intelligence?

In today's AI-driven world, we can easily get caught up in the latest technological advancements and overlook the importance of human intelligence. Even with AI's ability to analyze vast amounts of data and identify patterns, the human touch remains essential, if not more critical. We need people's ability to reason, to think creatively and critically to supplement AI's capabilities.

Both attackers and defenders employ AI to automate their operations, but it's only through human intelligence that we can strategize and deploy effective security measures, deciding how and when to use AI to stay ahead of the game.

Recent events, like the National Cybersecurity Strategy, have shown that defending our businesses and society against threats isn't just about using a single tool or hiring top-notch talent. The internet, which much like AI has sparked plenty of discussion about its merits and drawbacks, has made cybersecurity a collective challenge that demands collaboration between various stakeholders, including vendors, customers, researchers, and law enforcement agencies.

By sharing information and working together, we can build a more robust defense system capable of withstanding AI-powered attacks. To succeed, we must move away from a competitive mindset and embrace the cooperative spirit, combining our expertise in malware, understanding the attacker's mindset, and using AI to create products that can handle the ever-changing threat landscape. In the end, human intelligence is the icing on the cake that makes our AI-driven defenses truly effective.

Conclusion

Cybersecurity is a cat-and-mouse game between attackers and defenders. The attackers try new ways to bypass the defenses, while the defenders always try to stay one step ahead. The use of AI in malware is just another twist in this game. While there is no room for complacency, security vendors have played this game for decades, and some have become very good at it. At SentinelOne, we understand the immense potential of AI and have been using it to protect our customers for over ten years.

We believe that generative AI and LLMs, including ChatGPT, are just a tool that people can use for good or ill. Rather than fearing technology, we should focus on improving our defenses and cultivating the skills of the defenders.

To learn more about how SentinelOne can help protect your organization across endpoint, cloud and identity surfaces, contact us or request a demo.

Like this article? Follow us on LinkedIn, Twitter, YouTube or Facebook to see the content we post.

Read more about Cyber Security

- The Good, the Bad and the Ugly in Cybersecurity – Week 11
- SentinelOne's Cybersecurity Predictions 2023 | What's Next?
- Investing in Tomorrow | Why We Started S Ventures
- Apple's macOS Ventura | 7 New Security Changes to Be Aware Of
- Decoding the 4th Round of MITRE ATT&CK® Framework (Engenuity): Wizard Spider and Sandworm Enterprise Evaluations
- Bringing Identity to the Era of XDR