DARKTRACE

RESEARCH WHITE PAPER

The Next Paradigm Shift

AI-Driven Cyber-Attacks

Executive Overview

Every day Darktrace encounters advanced cyber-attacks in customer environments around the world. Targeting every industry, these attacks range from opportunistic, commodity malware, such as IoT botnets using password-guessing techniques, to covert intrusions persisting for several months, driven by highly-skilled human actors. In recent years, multiple aspects of commodity malware have become more sophisticated.

This report aims to examine the present state of the threat landscape and explore the role narrow artificial intelligence (AI) will play in cyber-offense, as well as the motivations behind its development. It will present three scenarios where advanced threats have achieved high levels of sophistication for specific, isolated characteristics. We have observed and caught these threats 'in the wild' over the last 12 months using Darktrace's Enterprise Immune System. By carefully examining what makes these threats so dangerous, we can extrapolate from them and predict what similar attacks, charged with AI and contextual awareness, will look like.

The scenarios in this report were selected based on the level of sophistication demonstrated. They act as the 'what-is' inventory from which we extrapolate to 'what-will-be' scenarios. Each threat covers a different phase of the attack lifecycle:

- Lateral movement
- Command & control traffic
- Data exfiltration

AI-driven or human-driven attacks?

What differentiates today's opportunistic, commodity malware from targeted, human-driven attacks? Opportunistic malware lacks the understanding and contextualization of a human attacker. While certain characteristics might be deemed clever, e.g. how it moves laterally or how it exfiltrates data, this kind of malware is mostly static according to its source code once it has been unleashed. But what if malware does not have to rely on human pre-programming to understand context for decision-making, but can do this on its own?

The most advanced human attackers try to blend into their target environment as much as possible. In order to do this, the human actor understands what normal behavior looks like and adjusts his/her techniques accordingly to 'live off the land'. It is entirely possible for malware to autonomously acquire this contextual understanding using AI, then adapt its behavior on the fly to blend into a target environment.

Therefore, we expect AI-driven malware to start mimicking behavior that is usually attributed to human operators by leveraging contextualization. But we also anticipate the opposite, i.e., advanced human attacker groups utilizing AI-driven implants to improve their attacks and enable them to scale better. We will explore the benefits of using AI-driven malware for cyber-attackers further in the conclusion. Ultimately, it is of little significance to for the defender whether a majority of the hack was carried out by a human, AI, or both.

For blue teams, it will be increasingly difficult to differentiate between the two during investigations, as they will start to blend into each other.

Max Heinemeyer
Director of Threat Hunting, Darktrace

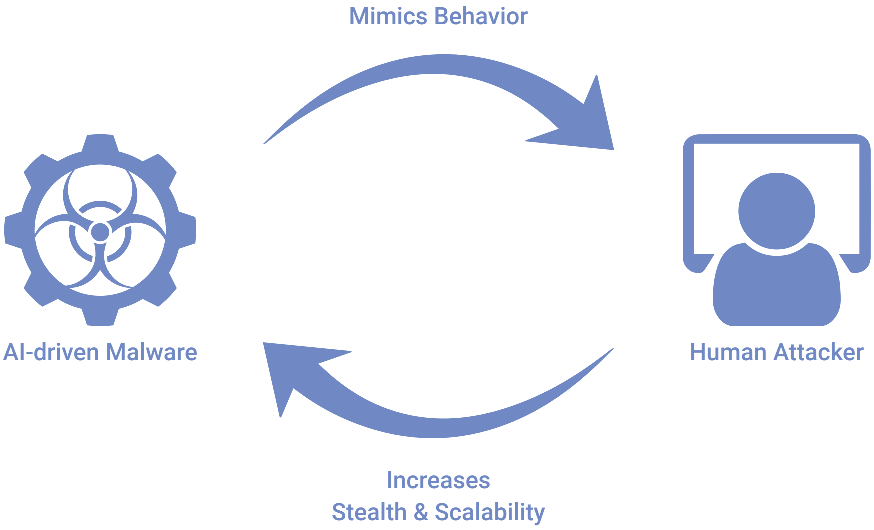

Figure 1: Reciprocal improvements in AI and human-driven cyber-attacks

AI-driven case studies

1. 'Autonomous' malware

Threat Discovered by Darktrace

An employee at a law firm fell victim to a spam campaign which led to an infection with Trickbot malware. Trickbot is opportunistic, information-stealing malware that, having gone through several iterations, remains under active development. The malware is modular and contains worming functionality that utilizes SMB exploits similar to WannaCry.

Within minutes of infecting patient zero, Trickbot had spread to over 20 other devices on the network with outdated SMB services. While the security team immediately detected the Trickbot infection, containing and cleaning up the attack was costly. The human security team could not react fast enough and the worming module kicked in.

Several hours after the initial infection, two machines showed signs of the Empire Powershell post-infection framework. This was a new development we had not seen before when first observed in June 2018. Trickbot is usually opportunistic malware that tries to infect as many victims as possible utilizing more 'traditional' C2. Empire is commonly used by advanced human attackers to facilitate covert hands-on-keyboard attacks.

The Empire Powershell modules were likely added to Trickbot for better persistence and to facilitate easier access to target systems for human attackers operating the Trickbot infrastructure. This allows the attackers to conduct manual intrusions conveniently, using Empire as a post-infection activity for high-value targets.

While Trickbot is best known for its ability to steal banking information, malware authors have started experimenting with locking modules. The attackers are likely diversifying their payloads to maximize profits – steal banking details first, then lock the victim out of their computer so they have to pay a ransom to get back in.

Darktrace alerted on the malware being downloaded from an unusual external source on the internet, the C2 channel was highlighted and the lateral movement identified immediately. It played an invaluable role in the clean-up operation and highlighted the post-infection Empire activity as soon as it started.

Indicators of Compromise

Trickbot SHA256: df53e89d8c19bce32f92a0778fc142e13
00f42e560c774aba06a481334711c34

Future AI Attack Scenario

In the future, AI-driven malware will self-propagate via a series of autonomous decisions, intelligently tailored to the parameters of the infected system. Imagine a worm-style attack, like WannaCry, which, instead of relying on one form of lateral movement (e.g., the EternalBlue exploit), could understand the target environment and choose lateral movement techniques accordingly. If EternalBlue were patched, it could switch to brute-forcing SMB credentials, loading Mimikatz or perhaps install a key-logger to capture credentials.

AI-driven malware will then choose whatever method appears most successful for the target environment and use this to move laterally. Instead of utilizing exploits, it might find PsExec is regularly used between certain devices at specific times of day. By learning this and then using PsExec for lateral movement, during times when it would normally be used, identification of the malware will become almost impossible. PsExec can of course be replaced by RDP, SSH or any other administrative toolkit that represents normal for a given environment.

The malware can learn context by quietly sitting in an infected environment and observing normal business operations, such as the internal devices the infected machine communicates with, the ports and protocols it uses, and the accounts which use it.

Able to make those decisions autonomously, no C2 channel will be required for the attack to propagate and complete its mission. By eliminating the need for C2, the attack will become stealthier and more dangerous.

Trickbot has displayed the first signs of utilizing multiple payloads for monetization – stealing banking details and locking machines for ransom. Malware authors can maximize their profits if their malware can choose autonomously which payload will yield the highest profit based on the context of the environment and infected machine.

As Trickbot is modular and under active development, why not add the capacity to make smarter decisions? Narrow AI can learn that if it infects the laptop of a VIP, such a user is likely to conduct a lot of email communication revolving around financial information. On a VIP's device, it will be more profitable to silently steal information or lock the machine and thus grind the company to a halt. However, if the malware identifies it has been dropped onto a server that is not processing any mission-critical information, it might just install a crypto- miner, as locking the server will only lead to investigation. Semantic analysis and contextual awareness allow software to make these distinctions and autonomously make these kinds of decisions.

How do we tell where an automated attack stops, and an interactive session starts? As this case of Trickbot leveraging Empire Powershell demonstrates, the previously clear distinction between automated, malware and human-driven attacks is no longer viable.

2. Intelligent evasion techniques

Threat Discovered by Darktrace

At a power and water company, Darktrace detected a device infected with malware that had taken steps to disguise its activity as legitimate. The attack utilized several techniques to stay covert.

A file was downloaded onto the device from the Amazon S3 cloud platform, which established a backdoor to the network. While establishing a backdoor is not uncommon, the program also showed signs of blending into the environment.

The malware utilized its own self-signed SSL certificate for windowsupdate.microsoft.com. The device then made subsequent HTTP(S) requests directly to an attacker-controlled IP address utilizing the fake Windows certificate. The communication was facilitated over ports 443 and 80, blending in with regular network traffic. Further Open Source Intelligence (OSINT) suggests that this particular threat actor utilizes alternative Doppelgänger techniques to reduce detectability in other infrastructure.

The self-signed certificate tricked traditional security controls and the activity was only picked up because Darktrace cyber AI distinguished the device's external communications as irregular, based on what it had learnt about that device's normal 'pattern of life'. While regular communications to Microsoft.com were seen in the Windows-dominated environment, the destination IP address for the HTTP(S) communication was very anomalous for the rest of the network, especially considering that a self-signed certificate was used.

Indicators of Compromise

Backdoor SHA256: b8ea95b3c66b9a121443ab76971 cdeec6dceda0ad54f8ce9b85fcc806f82a0ec

C2 IP: 212.69.36[.]86

Future AI Attack Scenario

Weaponized AI will be able to adapt to the environment it infects. By learning from contextual information, it will specifically target weak points it discovers, or mimic trusted elements of the system. This will allow AI cyber-attacks to evade detection and maximize the damage they cause.

Some of today's most targeted attacks attempt to blend in with their target's environment. Sometimes, this comes in the form of only establishing C2 channels during regular business hours, communicating over popular ports such as 53, 80 and 443 and leveraging standard protocols such as HTTP and HTTPS. Alternatively, attacks may use domain names for C2 communication and data exfiltration that closely resemble the target's own domain or company name. They may also utilize domain fronting for popular content delivery networks.

However, knowing what is normal for a given target environment is still mostly guesswork and assumptions made by the attacker. Malware that uses AI will be able to learn what constitutes normal as soon as the initial infection is successful.

Instead of guessing during which times normal business operations are conducted, it will learn it. Rather than guessing if an environment is using mostly Windows machines or Linux machines, or if Twitter or Instagram would be a better channel for steganographic C2 – it will be able to gain an understanding of what communication is dominant in the target's network and blend in with it.

3. Low-and-slow data exfiltration

Threat Discovered by Darktrace

In this scenario, Darktrace observed an extremely stealthy threat. Data was being exfiltrated from a medical technology company at such a slow pace, and in such small packages, that it avoided triggering the data volume threshold in legacy security tools.

The device in question was observed making multiple connections to a 100% rare external IP address on the internet over the course of 24 hours. Each connection was less than 1MB in size. However, the accumulated volume of hundreds of regular data transfers amounted to a significant breach.

The device had sent 15 GB of data to an attacker's external infrastructure. To further hide its malicious activity, the device utilized TCP ports 516 and 7897. While the use of non-standard ports seems suspicious, the target environment was very hectic, and a lot of high-ports were regularly used externally.

Darktrace AI learns the 'pattern of life' on a network over the duration of its deployment. When monitored continuously over the 24-hour period, the combined volume was clearly anomalous, allowing Darktrace to detect the outgoing data transfer. Darktrace immediately alerted the customer's security team to the ongoing data breach.

The data that left the company was sensitive information, including the names, addresses and medical history of patients. Had this succeeded without detection, the medical company's reputation would have been in jeopardy. In addition, a major breach would have exposed the organization to regulatory action, and further, personal, litigation. Darktrace was able to stop the breach from escalating into a crisis.

As no signs of a malware compromise were detected, this incident was most likely facilitated by a malicious insider. While it took a human attacker in this scenario to conduct a low-and-slow data exfiltration, there is no reason that automated attacks should not be able to replicate this behavior.

Future AI Attack Scenario

When we talk about machine-speed attacks, we generally imagine malware moving faster than humans can respond. The corollary, however, is just as dangerous; 'low-and-slow' attacks that evade detection because each individual action is too small for humans and traditional security tools to detect. While we looked at C2 and lateral movement techniques before, this scenario focuses on data exfiltration.

Most traditional security tools work in binary ways – was the upload bigger than 500MB? If yes, flag it as suspicious for later investigation. As this is known to attackers, they adapt their data exfiltration methods accordingly by chopping up the data into smaller chunks and gradually exfiltrating them over time. This data can be sent over the internet to either a single C2 server, or multiple different destinations.

Even though the example provided here was sophisticated enough to bypass all existing security tools except Darktrace, much stealthier data exfiltration scenarios are feasible. An attacker with a strong presence in the target's network does not have to conduct the exfiltration over the course of 24 hours – they could spread the exfiltration over 24 days if they wanted to.

Although it is already difficult for traditional tools to detect low-and-slow data exfiltration like this, it will become even harder once malware uses AI to understand the context of their target's environment. As soon as the malware no longer uses a hard-coded data volume threshold but is able to change it dynamically, based on the total bandwidth used by the infected machine, it will become much more efficient. Instead of sending out 20KB every 2 hours, it can increase the data volume exfiltrated during suitable times, e.g. when the employee whose laptop is infected is video-conferencing and sending out a lot of data anyway.

Another way that contextualization will help attackers is when deciding where to upload the stolen data. If video-conferencing is a common method of external communication in the target company, the malware could use such a video-conferencing system for data exfiltration, thus blending in with normal business operations.

Conclusion

This report demonstrates three scenarios of advanced cyber-attacks seen 'in the wild'. Each scenario shows isolated, sophisticated techniques to avoid detection – lateral movement at machine-speed, C2 that blends into the target environment, and low-and-slow data exfiltration.

The extrapolation of AI-driven attacks is entirely realistic. We see sophisticated characteristics in existing malware on the one hand – and narrow AI understanding context on-the-fly on the other. The combination of the two will mark a paradigm shift for the cyber security industry. Once the genie is out of the bottle, it cannot be put back in again.

In May 2017, we saw an initial paradigm shift when WannaCry was first seen in the wild. While worm-style attacks were long known in the cyber security industry, none were as effective as WannaCry nor did they use destructive payloads such as ransomware. After WannaCry, the cyber-crime ecosystem was quick to adapt its 'successful' techniques for other forms of cyber-attack such as worming crypto-mining malware, or banking trojans with the ability to move laterally. We anticipate similar copycat-like behavior once the first successful AI-driven attacks hit organizations.

What will motivate attackers to develop offensive cyber AI? Enthusiasm will be driven, in part, by the significant financial gains offensive cyber AI has the potential to bring. Better lateral movement capabilities mean hitting more devices with ransomware. Improved C2 methods will result in longer infiltration periods thus increasing the amount of data exfiltrated.

The other part of the answer is scalability. Human attackers can gain a high level of stealth due to their intelligence, situational awareness and decision-making capabilities. But they only scale so far – there is a finite number of capable hackers, even for nation states.

From a human capital perspective, advanced cyber operations are a highly intensive undertaking where several expert teams work in orchestration. For example, one team builds and maintains the C2 infrastructure, another team works on OSINT and keeping operational security high, hands-on operators use malicious implants and move through target networks, and backend analysts process exfiltrated data.

Being able to automate some of the most skill-intensive parts of these operations – like hands-on-keyboard intrusions – will result in a higher return on investment for attackers.

Companies are already failing to combat advanced threats such as new strains of worming ransomware with legacy tools. Defensive cyber AI is the only chance to prepare for the next paradigm shift in the threat landscape when AI-driven malware becomes a reality.

About Darktrace

Darktrace is the world's leading AI company for cyber security. Created by mathematicians, the Enterprise Immune System uses machine learning and AI algorithms to detect and respond to cyber-threats across diverse digital environments, including cloud and virtualized networks, IoT and industrial control systems. The technology is self-learning and requires no set-up, identifying threats in real time, including zero-days, insiders and stealthy, silent attackers. Darktrace is headquartered in San Francisco and Cambridge, UK, and has over 30 offices worldwide.

Contact Us

North America: +1 415 229 9100

Latin America: +55 11 97242 2011

Europe: +44 (0) 1223 394 100

Asia-Pacific: +65 6804 5010

info@darktrace.com

darktrace.com