AI-Augmented Attacks and the Battle of the Algorithms

Key Points

- Cyber-criminals will leverage offensive AI to create increasingly customized, targeted, and hard-to-detect cyber-attacks at scale.
- Open-source tools exist today to enable AI malware, and we will start to see these in the wild in the near future.
- AI will remove the human elements of the attack, making it harder for the perpetrators to be identified.
- Organizations will urgently need to employ AI defenses that can fight this new generation of attack on its own terms

Introduction

In discussions around the future of AI and cyber-threats, we often wonder when we can expect to see malicious or offensive AI attacks in the wild. While we have not yet seen conclusive evidence of execution, this report will show that all the tools and open-source research needed to facilitate an AI-augmented attack exist today. Therefore, we can anticipate that AI-driven cyber-attacks are not years away, but a very real possibility in the immediate future.

This report will document an end-to-end attack lifecycle, and how each stage could leverage elements of the AI 'toolkit' to improve and streamline the process. Attackers will, of course, evolve their tools to drive efficiency gains, however these tradecraft improvements are iterative and do not happen all at once. Furthermore, while it is likely that adversaries today are already leveraging AI in some capacity to improve individual attack phases, this report shows an end-to-end AI-driven attack purely as a thought experiment.

Cyber-crime gangs: an enterprise model

To illustrate how AI can be used to aid offensive capabilities, let's imagine a group of professional hackers dedicated to infiltrating a large organization. The criminals view themselves as cyber mercenaries working for the highest bidder, and have a team of around 15 people working for them remotely. Different members of the gang are specialized in different areas of expertise – there are social engineers, malware coders, hands-on intrusion operators, and post-breach data analysts.

Their crime-group is run like any other enterprise – each person carries out a different task, and in turn expects a return on their time investment. They have the same manpower restrictions as any other operation, and are always looking for ways to improve their attack efficiencies. Their victim is a military arms production company, and the adversaries are financially motivated – their main goal is to exfiltrate military specifications, weapon production data, and any other information they find along the way that could be sold or used for extortion or ransom demands.

In the following, the typical attack lifecycle waged against this hypothetical company will be examined. For each attack phase, we will first look at what traditional tools, techniques, and procedures look like. We will then compare it to the same attack phase being augmented by AI, and see how each of these attack stages can be substantially improved with existing tools and research.

Today's Attacks: Advanced but not Infallible

Attacks we see today often achieve success, but the threat actors must practice extreme care at every turn.

Stage 1: Trawling through social media

A team of human attackers create fake social media profiles over the course of several weeks during a reconnaissance period. They identify some of their targets manually or semi-automatically by crawling through LinkedIn, Instagram, and Twitter.

Some of the attackers carefully befriend some of the employees to gather more information about them on social media. This is a tedious and manual process.

In the meantime, another attack team is analyzing the victim's web presence looking for potential attack vectors. They are regularly slowed down by CAPTCHAs while browsing relevant websites, looking for vulnerabilities.

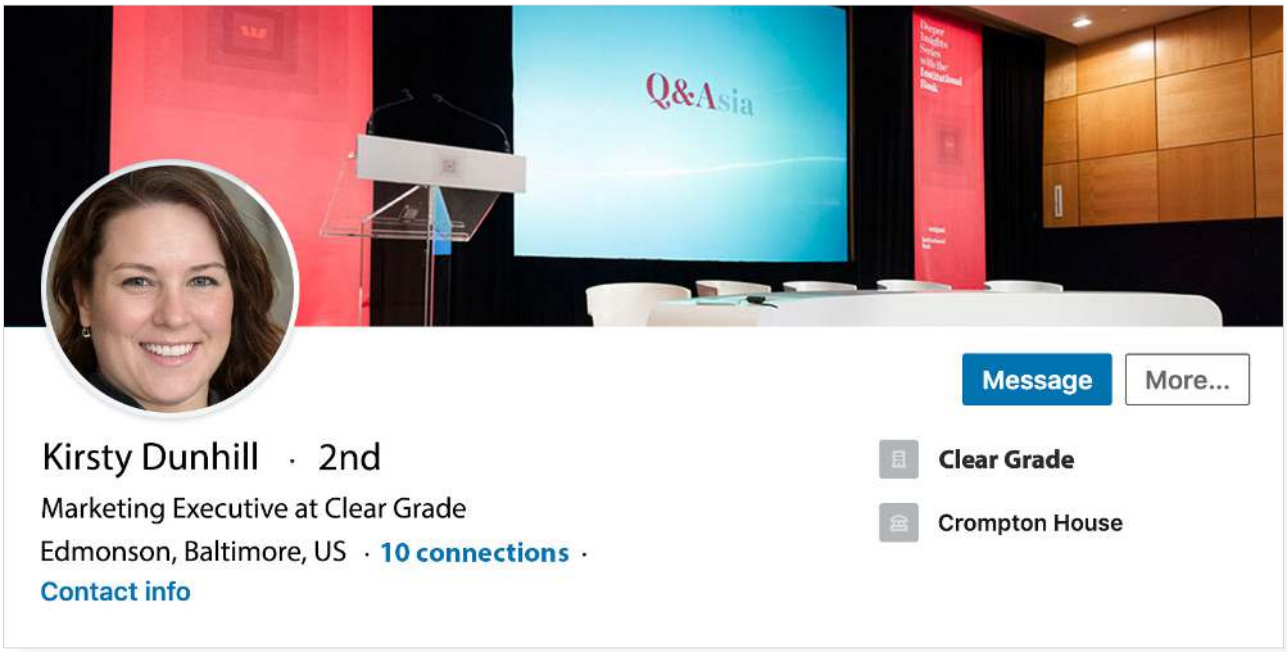


Figure 1: A fake social media profile used during reconnaissance

Stage 2: Spear phishing emails

Intelligence gleaned from social media is used for crafting spear phishing emails, which contain Office documents with malicious macros inside. These manually-crafted emails are based on the limited information gathered during initial reconnaissance, and not all of them are successful as the criminals missed some information on a few key employees – resulting in less believable phishing mails.

In one instance, the attackers fail to notice a change of job title, so use an outdated one in a targeted phishing email, prompting the security team to start an investigation.

The second attack team is now actively probing the victim's web servers looking for web-based vulnerabilities. They struggle as they are limited to known weaknesses and obvious gaps in the perimeter, and may not discover new, or hard-to-detect openings.

Stage 3: Malicious C2 channel detected

If the intrusion via phishing is successful, the malware establishes a command and control (C2) channel. They want to blend in with the target environment to avoid arousing suspicion, but their malware implant was hard-coded with specific C2 servers and ports. The attackers attempt to adapt the C2 behavior by manually observing the victim's network, but they lose some of their implants as the hard-coded external ports are blocked by the company's firewall.

Another infection is detected as the malware was pre-programmed to communicate only during US business hours – yet the infected machine that was discovered was being used in Europe, therefore operating on a completely different time schedule. The unusual, out-of-office-hours activity was detected by the security team. This was a costly attack phase for the cyber-criminals, who have to start this process again.

Stage 4: Brute-forcing passwords

Trying to achieve privilege escalation, the attackers then run keyloggers and try to loot the infected devices of their administrative credentials. They have some success finding several accounts that use weak passwords, but some of the more secure accounts take a very long time to brute-force with default password lists and dictionary attacks. The attackers are slowed down by these roadblocks.

Stage 5: Repetitive and arduous lateral movement

The harvested credentials are used to facilitate lateral movement. This is hit-and-miss – the attackers are able to successfully move laterally using pass-the-hash and Mimikatz. This process is repeated many times, the perpetrators are hacking one similar client machine after the next, trying to get hold of high-privilege accounts.

Every time credentials are obtained from a newly-compromised machine, the hackers analyze if this provides them with access to any new devices. This is a very manual process and requires heavy time investment from the adversaries.

> By exfiltrating more data than they need, hackers run the risk of tipping off the security team.

Stage 6: Packaging and exfiltrating data – a risky process

After a lot of lateral movement, the attackers have finally identified the data they were looking for. They also managed to get access to a database containing some files that appear to be related to military documentation. As the hackers can't sift through gigabytes worth of raw data, they decide to package all of it up and exfiltrate it in chunks out to their C2 server.

They plan to run the data analysis after the data is exfiltrated. This means that the vast majority of the exfiltrated data is useless for its purpose, as it may be totally unrelated to the sensitive military information that the attackers desire.

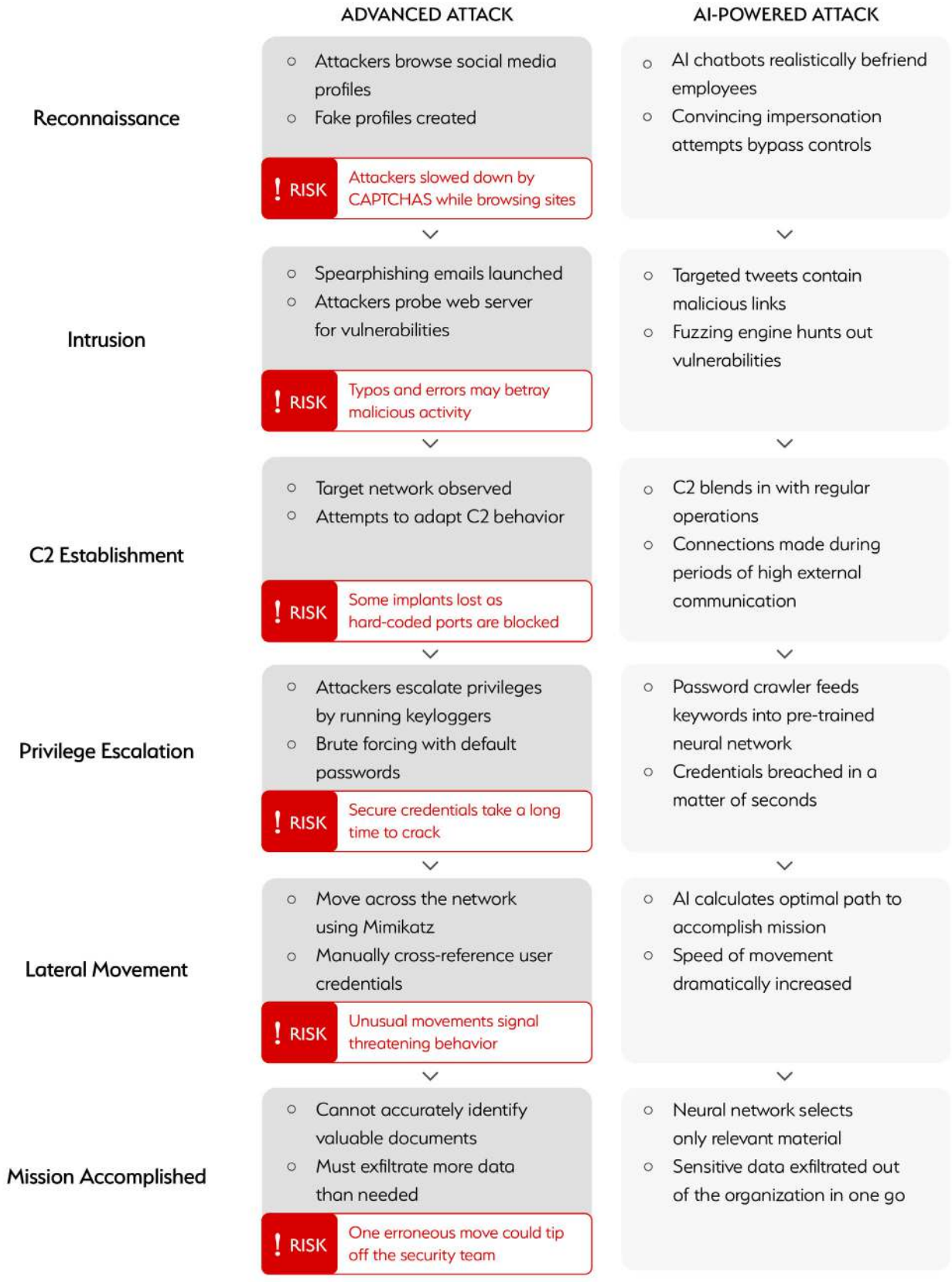
Instead of exfiltrating 100MB, they have to ship out several gigabytes of data, and by exfiltrating considerably more data than they actually need, the hackers run the risk of tipping off the security team.

They also miss some relevant data as the intrusion-operators are trained in hands-on-hacking, not in recognizing specialized weapon production or important military information. The attackers identified some potentially compromising material on VIP devices during their hunt for weapon blueprints, but don't have enough time to validate the authenticity of these documents.

While the mission did achieve partial success, the operation took several months and was very resource-intensive for the hacking group. They can afford to run only two such operations in parallel at most.

Anatomy of an attack

This infographic illustrates the attack lifecycle of this kind of advanced, yet non-AI-augmented cyber-attack, alongside an equivalent attack enhanced with the use of AI. It summarizes the tools and research that would be leveraged to achieve the augmentation.



AI-Augmented Cyber-Attack: The Next Generation

Let's now look at how an attacker might leverage AI tools to automate the traditional attack process, reduce risk factors and augment the yield.

Stage 1: Chatbots befriend the victim

Chatbots befriend employees of the target organization via social media – LinkedIn, Twitter, Instagram, and Facebook. The bots have previously learned what real social media profiles look like and have interacted with employees of the organization as well as creating believable content that appears genuine.

They use profile pictures of non-existent people created by an AI instead of re-using actual human photos. Once the chatbots have gained the trust of the victims at the target organization, the human attackers can gain valuable intelligence about its employees. At the same time, CAPTCHA-breakers are used for automated reconnaissance on the victim's internet-facing websites.

Stage 2: Customized phishing emails

The intelligence gathered from social media bots is then leveraged to craft convincing spear phishing attacks for an initial intrusion. An adapted version of SNAP_R is leveraged to create realistic tweets at scale and target several key employees. The tweets either trick the user into downloading malicious Office documents or contain links to servers which facilitate exploit-kit attacks. The AI classifier took all historic information for each individual target into account and was able to create highly targeted phishing messages at scale.

Meanwhile, an autonomous vulnerability fuzzing engine based on Shellphish is constantly crawling the victim's perimeter – internet-facing servers and websites – and tries to find new vulnerabilities for an initial technical foothold. The continuous fuzzing finally leads to an ephemeral test instance being discovered in the cloud – just a few minutes after it was created by a developer.

The machine is shut down after just two days, however this was enough time for the continuous crawler to find the new asset and for the autonomous fuzzer to find an exploitable vulnerability.

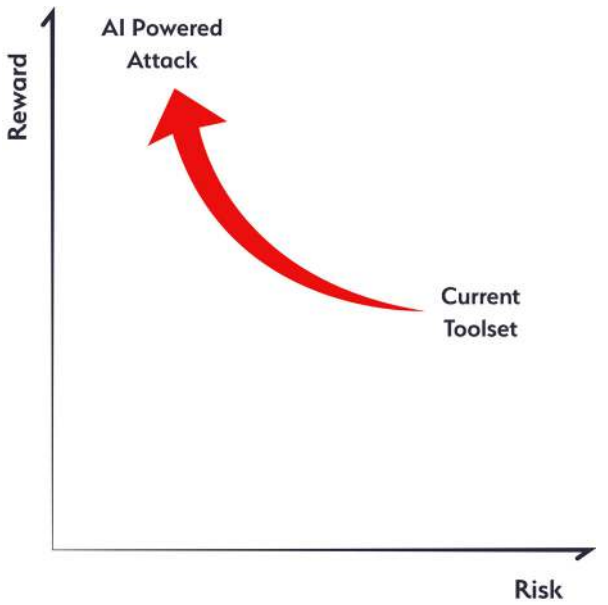


Figure 2: How adoption of AI increases yield and productivity

Stage 3: Mimicking regular business activity

Once the initial infection has taken place, either via a vulnerability discovered by the autonomous fuzzer or from automated social engineering via Twitter, the command and control (C2) channel is established. The popular hacking framework Empire is used to blend in with regular network operations, the malware sits and waits silently on the infected computer, learning its behavior.

The attackers have taken the idea of the statistic-based FirstOrder Empire module and have implemented an unsupervised clustering algorithm to learn what constitutes 'normal' on an infected device. It then auto-configures to replicate this 'normal' behavior, thus blending in with regular business operations and making it much harder to detect.

Thanks to this, the attackers avoid detection. The machine then uses some unusually high ports to communicate with specific APIs on the internet – the implant based on the augmented FirstOrder prototype has surfaced this as statistically significant and has auto-configured the malware to abuse this high port for C2 communication as it represents an exception on the firewall.

Stage 4: Passwords: challenging codes to crack

The Cewl tool creates a list of unique keywords based on the infected machine's documents and emails. It then feeds this basic password list of keywords into a neural network that was pre-trained on existing passwords, and uses supervised machine learning to create realistic permutations and potential passwords for advanced brute-forcing specific to the victim's context.

Even accounts that have strong, individual passwords can be cracked using this technique in a very short period of time.

Stage 5: Optimal pathways identified

Once accounts and passwords are obtained, lateral movement begins to get closer to the desired data. Moving laterally and harvesting accounts and credentials is an iterative process – identifying the optimal paths to accomplish the mission is critical to minimizing the intrusion time.

Parts of the attack planning can be accelerated by concepts from the CALDERA framework using automated planning AI methods. This greatly reduces the time required to reach the final destination.

Reconnaissance	CAPTCHA breaker
Intrusion	Shellphish SNAP_R
C2 Establishment	FirstOrder and unsupervised clustering algorithm
Privilege escalation	Cewl and neural network
Lateral Movement	MITRE Caldera
Mission Accomplished	Yahoo NSFW

Figure 3: The 'AI Toolbox' adopted by attackers

Stage 6: Successful infiltration of data

The mission is accomplished once engineering documents relating to the latest military technology and information are acquired. Instead of running a costly post-intrusion analysis operation to sift through gigabytes of data, a neural network pre-selects only relevant material for exfiltration. The neural network was trained on schematics, CAD drawings, and text-based documents containing 'weapon-related material', and therefore has a basic understanding of what such material constitutes and flags those for immediate exfiltration.

Furthermore, the Yahoo NSFW neural network is also abused, so that any compromising material downloaded onto corporate devices can be identified – no high-ranking VP wants word to get out that they have inappropriate images or documents saved onto their work computers.

As most of the steps in the attack lifecycle are automated or AI-augmented, the same attacking team that was confined to run a maximum of two in-depth operations in parallel without AI-augmentation can now run up to 200 in parallel with the same manpower as before – and with even better results. Instead of doing the labor-intensive manual work during the attack, the hackers now leave the heavy lifting to the machines and focus more on supervising the involved attacking tools than actually facilitating the majority of the hands-on intrusion tasks themselves.

> Attackers utilizing AI can run up to 200 operations in parallel – with significantly improved results.

Conclusion

We have to stop fighting yesterday's war, and look ahead to what we're going to be facing tomorrow. The motivation to develop better hacking tools clearly exists, as it enables attackers to scale better and increase their return on investment with intrusions. But this report establishes what an adversary could do with capabilities currently available in the public domain to augment their tradecraft with AI.

It is not far-fetched to hypothesize what nation-state groups and other heavily-funded cyber-adversaries might have developed behind closed doors. We have also seen a push for the adoption of Open Source tooling and Open Source communities, instead of just relying on in-house tradecraft.

To anticipate a further evolving threat landscape incorporating offensive AI attacks, we have to start adopting defensive AI solutions. The best adversaries are already trying to live off the land and blend in with regular operations. This will only increase once AI-methods are increasingly adopted by attackers.

Darktrace is able to detect even the most subtle breadcrumbs of an attack – this finally puts the advantage back in the hands of the defenders, as every small mistake caused by an attacker will show up to the security team or instantly be stopped by Antigena, Darktrace's Autonomous Response technology.

Only AI can fight AI, and the best algorithms will win. Cyber defense is an ongoing battle, and Darktrace cyber AI is leading the charge, allowing the human responders to take stock and strategize from behind the front line. A new age in cyber defense is just beginning, and the effect of AI on this battleground is already proving fundamental.

Autonomous Response

Autonomous Response is an AI technology created by Darktrace, the world-leading AI Platform for cyber defense. It is the only technology capable of taking action against in-progress cyber-attacks, when security teams need it most – whether that be over the weekend, at night, or simply when there is nobody around to respond to rapidly-spreading threats.

Darktrace Antigena, the world's first Autonomous Response solution, delivers a targeted, proportionate response as soon as it detects significant anomalous activity, containing the threat without interrupting daily business operations. It can work across every environment, no matter the complexity.

Over 3,000 customers across all industry verticals use Darktrace to help defend their networks across every corner of their organization. Every 3 seconds, Darktrace Antigena responds to a cyber-threat.

Autonomous Response represents a major step forward in the development of cyber AI defense, as it offers, for the first time, the possibility of a 'self-healing' network. Reacting in seconds, it defends organizations against even the most advanced attacks, on a 24/7 basis.

Available across email, network, IoT, and cloud, Antigena is part of a wider, data-agnostic Cyber AI Platform that works across an organization's entire digital infrastructure. Autonomous Response allows human security teams the time to focus on what really matters, while keeping attacks at bay.

About Darktrace

Darktrace is the world's leading cyber AI company and the creator of Autonomous Response technology. Its self-learning AI is modeled on the human immune system and used by over 3,000 organizations to protect against threats to the cloud, email, IoT, networks and industrial systems.

The company has over 1,000 employees and headquarters in San Francisco and Cambridge, UK. Every 3 seconds, Darktrace AI fights back against a cyber-threat, preventing it from causing damage.

Contact Us

North America: +1 (415) 229 9100

Europe: +44 (0) 1223 394 100

Asia-Pacific: +65 6804 5010

Latin America: +55 11 97242 2011

info@darktrace.com | darktrace.com

@darktrace