The CISO's Guide to Cyber AI

Categorizing the Use of AI in Cyber Security

CONTENTS

Abstract

The use of artificial intelligence (AI) in cyber-attacks is becoming increasingly common, with generative AI and large language model (LLM) tools opening doors to providing offensive methods to more novice threat actors.

As a result, the cyber security industry is increasingly implementing AI into defensive technology across prevention, detection, and response, and recovery, recognizing this as a necessity to effectively combat the evolving threat landscape.

However, not all AI is created equal: different types of AI have their own strengths and weaknesses when applied to different cyber security use cases. Most security solutions in this area rely on AI trained on known attack data, which is limited, as it is designed to recognize only the same or similar types of attacks it has seen before. In a future where novel cyber-attacks at speed and scale are the new normal, this supervised learning approach needs to be accompanied with AI that can recognize unknown threats.

This white paper categorizes the different applications of AI in cyber security and explains how Darktrace's approach to cyber AI uniquely understands the business in order to protect it from all types of attacks, including sophisticated and targeted ones created by generative AI.

The Era of AI-Powered Attacks

For some years, human security teams have struggled to keep up with the rate and pace of threats facing businesses.

They are managing increasingly complex digital environments spanning multiple clouds, networks, endpoints, and apps, with vast amounts of valuable and potentially vulnerable data. And the attackers are constantly innovating.

The widespread availability of generative AI has changed the threat landscape and opened new doors for attackers by making it possible for machines to deploy unique and sophisticated attacks at scale – continuously morphing at machine speed.

We can expect to see more novel phishing attacks, new automated creation of malicious code, sustained attack campaigns, and even deep fakes designed to elicit human trust.

Generative AI and other AI toolkits can also augment malicious actors at every stage of the attack kill chain.

Darktrace researchers found a **135% increase** in novel social engineering attacks from January to February 2023, corresponding with the widespread adoption of ChatGPT.

Businesses are also concerned about privacy issues, questioning how generative AI models are trained and whether that data is trustworthy and safe to use. One possible solution is to contribute their own data, however this poses the question: how can they use business data to train AI without putting it at risk of exposure?

Referencing business data in AI inputs and directives also carries risk. Increasing use of generative AI tools among employees can lead to inadvertent IP loss or data leakage. In one instance in May 2023, Darktrace detected and prevented an upload of over 1GB of data to a generative AI tool by an employee of one of its customers.

In this new threat landscape, security teams can't defend against novel attacks or new avenues of potential data loss with only a database of known attacks. Machine-speed defenses and comprehensive understandings of normal business activity are needed to counter these rapidly generated unknown attacks and insider threats.

AI Attack Tactics & Techniques

Various types of offensive AI have the potential to augment cyber adversaries at every stage of the attack kill chain.

Reconnaissance

- CAPTCHA-breaking AI tools can gather information on an organization's attack surface.
- Generative AI can perform OSINT collection on specific targets.
- Supervised machine learning can assist in processing and categorizing technical stack information for potential vulnerabilities.

Weaponization and Delivery

- Generative AI can create realistic fake personas to interact with employees of the target organization via social media or highly personalized phishing emails, tricking them into downloading malicious documents that contain links to servers which facilitate exploit-kit attacks.
- Generative AI can identify command and control (C2) infrastructure, like domains available for purchase, as well as creating content for the website to appear legitimate.
- AI can automate the exploitation of vulnerabilities, carrying out network scans and collecting intelligence.

Command and Control

- Generative AI can find creative ways 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 could blend in with regular network operations with the malware sitting and waiting silently on the infected computer, learning its behavior.

Lateral Movement / Privilege Escalation

- Password-cracking tools can generate lists of unique keywords based on the infected machine's documents and feed into a neural network that uses supervised machine learning to create realistic permutations and potential passwords.

Exfiltration & Encryption

- Generative AI can summarize content so that the adversary can selectively choose what to exfiltrate, thus lowering their footprint on the network.
- Image classification tools can speed up exfiltration by identifying sensitive documents from which attackers can profit.

For more information on how AI can be used to augment the attack kill chain, please refer to our white paper:

Navigating the New Threat Landscape.

The Modern Attack: One of One, not One of Many

In the past, cyber-attacks were often "one of many." Attackers would use known techniques and tools to target a variety of victims. However, new techniques and advances in offensive AI are making it possible for attackers to create "one of one" attacks. These attacks are entirely unique, tailored to the specific victim and their environment. This makes them much more difficult to defend against.

For example, attackers can use AI to generate custom malware that is specifically designed to exploit a victim's vulnerabilities. They can also use AI to automate the attack process, making it faster and more efficient. As a result, we expect cyber-attacks will become increasingly sophisticated and difficult to defend against.

The use of AI in cyber-attacks is still in its early stages, but it is rapidly evolving. AI-powered attacks will become more common, and they are expected to continue to grow in popularity in the future.

In order to protect themselves, organizations need to adapt their security to be able to cope with "one of one" threats. In practice, this means moving away from a perimeter-based approach that relies on rules and signatures: stopping activity from occurring if it associates an incident with something malicious it has witnessed before. Even if AI is used to automate this approach, it is still fundamentally stuck in the past, looking backwards at previously recognized threats and forever playing catch up.

By contrast, self-learning AI approaches learn what constitutes 'normal' by continuously analyzing every device, every user, and the millions of interactions between them, this type of AI can understand 'self' for an organization. Once it knows 'self,' it can piece together subtle deviations from 'self' and connect the dots of a cyber-attack.

This way, it can adapt and evolve at the same rate as threats, identifying unfamiliar and novel attacks.

Categorizing Cyber AI

The legacy approach to cyber security entails piping data from thousands of environments and storing this in large databases hosted in the cloud, where attack patterns can be identified, and threats can be stopped when they reoccur.

But when novel and targeted attacks are the norm, protection from known and previously encountered attacks is no longer enough.

One on one security learns from your enterprise data to protect you from all threats.

	Signature-Based, Leveraging Known Attacks	DARKTRACE
AI Type	Supervised Learning	Self-Learning Cyber AI
Data Source	Large data lake	Your business data
Speed	Processing results in latency	Real time - No latency
Privacy	High concern (Data sent to public cloud)	Lower concern (Data remains in place)
Data Cost Transfer	High	Low
Use Cases	✓ Known Attacks	✓ Known Attacks ✓ Unknown Attacks ✓ Nation State Attacks ✓ AI Attacks ✓ Insider Threats

Figure 1: Comparing the micro-view of your business data offered by Darktrace's Cyber AI with the macro view of legacy security tools

Known Attack Data / Supervised Machine Learning

One of the most common types of AI in security today is supervised machine learning models that are trained on known attack data and attacker behavior. Supervised machine learning is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately.

These models are commonly found in Extended Detection and Response (XDR) solutions. They are trained on massive volumes of structured, labeled attack data and threat intelligence, and they perform extremely well at stopping those known attacks. This makes them a good starting point for any security stack.

However, these models can fall short when they encounter something they haven't seen before. If the model hasn't been trained on a specific pattern, it can easily miss it. Additionally, co-mingled benign or legitimate data (syslog, network traffic, etc.) can cause big problems in the efficacy and accuracy of this AI's performance. And, just like most AI, testing and validation is crucial to ensuring accurate outcomes.

Generative AI and LLMs

Publicly available text-based generative AI systems, which are powered by LLMs, are pre-trained on massive volumes of internet data and can be applied to human language, machine language, and more. For consumers, LLMs are game-changing, as we are already seeing. However, many LLMs will likely remain the domain of existing Big Tech players, as training these systems requires access to massive volumes of data and computing, as well as significant manual labor for validating, cleansing, and preparing the data for processing.

Generative AI and LLM tools promise increases in productivity and new ways of augmenting human creativity. Generative AI excels at learning the fluidity of language and impersonating humans. This will be a large net gain in the future for translation abilities.

However, employee adoption of these tools introduces risks around privacy, especially through the lenses of data exfiltration, giving away business strategies or competitive advantages, and carries legal implications. For example, an employee could input proprietary information as part of a prompt to ChatGPT.

In security, generative AI can give defenders contextual understanding from a known environment on a massive scale and automate key tasks like summarizing or reporting on incidents. Generative AI and many LLMs are trained on billions of parameters of static data scraped from the internet. This makes them ideal for tasks like emulating sophisticated phishing attacks for preventative security or creating simple to use querying mechanisms for better human interaction.

However, if not applied responsibly, generative AI can cause confusion by "hallucinating," where it references invented data, or by providing conflicting responses due to confirmation bias in the prompts written by different security team members.

In addition, 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 SQL injections and proving very difficult to defend against.

Self-Learning AI

Self-Learning AI is a multi-layered AI approach including dozens of AI techniques and hundreds of models. At its core, Darktrace's Self-Learning AI has a foundation of multiple unsupervised machine learning techniques including neural networks, Bayesian meta-classification probabilistic models, various clustering techniques, large-scale computational regularization techniques and many others along with supervised machine learning models to learn in real-time an organization from the inside out.

AI is used to understand what is normal for an organization, device, account, user, and/or cluster as well as to identify anomalous behavior, patterns, and activity and respond in real time to contain misuse, misconfigurations, incidents, or anomalies. This AI is trained on each business's real-time data, allowing it to understand normal behavior and therefore identify the abnormal, including evergreen novel threats like zero days, insider threats, nation-state attacks, cloud and SaaS-based attacks, and even generative AI-powered attacks like sophisticated, scaled phishing campaigns.

Darktrace's Self-Learning AI was built around four core principles:

- It learns 'on the job' – it does not depend upon knowledge of previous attacks.
- It learns on real business data and thrives on complexity and diversity of modern businesses.
- It constantly revises assumptions about behavior, using probabilistic mathematics.
- It is always up to date and not reliant on human input.

Applying the Right AI to the Right Cyber Security Challenge

Core to the Darktrace approach to cyber defense is the belief that the right type of AI must be applied to the right use cases.

Self-Learning AI is central to Darktrace's approach and critical to identifying novel cyber-threats that most other tools miss. However, this is used in combination with different AI methods, including LLMs, generative AI, and supervised machine learning to support the Self-Learning AI. We build technology by looking at where AI can best augment the people in a security team and where it can be used responsibly to have the most positive impact on their work.

It will take a growing arsenal of AI to fight back in the age of offensive versus defensive AI.

We're excited about the potential of generative AI, and our team in the Cyber AI Research Centre is continually working on new features and functionality that we plan to make available across all our products to play to the strengths of new and emerging AI techniques including generative AI.

By developing and using new AI models and capabilities, including our own proprietary LLMs, alongside Self-Learning AI, Darktrace is helping customers to prepare for and fight back against increasingly sophisticated cyber threats – including the increased use of AI by cyber-attackers.

Here's how Darktrace applies various types of AI:

Self-Learning AI

Darktrace's self-learning technology is unique in its ability to understand and detect novel attacks. This makes it an absolute must-have in any future-facing security stack. For example, traditional email security tools that rely on knowledge of past threats take an average of 13 days from an attack being launched to detect the attack.

[1]By contrast, Darktrace/Email™, powered by Self-Learning AI, is capable of spotting and stopping threats as soon as they are launched.

Today, we are applying our Self-Learning AI to augment human security teams throughout each stage of cyber resilience:

- Prevention: With Darktrace PREVENT™, we are applying AI to help security practitioners harden security inside and out. The AI continuously monitors the attack surface for risks, high-impact vulnerabilities, and external threats. It also looks inside the environment to expose potentially vulnerable attack paths and high value targets.
- Detection: Darktrace DETECT™ uses AI to uncover threats within a network by analyzing thousands of metrics in real-time and revealing subtle deviations which a human wouldn't see but that may signal an evolving threat – even unknown techniques and novel malware that may bypass all other security controls.
- Response: Our AI enables Darktrace RESPOND™ to autonomously take action against attacks, at machine speed, bringing response times down from hours and days to mere seconds.
- Recovery: Darktrace HEAL™ can help organizations assess their readiness for an attack and practice with real-world scenarios. During an attack, AI helps to prioritize remediation actions to augment human teams. Post attack, our AI allows businesses to recover from cyber-attacks and get back to full operations faster and more confidently than a human team can alone.

Applied Supervised Machine Learning

Pages 4 covers the limitations of supervised machine learning in raw threat detection: when restricted only to historical attack data, a supervised machine learning engine fails to spot novel, never-before-seen threats. But in another application of cyber security, supervised machine learning can excel: specifically, automating the initial analyst triage of anomalies on the network, establishing which are dead ends, and which may be part of a wider and more significant security incident.

Darktrace has incorporated supervised machine learning into its product stack through its Cyber AI Analyst: an investigation tool trained on years of expert analyst behaviors, created to significantly expedite triage and reduce time-to-meaning for security teams.

Cyber AI Analyst™ takes individual anomalous events discovered by Darktrace's core Self-Learning AI, and then applies a second layer of AI to these findings, using supervised machine learning to piece together disparate signs of an attack and then prioritizing that for the human responder. It generates incident summaries that highlight every stage of the attack and includes a natural language summary that even a non-technical person can understand.

Security teams have seen a reduction in triage time by over 90% using this technology.

[1] Thirteen days mean average of phishing payloads active in the wild between the response of Darktrace/Email compared to the earliest of 16 independent feeds submitted by other email security technologies.

Natural Language Processing (NLP)

As AI becomes more widely adopted, explainable AI will become more critical than ever to clarify the decision-making roadmaps and outputs of other AI agents. It is essential not only to understand how valid AI models' outputs are, but also how the AI arrived at its conclusions. AI-driven natural language processing provides a clear explanation of the actions taken.

In addition to generating a coherent, easily-understandable reports on significant cyber security incidents, Darktrace's Cyber AI Analyst makes its investigation process transparent – showing what questions it asked before arriving at its conclusions.

This is in stark contrast to the 'black box' model that risks eroding trust between humans and AI, which can present problems with compliance and audit requirements.

Darktrace uses NLP elsewhere in its product suite to intelligently map Advanced Persistent Threat (APT) capabilities – to identify threat actors likely to target an organization and assess the organization's susceptibility to their known approaches and methods, allowing the organization to take the appropriate preventative actions.

Darktrace has a history of using techniques from the NLP space which have formed the basis on which LLMs emerged.

We use the best technique which has often been borrowing from NLP and adjacent technologies and experimenting with techniques such as n-grams, Long Short-Term Memory networks (LSTMs), and transformers for novel use cases.

LLMs

Page 4 covers the limitations of LLMs: the integrity of their outputs relies on the data on which they are trained. That is why the Darktrace models have been trained on our proprietary security data and are then applied to each customer's specific environment to build contextual datasets. This ensures both quality and relevance.

Darktrace models have been trained on our proprietary security data and are then applied to each customer's specific environment to build contextual datasets. This ensures both quality and relevance.

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 now uses LLMs to augment attack engagements designed to proactively harden defenses while providing security awareness. Bringing LLM-generated attack emulation capabilities alongside existing NLP-derived attack engagements allows attacks to be emulated in a wider range of sophistication and control the level of complexity to meet customer needs.

Darktrace has also been using LLMs in Cyber AI Analyst to better understand complex attack patterns, using enhanced LLMs that have been trained on security, network, and engineering data so that it can better identify anomalous behavior, services, and endpoints. This enables Cyber AI Analyst to provide even more precision for human analysts, continuously learning and improving autonomous investigations in real time.

These new capabilities bring several improvements, and quality-of-life changes for existing Darktrace customers – including to Cyber AI Analyst's proprietary LLM classifier to categorize malicious communications based on textual properties.

Crucially, these potential security incidents do not just relate to malicious activity by attackers attempting to breach networks from outside – they can help to detect data leaks by employees too.

Self-Learning AI in the Real World

For 10 years, Darktrace has been applying AI to solve real customer problems within the cyber security space. Over 8,000 organizations globally rely on Darktrace to protect themselves from cyber disruption.

For example:

- The City of Las Vegas uses Darktrace to save its cyber security team human-hours in threat investigation and response, as well as to spot the subtle signs of emerging cyber incidents while still in their early stages.
- Cruise giant Royal Caribbean relies on Darktrace to protect its fleet of cruise ships from a range of cyber-attacks, providing visibility across a diverse range of IT infrastructure and securing these floating smart cities.
- Darktrace's AI helped provide cyber protection to the 2022 Qatar World Cup. The security team used Darktrace to rapidly synthesize information

Figure 2: The City of Las Vegas Relies on Self-Learning AI to Protect its Infrastructure

Conclusion

Cyber defense needs to evolve at the same pace as that of attacks in order to protect organizations against the new wave of AI-driven cyber-threats. In the "Cost of a Data Breach Report 2023," IBM reported that the global average cost of a data breach was USD $4.45 million.

The potential for significant financial expsoure means businesses cannot afford to rely only on supervised machine learning or haphazardly apply generative AI to workflows, as these types of AI cannot protect against novel attacks or new avenues of data loss. Instead, organizations must apply the right types of AI to the right areas of their security stacks to best position themselves for the future.

That same IBM report found that the average savings for organizations that use AI in security is USD $1.76 million .

While it is supplemented by other techniques, at the core of Darktrace's products is a unique Self-Learning AI that trains in real-time on each customer's specific business data to determine patterns of life in that unique business to help identify and stop cyber-attacks, even unknown ones, and reduce cyber security risk.

Darktrace promotes the use of AI while providing methods to enforce safeguards in its application and explain its outputs to verify its accuracy.

As threat actors increasingly attempt to abuse AI for malicious purposes, augmenting humans with machine-based autonomous response is critical.

With a unique approach to cyber security and new features consistently added across our product suite, Darktrace helps augment human teams to protect against evolving attacks in real time – even against the latest, most innovative cyber threats. In this era of AI, Darktrace facilitates positive and efficient human-AI partnerships to customize robust cyber security that meets each organization's unique needs and policies.

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