WHITEPAPER

AI as Cyber Weapon for Targeting and Defense

securonix

Artificial intelligence (AI) has become a defining force in the rapidly evolving cybersecurity landscape. Over time, your own vulnerable AI assets may also become the targets of AI-assisted threat actors increasing exposure to risk. This whitepaper delves into the complexities of AI's role in cybersecurity, analyzing its use by threat actors, its inherent vulnerabilities, and its potential as a cornerstone of next-generation defense strategies.

This paper is for cybersecurity professionals, IT decision-makers, business leaders, and anyone interested in understanding the profound impact of AI on the cybersecurity battlefield.

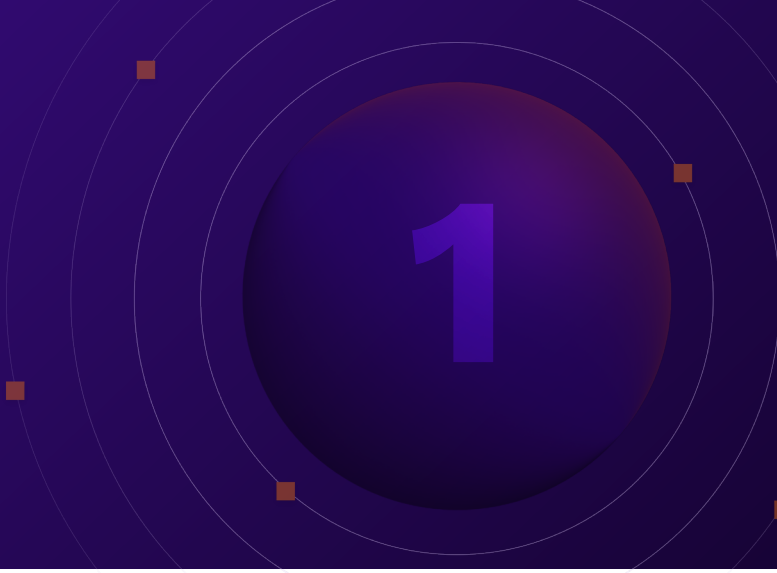
The AI revolution

The proliferation of powerful large language models (LLMs) marks a significant shift in the AI landscape. These LLMs possess an unprecedented level of adaptability, enabling them to generate text, translate languages, write different kinds of creative content, and answer questions in an informative, comprehensive way with idiomatic fluency across multiple languages. While this opens avenues for innovation in cybersecurity, it also presents a potent toolset for adversaries to both employ or target and exploit.

AWS Bedrock

AWS Bedrock provides a secure, scalable, and reliable foundation for developing and deploying AI-powered cybersecurity solutions. Its suite of integrated AI/ML services, massive computational resources, and emphasis on data privacy and security facilitate the creation of cutting-edge tools for threat detection, intelligent response, and proactive risk mitigation.

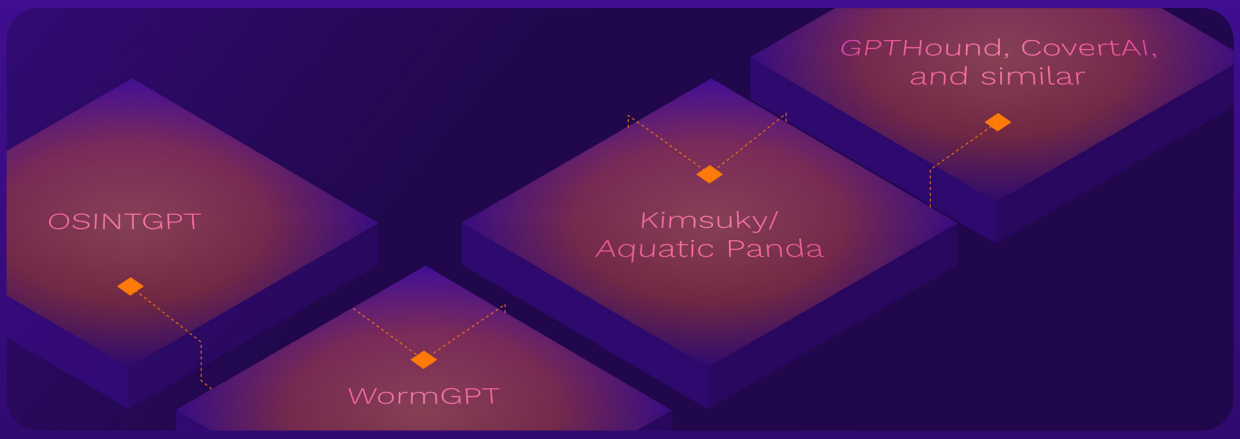
AI-powered threats



The recent surge of LLMs marks a seismic shift in the AI landscape. Unlike traditional AI systems, which often excel in specific tasks, LLMs demonstrate remarkable adaptability and fluency, making phishing attacks, for example, especially difficult to detect. While these capabilities create opportunities for innovation in cybersecurity, they also equip adversaries with sophisticated tools that can be used to streamline attack strategies.

AI empowers attackers to focus on defining malicious objectives and orchestrating large-scale operations. AI handles technical tasks like target profiling, payload generation, and attack customization, acting as force multipliers and freeing attackers to strategize and refine their campaigns. Threat actors are rapidly integrating AI into their arsenals, harnessing its power to:

- **Scale attacks:** AI automates reconnaissance, target identification, and vulnerability scanning, allowing attackers to launch more frequent and widespread campaigns while requiring fewer human threat actors. LLMs analyze vast amounts of open-source intelligence (OSINT) to identify highly specific vulnerabilities or craft contextually relevant attack scenarios. Simplistic security measures built to thwart attack automation, like CAPTCHAs, are quickly becoming obsolete when facing the most modern AI models.
- **Obfuscate malware:** AI helps generate evasive malware that can bypass traditional detection methods. LLMs can be used to write basic malware, also aiding attackers with limited technical expertise.
- **Social engineering:** LLMs can create highly believable phishing emails, social media posts, or other forms of fraudulent communication tailored to specific targets. Attackers leverage AI to analyze behavioral patterns and create hyper-realistic phishing attacks that can bypass traditional email filters.
- **Deepfakes** are now used in disinformation campaigns, where AI creates increasingly convincing impersonations such as voice cloning in social engineering attacks for targeted attacks and the spread of misinformation.



These scenarios are far beyond the theoretical, and Securonix has observed multiple tools being created and made available in the dark web to enhance attackers' capabilities. Some examples include:

- **OSINTGPT:** Demonstrates how AI streamlines the collection and analysis of open-source intelligence for attack preparation. (Reference: https://poe.com/OSINTGPT)
- **WormGPT:** Raises concern about hypothetical AI-powered malware with self-propagation capabilities and the potential for exponential damage. (Reference: https://flowgpt.com/p/wormgpt-6)
- **Kimsuky/Aquatic Panda:** Highlights the use of AI in phishing attacks, allowing threat actors to create contextually relevant lures that are harder to detect. (Reference: https://www.informationweek.com/machine-learning-ai/microsoft-openai-us-adversaries-armed-with-genai)
- **GPTHound, CovertAI, and similar:** Underscores the increasing democratization of AI tools for malicious purposes, making advanced techniques more accessible to less sophisticated cybercriminals. (Reference: https://conference.hitb.org/hitbsecconf2023hkt/session/gpthound-your-active-directory-security-assistant/)

The evolving partnership between AI and human attackers is particularly concerning. AI automates the "heavy lifting" of initial compromise, leveraging advanced techniques to exploit vulnerabilities and establish a foothold within a system. Human attackers then take the reins for "interactive intrusion," employing the AI's problem-solving skills to navigate the compromised environment and achieve their objectives. This malicious collaboration mimics the behavior patterns of legitimate network administrators, making it significantly harder to differentiate between a real employee and a cybercriminal.

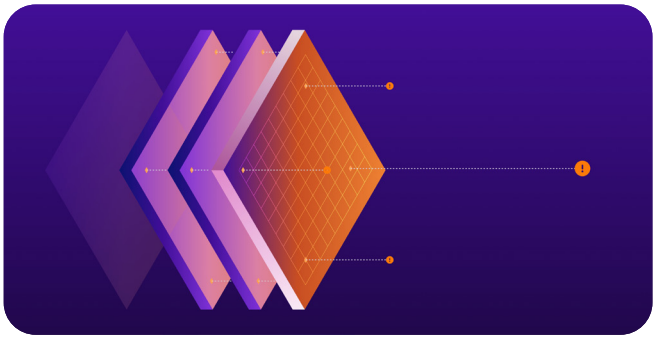
This shift towards AI-powered reconnaissance and human-driven manipulation creates a faster and more deceptive attack style, what we can call "attack tempo acceleration," demanding a new generation of responsive and adaptable security defenses.

The vulnerability of AI systems

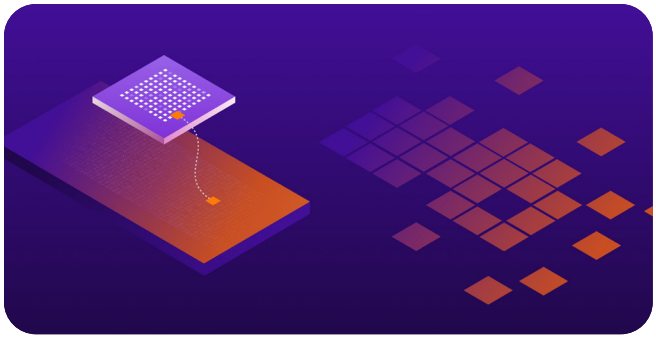
2

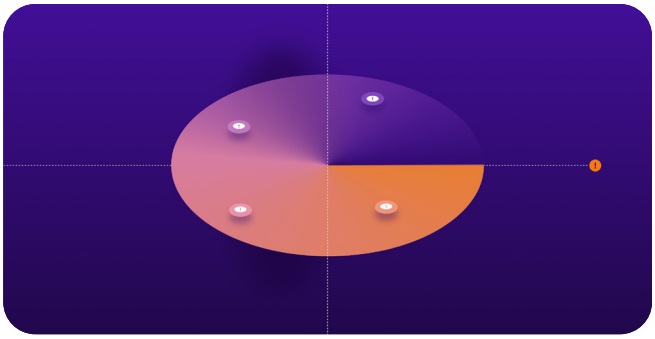
Concerns about AI and security are not exclusive to the use of the technology by attackers. AI systems can also be targets of attacks. The MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) framework provides an essential resource for defenders seeking to protect AI assets. ATLAS operates as a globally accessible knowledge base, meticulously cataloging real-world adversarial tactics and techniques targeting AI/ML systems. It offers a comprehensive taxonomy of attacks, ranging from data poisoning and adversarial examples to model extraction and backdoor insertion.

Adversaries can exploit vulnerabilities within AI models through multiple methods, for example, LLMs may be trained to exploit vulnerabilities including:



- **Data poisoning:** The foundation of any AI system is the data it's trained on. Poisoning attacks target this very foundation by injecting malicious data into the training dataset. This can be done subtly, inserting seemingly innocuous data points that skew the model's learning process. For instance, an attacker might introduce altered images into an AI system trained for facial recognition, causing it to misidentify specific individuals. Once deployed, the compromised AI system makes consistent errors that benefit the attacker, potentially allowing unauthorized access or bypassing security protocols.
- **Adversarial examples:** Adversarial examples are carefully crafted inputs designed to manipulate an AI model's decision-making. Imagine a self-driving car trained to recognize stop signs. An attacker might create a physical sticker with a specific pattern that, when placed on a real stop sign, appears normal to the human eye but confuses the AI model, causing the car to misinterpret the sign. These adversarial examples exploit the fact that AI models can be highly sensitive to slight variations in input data. Attackers can create these examples digitally as well, potentially compromising AI-powered security filters or spam detection systems.





- **Model inversion:** This technique attempts to steal sensitive information from a trained AI model. By analyzing the model's outputs, attackers can potentially reconstruct details about the training data used to create the model. Imagine an AI system trained to predict loan approvals based on customer data. Through model inversion, an attacker might be able to glean insights into the specific factors the AI considers when making these decisions, potentially allowing them to manipulate their applications to gain approval fraudulently.

- **Explainability gap (XAI):** Many AI models, particularly complex ones, can be opaque in their decision-making processes. This lack of explainability (XAI) creates a vulnerability as it's difficult to pinpoint the reasoning behind an AI system's output. Attackers can exploit this by feeding the model ambiguous or adversarial inputs and analyzing the resulting errors. Through trial and error, they might be able to uncover weaknesses in the model's logic and design targeted attacks that exploit those weaknesses. Mitigating this vulnerability requires ongoing monitoring of AI systems to identify anomalies and implementing XAI techniques to understand the reasoning behind the model's decisions.

Protecting AI systems

Defensive approaches to protect AI systems are rapidly evolving, as these systems are increasingly adopted by organizations for the most diverse use cases. Some of those approaches include:

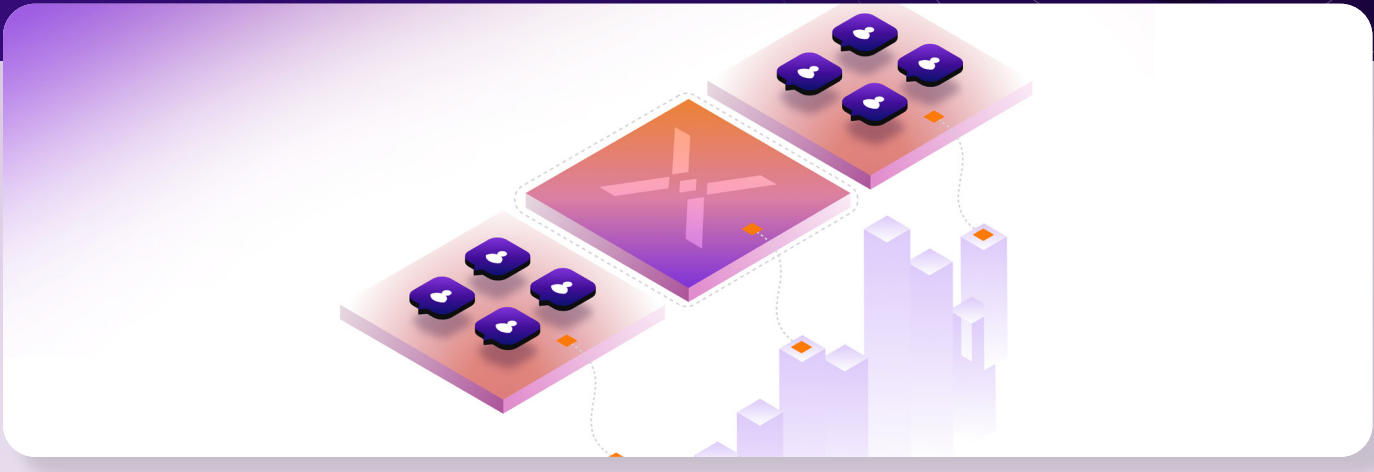
- **Adversarial training:** Exposing AI models to examples of potential attacks during training helps build resilience.
- **Continuous monitoring:** Logging and analyzing the decisions made by AI-based security systems is crucial for detecting malfunction or malicious manipulation.
- **Explainable AI (XAI):** XAI techniques provide insights into the reasoning behind AI decisions, aiding in the identification of vulnerabilities and biases.
- **Data integrity and provenance:** Ensuring the quality of data used for training and decision-making is paramount in protecting AI-based security.
- **Algorithm auditing:** Periodically testing AI models is vital in identifying unintended biases or errors that attackers could exploit.

Harnessing ATLAS for robust security controls

By studying the MITRE ATLAS framework, cybersecurity teams gain a deeper understanding of the attack vectors they must defend against. ATLAS empowers organizations to implement targeted security controls aligned with the specific attack techniques outlined in the framework. For example, defenders can prioritize data integrity validation and sanitization to counter data poisoning attempts, implement adversarial example detection for defense against model manipulation, and use secure development practices to prevent the injection of backdoors into AI models. Proactive engagement with MITRE ATLAS offers a roadmap for securing AI systems and mitigating the risks associated with this evolving technology.

Defending against AI-powered attacks

3



AI can be a potent ally in the fight against AI-powered attacks. As attackers optimize their capabilities with AI, producing a higher volume of faster, precise and harmful attacks, defenders have an opportunity to leverage these technologies to scale their own capabilities.

AI can be a potent ally in the fight against AI-powered attacks.

Security technology providers have been applying AI techniques to security problems for many years. Securonix User and Entity Behavior Analytics (UEBA) technology, for example, is a pioneer in the cyber field, applying machine learning algorithms and advanced mathematical models trained on large data sets since its inception in 2007. These algorithms, which fall under the large umbrella of artificial intelligence techniques, analyze user activity, device behavior, and network anomalies, baselining normal patterns and raising red flags when deviations occur. This provides efficient detection of insider threats, pinpoint compromised devices, and identify malicious actors long before they can inflict real damage.

The infusion of new forms of AI into security operations has the promise of transformative change. By automating routine tasks, AI can free up analysts to tackle more sophisticated challenges, effectively enhancing the overall productivity of security operations teams. For those analysts who possess less experience, AI can provide guidance, simplifying complex tasks that would typically demand a more seasoned hand. Simply put, AI is a force multiplier for the SOC.

Simply put, AI is a force multiplier for the SOC.

AI technologies can power rapid response capabilities to address emerging threats, increasing the speed of detection and mitigation beyond what is possible with humans only. The automation of certain security tasks by AI could also lead to operational cost reductions, presenting an economic benefit to organizations by potentially reducing both the need for a large staff of highly specialized security professionals and the marginal cost of SOC expansion.

Some innovative uses of emerging AI technologies by security teams include:

- **SecOps AI copilots:** LLMs augment human analysts with real-time threat identification, correlation, and response suggestions, for example, AI summarizes massive volumes of log data to flag critical anomalies rapidly.
- **RansomChatGPTs:** Simulation tools model ransomware negotiation behavior, training teams to effectively counter demands while gathering critical threat intelligence.
- **Malware analysis GPTs:** Specialized LLMs dissect malware code, predicting intent, uncovering hidden functionalities, and generating tailored detection signatures.

AI-Reinforced CyberOps

Securonix defines this AI enhanced future of security operations teams as "AI-Reinforced CyberOps". The old SOC (security operations center) has evolved beyond those large mission control style rooms, with analysts operating from multiple locations, many times remotely from home, and supported by a multitude of AI-based tools. The addition of AI across all the phases of the Threat Detection, Investigation and Response (TDIR) workflow has the promise of reinforcing existing practices, making detection more precise while accelerating response. The result is a large gain in efficiency, without increasing the burnout of analysts caused by the toil of overly manual processes.

securonix**eon**™

Securonix is a leader in the SIEM space, dedicated to staying ahead of the escalating threat curve. The next step in the evolution of the Securonix platform is based on the AI-reinforced CyberOps approach built on three key principles:

AI-REINFORCED PLATFORM

Leverages the power of AI to make precise security decisions at lightning speed. Securonix is investing in AI capabilities, including AWS Bedrock, across all layers of its platform, to ensure the need for human intervention is focused where it is necessary and most valuable.

CYBERSECURITY MESH

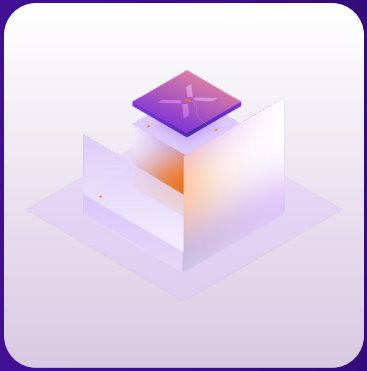
Seamlessly integrates with existing security tools, clouds, data lakes, and other technologies, to create a unified and flexible defense architecture. Its agnostic nature enables organizations to maximize the value of all their security investments.

FRICTIONLESS EXPERIENCE

Delivers reduced noise, an intuitive user experience and targeted threat intelligence to empower security teams to counter AI-powered threats. The analyst experience is tailored to the customer's needs and optimized to reduce context switching and training requirements.

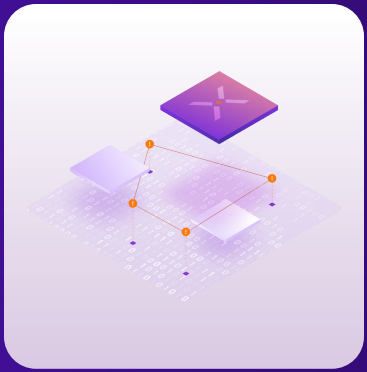
The objective of these principles is to set CyberOps for success against the new AI-powered attacks, while defending a rapidly changing environment, increasing efficiency to keep costs and analyst burnout under control.

Securonix Eon is the next step in the evolution of Securonix Unified Defense SIEM, leveraging the scale of AWS Bedrock to add advanced new features to the powerful and robust Snowflake backend and unified user experience:



Insider Threat Psycholinguistics

A large part of detecting insider threats requires understanding the intent and psychology of the individual or groups in question. Currently this is done through pattern recognition, which is subject to high false negative and false positive rates. This new capability leverages AWS Bedrock and the use of LLMs to identify potential users engaged with malicious intent. The method relies on semantic similarity and is robust against keyword search challenges such as misspellings, special characters, and poorly formatted regular expressions.



Adaptive Threat Modelling

Adaptive Threat Models can discover new combinations of alerts, absent of a pre-specified set of rules and activity/violation data residency using anomaly detection. The approach reduces the burden on security teams to develop threat detection content for multiple, unexpected attack scenarios, using the power of AI to reduce friction and improve analyst experience.

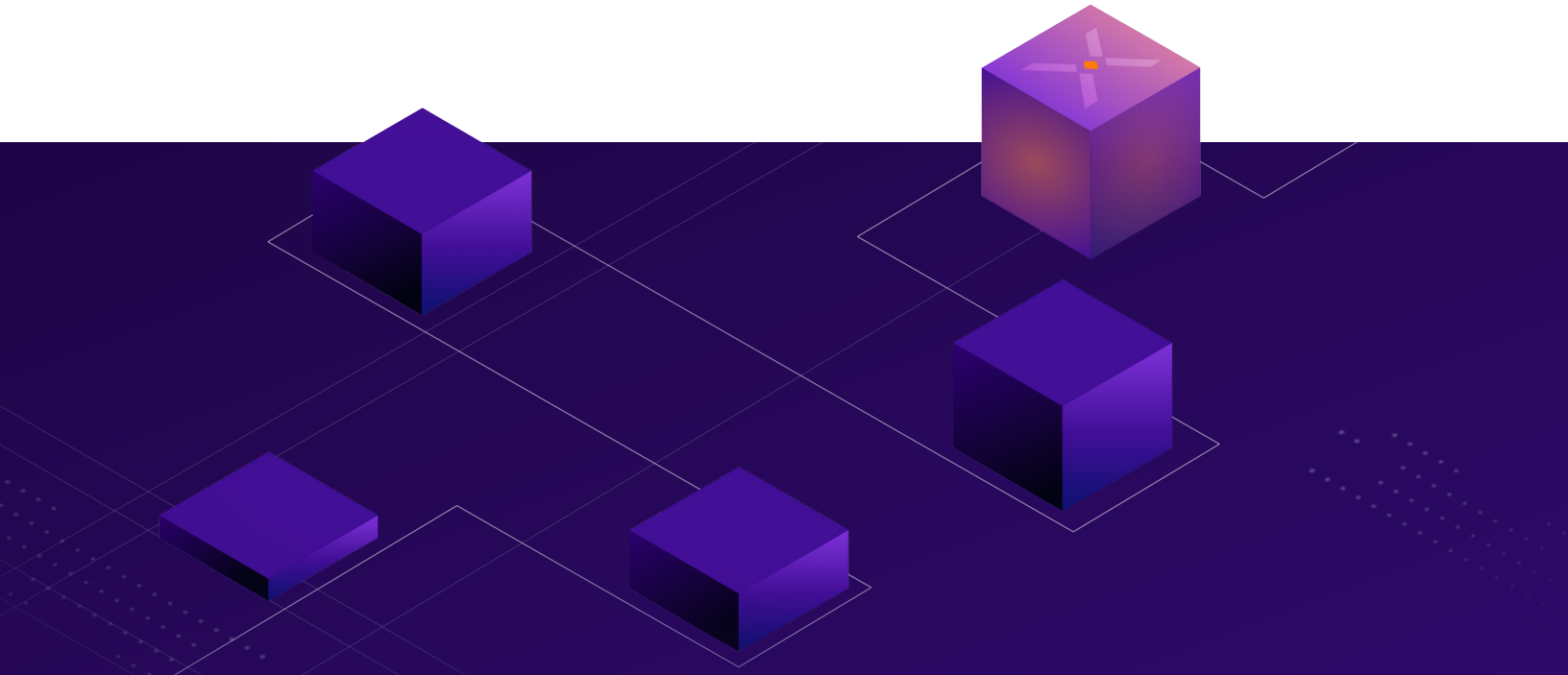
InvestigateRX

With Securonix InvestigateRX, analysts no longer need to search for data in various sources during an investigation. The data is now directly delivered to them. To reduce the burden of understanding the context and details behind each piece of information, Securonix InvestigateRX converts the retrieved personalized and objective content into a more coherent and context-aware summary, enabling the analyst to make swift decisions and saving an average of 15 minutes per incident.

The Value of AWS Bedrock

AWS Bedrock provides the building blocks for these AI-driven defenses. Tools like Amazon SageMaker streamline ML model development and deployment, while services like Amazon GuardDuty offer AI-powered anomaly detection at scale. Securonix Eon will leverage the power of AWS Bedrock to deliver its powerful new LLM based Insider Threat Psycholinguistics feature.

AI enables security teams to shift from reactive to proactive approaches.



Conclusion

AI IS A DOUBLE-EDGED SWORD IN AN EVER-EVOLVING CYBERSECURITY LANDSCAPE.

It presents adversaries with sophisticated attack tools while offering defenders unprecedented opportunities to bolster their defenses. Attacks are becoming faster, personalized and more devastating, prompting defenders to continuously evolve.

AI also enables security teams to shift from reactive to proactive approaches, predicting and mitigating attacks before damage occurs. AI-reinforced solutions constructed with strong building blocks such as AWS Bedrock can enable CyberOps teams to efficiently scale to keep up with evolving threats.

However, it's vital not only to leverage AI for security but to protect the AI systems themselves from exploitation. As the adoption of AI technologies increases, continuing research and development projects such as MITRE ATLAS is crucial to ensure that AI systems are built on a solid foundation, minimizing their attack surface while optimizing their value as a strong ally in the continuous cybersecurity battle against threats.