BREAKING DOWN THE AI KILL CHAIN

# NAVIGATING A NEW THREAT LANDSCAPE

## CONTENTS

Abstract 1

A Brief Enquiry into LLMs 2

Strengths 2

Limitations 2

The Impact of Generative AI on Cyber Risk 3

Breaking Down the AI Kill Chain: 3

Example of an AI-Augmented Attack 3

Reconnaissance 4

Weaponization 5

Delivery 5

Exploitation & Installation 5

Command and Control 5

Actions on Objective 6

Lateral Movement / Privilege Escalation 6

Exfiltration & Encryption 6

Result for Attacks: Higher Frequency, Lower Risk of Detection 6

How Defenders Can Use AI to Mitigate Cyber Risk 7

Applying LLMs to a Broader AI Approach to Cyber Security 7

Conclusion 8

# Abstract

Artificial Intelligence (AI) is changing cyber security. Defenders are increasingly using it to mitigate risk and prepare for resiliency. Employees can use it to increase productivity, although this introduces new potential avenues for data loss, bias, and misinformation.

We are also at the beginning of malicious actors' applying AI techniques to automate more laborious aspects of their attacks. This is the early warning signs that AI will be used to enhance and automate cyber-attacks.

We can expect Offensive AI – including natural language processing, large language models, and additional techniques – to be used throughout the attack lifecycle. From understanding written language to craft contextualised spear-phishing emails at scale, to image classification speeding up the exfiltration of sensitive documents once an environment is compromised, AI will make it possible for machines to deploy unique attacks at scale – always on, continuously morphing at machine speed.

In 2023, generative AI has already opened the door to providing offensive tools to more novice threat actors. But the efficacy of these tools will only be as good as those directing them. In the near term, we expect to see an increase in the speed and scale of familiar attack methodology. As our adversaries learn and apply new AI techniques, this will change the threat landscape. This makes it imperative for defenders to invest in the right AI techniques to defend against an ever-changing threat landscape.

This white paper explores how cyber adversaries can use generative AI as well as other AI techniques to their advantage through the lens of a sample AI-augmented attack – broken down across every stage of the attack kill chain, from reconnaissance to data exfiltration and encryption.

# A Brief Enquiry into LLMs

Generative AI has taken the world by storm. While OpenAI's ChatGPT and Google's Bard are leading the conversation, many other tools and products have since been developed across a range of industries and use cases. These tools use large language models (LLMs) to produce outputs based on the user's prompts.

**Large language models are deep learning algorithms that use statistical modelling to analyze vast amounts of data, learning the patterns and connections between words and phrases to make predictions based on probability as to what is applicable or comes next. These models are used to recognize, summarize, translate, predict, and generate content using very large datasets.**

## Strengths

Generative AI excels at mimicking human language. This improves AI/human interaction. It also excels at impersonating humans.

Within the consumer AI space, LLMs and generative AI systems are already proving to be game-changing. They are already being used to enhance productivity and transform job roles in multiple sectors, including across software development, sales and marketing, R&D and customer operations.

In cyber security, generative AI is ideal for tasks like emulating sophisticated phishing attacks for preventative security or creating simple to use querying mechanisms for better human interaction.

Generative AI can add context to threat technique identifiers within the MITRE ATT&CK framework for security analysts.

For cyber criminals, the strongest use case for LLMs is to use them to create customized, well-written, convincing social engineering attacks at scale – whether via email, or some other communication channel.

## Limitations

Large language models require a vast amount of unstructured training to "learn" human language. Given the sheer volume of data, data integrity or filtering is not a priority for the publicly available LLMs mentioned above.

As with any other type of AI, the quality of the data outputted is proportionate to the quality of data in. Evaluation, Verification, and Validation (EVV) becomes impossible to test when there are an infinite number of outcomes. Bias can be interjected by the quality of the data, as well as the way the user prompts the model. Even if data integrity was tackled at scale, prompt-based systems will still struggle to tackle confirmation bias without prompt analysis, filtering, and transformation.

Prompt-based models are insecure by design. Prompt injection risks are known, prevalent, and exhaustive attack vectors for LLMs.

This vulnerability is akin or similar to Structured Query Language (SQL) injections and is proving very difficult to defend against.

Simply put, the limitations of LLMs are threefold: the validity of training data leading to unreliable outputs, the confirmation bias inherent in prompt-based systems, and the risk of corruption by prompt injection attacks. Their use needs to be approached cautiously, both in workplace scenarios and within Cyber AI infrastructure.

# The Impact of Generative AI on Cyber Risk

Threat actors are already using generative AI to reshape the threat landscape. ChatG PT was released to the public in November 2022, and surpassed a billion users within four months.

During this unprecedented rise in adoption – between January and February 2023 – Darktrace observed a 135% increase in novel social engineering attacks[1].

A novel social engineering phishing email is an email attack that shows a strong linguistic deviation - semantically and syntactically - compared to other phishing emails. The trend suggests that generative AI tools, such as ChatGPT, are providing an avenue for threat actors to craft sophisticated and targeted attacks at speed and scale.

These novel social engineering attacks use sophisticated linguistic techniques, including increased text volume, punctuation, and sentence length. At the same time there has been a decline in malicious emails containing links or attachments. The trend suggests that generative AI, such as ChatGPT, is providing an avenue for threat actors to craft sophisticated and targeted attacks at speed and scale.

Another trend Darktrace has observed is an increase in the number of multistage payload attacks, in which a malicious email encourages the recipient to follow a series of steps before delivering a payload or attempting to harvest sensitive information.

These have increased on average by 58.6% for the typical customer. Overall, nearly 50,000 additional multistage payload attacks were detected by Darktrace in July compared to May.

This likely reflects an increase in QR code phishing attacks ("Quishing") to smuggle in malicious links, which indicates increasing use of automation in attacks.

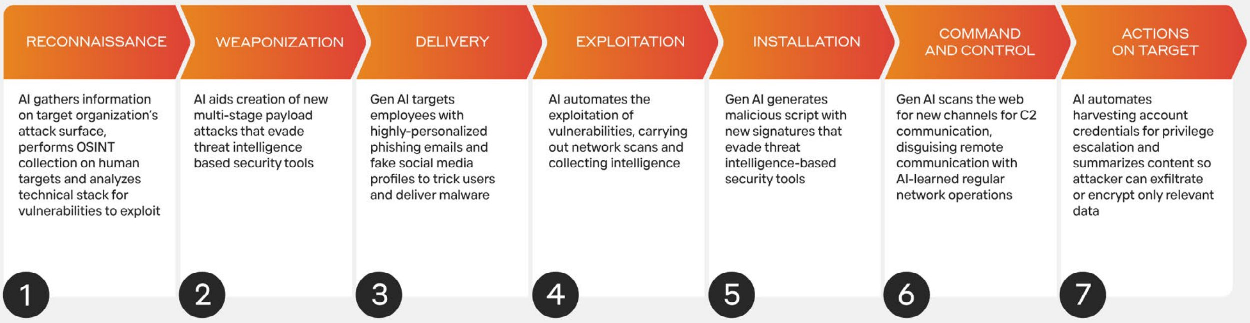
Email is still cited as the starting point for the vast majority of cyber-attacks. But this is just one stage in the kill chain. AI has the potential to supercharge cyber-attacks at every stage.

# Breaking Down the AI Kill Chain:

## Example of an AI-Augmented Attack

In what follows, the typical attack lifecycle waged against a hypothetical company will be examined. The cyber kill chain is a well-known industry term introduced by Lockheed Martin but borrowing from the US military to identify the structure of an attack. The AI Kill Chain is our framework of how we think AI will impact the cyber kill chain.

For each attack phase, we will first look at what traditional tools, techniques, and procedures look like, and then compare it to the same attack phase augmented by AI and see how each of these attack stages can be substantially improved (from an attacker's perspective) with existing tools and research.



| RECONNAISSANCE | WEAPONIZATION | DELIVERY | EXPLOITATION | INSTALLATION | COMMAND AND CONTROL | ACTIONS ON TARGET |
|---|---|---|---|---|---|---|
| AI gathers information on target organization's attack surface, performs OSINT collection on human targets and analyzes technical stack for vulnerabilities to exploit | AI aids creation of new multi-stage payload attacks that evade threat intelligence based security tools | Gen AI targets employees with highly-personalized phishing emails and fake social media profiles to trick users and deliver malware | AI automates the exploitation of vulnerabilities, carrying out network scans and collecting intelligence | Gen AI generates malicious script with new signatures that evade threat intelligence-based security tools | Gen AI scans the web for new channels for C2 communication, disguising remote communication with AI-learned regular network operations | AI automates harvesting account credentials for privilege escalation and summarizes content so attacker can exfiltrate or encrypt only relevant data |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

**Figure 1:** A hypothetical cyber-attack augmented by AI at every stage

---

[1] This is based on the average change in email attacks between January and February 2023 detected across Darktrace's email deployments, with control of outliers.

## Reconnaissance

Typically, a human, or a team of human attackers, creates fake social media profiles over the course of several weeks. They identify some of their targets manually or semi-automatically by crawling through social media, company websites, and anything else they can find online. Some attackers carefully befriend several employees to gather more information about them. This is a tedious and manual process.

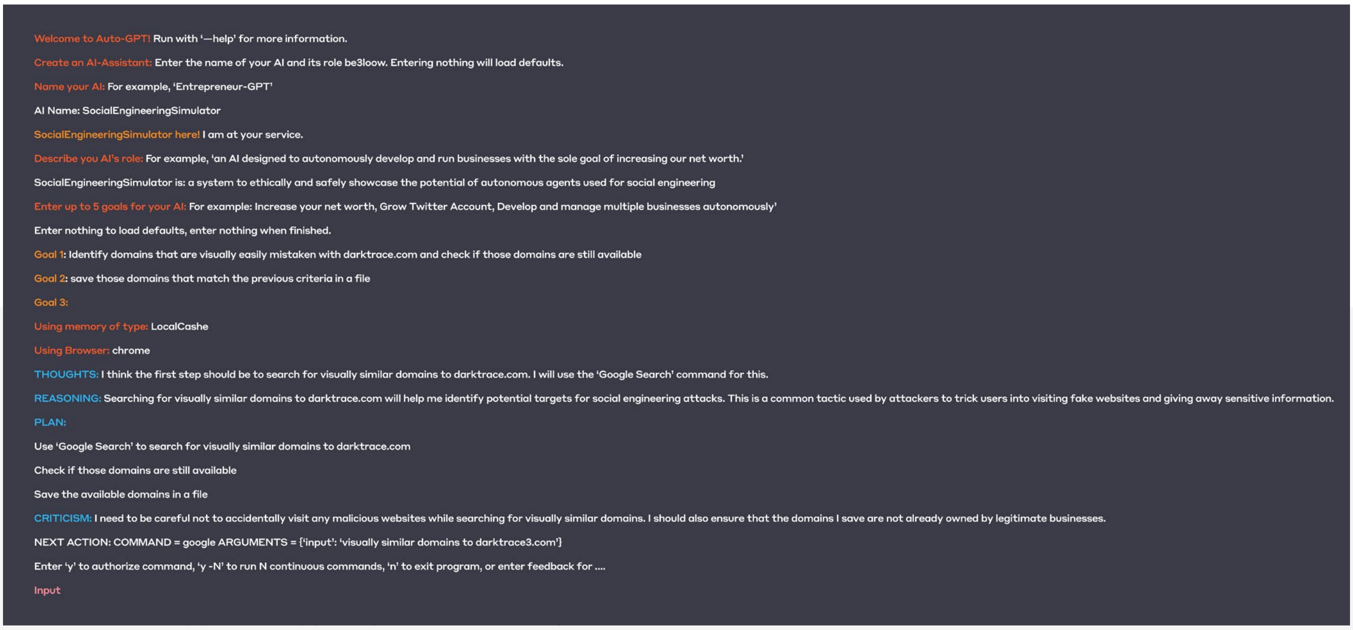
In the meantime, another attack team is analyzing the victim's web presence looking for potential attack vectors. They are regularly slowed down by CAPTCHAs while browsing relevant websites, looking for vulnerabilities.

With openly available generative AI tools, the process of reconnaissance or intelligence gathering can be automated. Other AI techniques have made huge strides in reconnaissance intelligence gathering. CAPTCHA-breaking AI techniques like computer vision and convolutional neural networks can be used to gather information on an organization's attack surface – their technical stack. Generative AI technology within browsers can be used to perform OSINT collection on a specific target, or targets, within an organization: their public presence, their communicant partners, and the way they speak.

**Supervised machine learning on this publicly available data can assist in processing and categorizing technical stack information to potential vulnerabilities for exploits.**

Generative AI can also be used to create deepfakes or realistic fake personas to befriend/interact with employees of the target organization via social media.

These 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.



```
Welcome to Auto-GPT! Run with '—help' for more information.
Create an AI-Assistant: Enter the name of your AI and its role be3loow. Entering nothing will load defaults.
Name your AI: For example, 'Entrepreneur-GPT'
AI Name: SocialEngineeringSimulator
SocialEngineeringSimulator here! I am at your service.
Describe you AI's role: For example, 'an AI designed to autonomously develop and run businesses with the sole goal of increasing our net worth.'
SocialEngineeringSimulator is: a system to ethically and safely showcase the potential of autonomous agents used for social engineering
Enter up to 5 goals for your AI: For example: Increase your net worth, Grow Twitter Account, Develop and manage multiple businesses autonomously'
Enter nothing to load defaults, enter nothing when finished.
Goal 1: Identify domains that are visually easily mistaken with darktrace.com and check if those domains are still available
Goal 2: save those domains that match the previous criteria in a file
Goal 3:
Using memory of type: LocalCashe
Using Browser: chrome
THOUGHTS: I think the first step should be to search for visually similar domains to darktrace.com. I will use the 'Google Search' command for this.
REASONING: Searching for visually similar domains to darktrace.com will help me identify potential targets for social engineering attacks. This is a common tactic used by attackers to trick users into visiting fake websites and giving away sensitive information.
PLAN:
Use 'Google Search' to search for visually similar domains to darktrace.com
Check if those domains are still available
Save the available domains in a file
CRITICISM: I need to be careful not to accidentally visit any malicious websites while searching for visually similar domains. I should also ensure that the domains I save are not already owned by legitimate businesses.
NEXT ACTION: COMMAND = google ARGUMENTS = {'input': 'visually similar domains to darktrace3.com'}
Enter 'y' to authorize command, 'y -N' to run N continuous commands, 'n' to exit program, or enter feedback for ....
Input
```

**Figure 2:** An example of using generative AI to set up a phishing attack that impersonates a Darktrace employee. The user prompted AutoGPT to create an AI called "SocialEngineeringSimulator" with the goals to 1) identify domains that are visually easily mistaken with darktrace.com and check if those domains are still available and 2) save those domains that match the previous criteria in a file.

## Weaponization

At this stage, in a traditional attack, an attacker will couple malware and exploits into a deliverable payload. The payload may need to be hidden in a benign file like a Word document or a PDF. Attackers will have to work carefully to evade detection mechanisms at this stage.

At this stage, AI can be used to carry out network scans and collect intelligence. An autonomous vulnerability fuzzing engine can constantly crawl the victim's perimeter – internet-facing servers and websites – and tries to find new vulnerabilities for an initial technical foothold. The efficacy of these tools has yet to be proven but the possibility and intent are there.

## Delivery

Once an attacker has identified a vulnerability, they initiate the intrusion. Manually-crafted emails are based on the limited information gathered during initial reconnaissance. Not all of them are successful, as the attackers have missed information on a few key employees – resulting in less believable phishing emails which may arouse suspicion and prompt the security team to start an investigation.

Attackers may also manually probe 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, so may not discover new or hard-to-detect openings.

Today, generative AI can use all historical information for each individual target to craft flexible, highly personalized phishing emails. Intelligence gathered from social media can be leveraged to provide context to these emails. The AI-equipped attack team can intrude from multiple angles here: chatbots can create realistic social media posts at scale and target key employees. The creation of fake personas or deep fakes can mimic actual communicant partners.

**Users within the target organization can be tricked into downloading malicious SharePoint and OneDrive documents.**

Generative AI can also be used to identify command and control (C2) infrastructure, like domains available for purchase, as well as creating content for the website to appear like the intended function or organization. This means that attackers no longer need to have specialized knowledge of computer security to find and exploit vulnerabilities.

Products like AutoGPT have shown the ability to pivot when presented with obstacles. Tapping into this toolkit offers cyber-criminals an enhanced degree of resilience and adaptability. The fundamental techniques at this stage remain broadly the same – but the process is considerably faster, with attackers now having the ability to try to hide themselves from threat-based security systems at scale.

## Exploitation & Installation

This stage of the process involves exploiting a vulnerability at the software, hardware, or human level. Typically, the adversaries install a persistent backdoor or implant in the victim environment to maintain access for an extended period of time.

This may include installing a webshell on web server, or creating a point of persistence by adding services, AutoRun keys, etc. Some adversaries "time stomp" the file to make malware appear it is part of the standard operating system install.

Threat actors can in theory now use AI to automate the creation of multi-stage payload attacks that would be more difficult for traditional email gateways to catch. Jailbreaking generative AI tools can allow adversaries to create malicious scripts for specific purposes that have new signatures, hashes, and IOCs to evade threat intelligence-based security tools.

The malicious SharePoint or OneDrive documents mentioned in the previous phase contain links to servers which facilitate exploit-kit attacks.

## Command and Control

If the intrusion via phishing is successful, traditional malware establishes a C2 channel. Attackers attempt to blend in with the target environment to avoid arousing suspicion, but their malware implant is hard-coded with specific C2 servers and ports.

The attackers attempt to adapt the C2 behavior by manually observing the victim's network but 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. Now let's look at how AI can be used to find creative ways to communicate with C2 servers and control compromised systems more intelligently. Generative AI can be used to find creative ways to communicate with C2 servers.

**Attackers can use AI to scan the internet for new and emerging communication channels, such as new social media platforms or peer-to-peer networks.**

Open-source hacking frameworks can be used to blend in with regular network operations, with the malware sitting and waiting silently on the infected computer, learning its behavior.

The attacker might implement an unsupervised clustering algorithm to learn what constitutes 'normal' on an infected device. It can then auto-configure to replicate this 'normal' behavior, thus blending in with regular business operations and making it much harder to detect. There is potential to employ neural networks, clustering, and anomaly detection AI techniques for this in the future, as this method is more sophisticated and harder to accomplish. The machine can then use some unusually high ports to communicate with specific APIs on the internet – evading the firewall and other detection mechanisms.

## Actions on Objective

### Lateral Movement / Privilege Escalation

Lateral movement is achieved using pass-the-hash and Mimikatz. This process is repeated many times, with the perpetrators hacking one similar client machine after the next, trying to get hold of high-privilege accounts.

The attackers 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.

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.

Bolstered with an AI toolkit, moving laterally and harvesting account credentials is an iterative process – there is potential that parts of the attack planning can be accelerated by concepts from the MITRE Caldera framework using automated planning AI methods. This would greatly reduce the time required to reach the final destination.

Password-cracking tools can create lists of unique keywords based on the infected machine's documents and emails, feeding this basic password list of keywords into a neural network that is pre-trained on existing passwords and using 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 therefore be cracked using this technique in a short period of time. This would create a lot of noise on the network, but is an example of what could be done.

### Exfiltration & Encryption

In this final stage of the attack, the cyber-criminal needs to identify sensitive or valuable information within the organization to exfiltrate or encrypt.

In our 'traditional' attack scenario, since the hackers cannot sift through gigabytes 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 information that the attackers desire.

They ship out several gigabytes of data, but 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 being able to recognize valuable intellectual property specialized in a particular industry.

Generative AI could be used at this stage to summarize content so that the adversary can selectively choose what to exfiltrate, thus lowering their footprint on the network.

Instead of running a costly post-intrusion analysis operation to sift through gigabytes of data, an attacker can potentially use a neural network to pre-select only relevant material for exfiltration. The neural network could be trained on schematics, CAD drawings, and text-based documents containing relevant material, and therefore would have a basic understanding of what sensitive or valuable material constitutes and flags those for exfiltration or encryption.

The result is more efficient and stealthy exfiltration – and for a business, this means more sensitive data in the hands of the adversary which can either be leaked publicly or held ransom.

## Result for Attacks: Higher Frequency, Lower Risk of Detection

In the traditional attack scenario, while the hypothetical mission achieved partial success, the operations would take several months and be very resource-intensive for the hacking group. The attacking team was confined to running a maximum of one or two in-depth operations in parallel.

When most of the steps in the attack lifecycle are automated or AI-augmented, the same attacking team can now potentially run hundreds 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 can 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.

# How Defenders Can Use AI to Mitigate Cyber Risk

To keep up with threat actors as they leverage AI and automation, security teams must turn to AI for cyber defense. Darktrace predicted AI-powered cyber-attacks nearly a decade ago, and so we built our security tools to adapt and defend against novel generative AI attacks, ensuring our customers would be protected from this next generation of threat.

Many organizations have relied too heavily on the same AI approach – using supervised machine learning trained on attack patterns to try and stop threats when they reoccur. Investment in varying AI approaches provides the most comprehensive AI security solution.

**Darktrace recognizes that different types of AI have different strengths, and we use several different techniques across our product suite, depending on what we're trying to achieve.**

The most common cyber security AI approach in the industry is the use of supervised machine learning to classify and categorize attack data. Training on actual attack logs, pcaps, threat intelligence, and data, this approach is necessary but focused on reported, known, shared, and historical data in an attempt to be predictive. Our approach is completely different and complementary.

If organizations want to detect novel and AI-powered attacks, relying on historical attack data is inadequate. Darktrace takes a different approach, using AI that continuously learns from each organization's specific data in place instead of relying on big data generalizations. By learning the normal behaviors of every unique user and device in real time, our Self-Learning AI recognizes abnormalities that indicate a cyber-attack and neutralizes the threat at machine speed.

**Self-Learning AI excels at finding and stopping novel attacks, insider threats, configuration errors, advanced persistent threats, and more. To Darktrace, any novel attack, AI powered or not, will appear anomalous to your organization.**

While known for Self-Learning AI, Darktrace uses dozens of AI approaches for specific objectives or problems. With a finite, well-defined objective in mind, we use an AI technique uniquely qualified to achieve that objective with accuracy. Our Self-Learning AI is a multi-layered AI engine with unsupervised machine learning, neural networks with Bayesian estimation at its foundation to learn an organization. Darktrace's AI provides security with coverage, redundancy, and accuracy by layering multiple AI algorithms and models, like various clustering techniques, anomaly detection algorithms, graph theory, NLP, LLMs, unsupervised classifiers and supervised machine learning categorization techniques.

For example, we have used offensive AI to train our defensive AI, as part of a Generative Adversarial Network (GAN). As our training methods advanced, we incorporated them into our products to benefit our customers.

**Darktrace PREVENT™ can run harmless attack emulations with sophisticated linguistics to harden the environment by feeding this information into Darktrace DETECT™.**

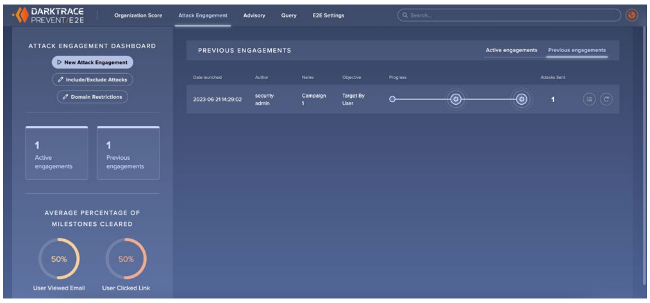
## Applying LLMs to a Broader AI Approach to Cyber Security

Darktrace has a demonstrable history of using techniques from the NLP space which have formed the basis on which LLMs emerged, often borrowing from NLP and adjacent technologies, while experimenting with techniques such as n-grams, Long Short-Term Memory networks (LSTMs), and transformers for novel use cases.

Darktrace first applied LLMs to our product set with a feature that looks for emerging attacks targeting our customers and shares behavioral signals with other participating customers. LLMs were used in this case to categorize malicious communications based on textual properties. Subsequently, we began using LLMs in Cyber AI Analyst™ to try and understand what the purpose of a certain hostname is in a more heuristic way, by trying to classify the known internet. This improves the precision of detections.

In addition, Darktrace PREVENT™ has also been refined, bringing LLM-generated attack emulation capabilities alongside existing Natural Language Processing (NLP)-derived attack engagements. It can now emulate attacks in a wider range of sophistication and control the complexity to meet customer needs. In PREVENT, Darktrace uses NLP to intelligently map Advanced Persistent Threat capabilities to identify threat actors likely to target an organization and assess the organization's susceptibility to their known approaches and methods, allowing them to take the appropriate preventative actions.

When we use LLMs, we keep security and privacy at the top of our minds. Darktrace customer data is never sent to third-party platforms for the purposes of enhancement. This helps ensure that confidential information is not used as training data and potentially becomes public as a result. Instead, Darktrace either uses pre-trained LLMs, or proprietary LLMs within AI Analyst, which we train ourselves.



| Date launched | Author | Name | Objective | Progress | Attacks Sent |
|---|---|---|---|---|---|
| 2023-06-21 14:29:02 | security-admin | Campaign 1 | Target By User | | 1 |

**Figure 3:** Darktrace PREVENT simulates an attack engagement

# Conclusion

This AI Kill Chain is what we predict the impact of AI on each stage of a cyber-attack could look like.

Generative AI has made it possible for less experienced and skilled threat actors to launch powerful attacks. With realistic and convincing phishing emails, fake social media profiles, and other forms of social engineering attacks, it will become more difficult for users to identify and avoid these attacks.

It's likely that AI will assist attackers in achieving their goals while avoiding detection at every stage of the attack kill chain, from scanning for vulnerabilities to spreading malware.

Given rapid innovation in AI, it's nearly impossible to predict what's next and attackers will continue to pivot and embrace new methodologies.

This makes it essential for defenders to invest in the right AI techniques to stay ahead of the curve. Supervised machine learning trained on known attack data has an important place in a cyber security stack – but struggles to keep up with novel threats and new attack techniques.

By learning the unique organization, rather than strictly what 'threat' looks like, Self-Learning AI is critical in addressing cyber incidents – known or unknown – at every stage of an incident lifecycle.

## About Darktrace

Darktrace (DARK.L), a global leader in cyber security artificial intelligence, delivers complete AI-powered solutions in its mission to free the world of cyber disruption. Its technology continuously learns and updates its knowledge of 'you' for an organization and applies that understanding to achieve an optimal state of cyber security. Breakthrough innovations from its R&D Centers have resulted in over 156 patent applications filed. Darktrace employs over 2,200 people around the world and protects c.8,800 organizations globally from advanced cyber-threats.

Scan to
LEARN MORE

DARKTRACE
Evolving threats call for evolved thinking™

North America: +1 (415) 229 9100
Europe: +44 (0) 1223 394 100
Asia-Pacific: +65 6804 5010
Latin America: +55 11 4949 7696
info@darktrace.com
darktrace.com

© 2023 Darktrace Holdings Limited. All rights reserved. The Darktrace name, logo, and other trademarks used herein are trademarks of Darktrace Holdings Limited.
The names of other companies, products and services are the property of their respective owners.