the-tech-trend.com

Understanding AI in Cybersecurity and AI Security: AI in Cybersecurity (UCSAISec-01)

Arash Habibi Lashkari

15–18 minutes

Artificial intelligence (AI) is used in more tools to protect our data. With its ability to crunch vast volumes of data, recognize patterns, and adapt to new threats, AI helps cybersecurity shift from reactive to proactive. It's not just about blocking known threats anymore – AI helps us spot suspicious behavior before something goes wrong. This article examines how AI is changing the game in cybersecurity across different areas and why AI has become a new target for attackers.

Threat Predictions

Over the past few years, we have witnessed significant shifts in the cyber threat landscape. Cyber-attacks have become much more sophisticated, targeted, and dangerous. As we enter 2025, it is evident that cyber resilience, zero-trust frameworks, and proactive threat intelligence are no longer optional but essential for safeguarding digital assets.

Several trends are emerging:

- **Misuse of Generative AI**: Large Language Models (LLMs) like GPT-4 make it easier to automate attacks, write phishing emails, build malware, and even create deepfakes. These tools lower the bar for less-skilled attackers to launch sophisticated campaigns.
- **Exploiting File Transfer Tools**: Managed File Transfer (MFT) platforms, like those in the MOVEit and GoAnywhere breaches, are often targeted because they handle sensitive business data.
- **Insider Threats**: Employees, contractors, or partners with legitimate access to an organization's critical assets can intentionally or unintentionally allow attackers to gain access. Up 47% in the past two years, these threats are costly to contain.
- **QR Code Scams**: QR codes are everywhere, and people trust them, but cybercriminals place fake QR codes that distribute malware or lead victims to fake websites.
- **Edge Devices**: Routers, firewalls, and switches are targeted by especially dangerous Advanced Persistent Threats (APTs) due to their inherent vulnerabilities and lack of intrusion detection capabilities.
- **Excel with Python**: A new replacement for malicious macros.
- **Signed Drivers**: Legitimate drivers can be exploited to gain kernel-level access. Despite some mitigation efforts, such as Microsoft's Vulnerable Driver Blocklist, these attacks remain simple to execute.

The Cyber Attacker's Motivation

Cyber attackers have all sorts of reasons for what they do. Understanding these motivations is critical for devising effective

defense mechanisms to protect an organization's assets. Some common motivations for cyberattacks include:

- Financial Gain
- Espionage
- Hacktivism
- Revenge
- Intellectual Property Theft
- Challenge/Thrill
- Cyber Warfare

The Traditional Approach to Cybersecurity

Before AI, cybersecurity relied on rule-based systems and signature-based detection to spot known threats like phishing, network breaches, and malware. These methods struggled with zero-day vulnerabilities and complex attacks like Advanced Persistent Threats (APTs).

With massive data and security logs, human analysts cannot hope to audit logs manually to find suspicious activities. Automated cyber threat detection systems were needed to proactively and dynamically manage new malware and zero-day attacks.

Also read: Future of AI Cyber Defense: How to Identify AI Cyber Attacks

6 Ways AI-Centric Cybersecurity Improves on Traditional Methods

AI has revolutionized cybersecurity in ways traditional methods

can't match. Here's how:

1. **Adapts on the Fly**: Unlike rigid rule-based or signature-based systems, AI updates itself as new threat data comes in, to stay ahead of evolving threats.
2. **Automates Tasks**: AI handles threat analysis, detection, and response automatically, reducing manual work and speeding up reaction times.
3. **Scales Easily**: AI manages massive data loads from modern networks, where traditional methods struggle and slow down due to manual rule updates.
4. **Spots New Threats**: Using machine learning and deep learning, AI catches anomalies and zero-day attacks that signature-based systems miss.
5. **Tracks Behavior:** AI is excellent at analyzing and baselining user and system behavior to spot insider threats or sneaky attacks by noticing abnormalities.
6. **Stays Ahead**: AI predicts threats based on past trends, enabling proactive defenses – unlike traditional approaches, which only respond after an attack.

Selecting AI Models for Cybersecurity

AI, including machine learning (ML) and deep learning, is now widely used to tackle cybersecurity challenges like spotting malware, detecting network breaches, identifying phishing attempts, and finding software weaknesses. These tools can be applied across areas like network security, app security, cloud security, and IoT/OT security, using everything from basic models

like linear regression to advanced ones like BERT or GPT-4.

Choosing the right model depends on factors like data type, performance needs, and threat complexity, as explored below:

Threat Detection with ML Models

The choice of ML model hinges on the data–structured files, binary images, code, logs, or network packets. Tabular data with numbers or categories suits models like decision trees, random forests, or SVMs. Images or text align better with deep learning options like CNNs, RNNs, or Auto Encoders. Complex or large datasets may require transformer models like GPT or BERT.

Speed, Scale, and Updates

Time and scalability influence model selection. Fast, real-time demands favor lightweight models with fewer parameters, especially for large, varied datasets. High-powered computing – on-site or in the cloud – boosts speed. Adaptability to new threats through updates is also key.

Handling Labeled vs. Unlabeled Data

Security data is often difficult or costly to label accurately, but the amount of labeling impacts model choice. Labeled data best fits supervised learning. A mix of some labeled but more unlabeled data leans toward semi-supervised models. The absence of labeled data, common in real-world scenarios, calls for unsupervised learning.

Model Clarity

Clarity in a model matters. Simple models like linear regression or decision trees offer built-in understanding, making them valuable when explainability is essential.

How to Secure AI Models

As AI becomes a more significant part of daily life, securing these models against attacks is critical. Adversaries target both data and models with threats that can weaken performance, security, and trust in ML systems:

- **Data poisoning** involves slipping insufficient data into training sets to skew results or bias predictions.
- **Membership inference** exploits model outputs to guess if specific data was used in training, risking privacy.
- **Model inversion** reverse-engineers a model to pull out sensitive information.
- **Model evasion** tricks the model with subtle input changes that look normal but cause errors.
- **Model theft** is stealing or copying proprietary models, threatening intellectual property.

Tackling these risks requires strong validation, constant monitoring, privacy-focused methods, and clear, accountable development. Adding software engineering habits like testing, patching, and quality checks can boost model reliability and accuracy.

AI Used for the Offensive and Defensive

AI is a dual-edged sword in cybersecurity, with defenders and

adversaries leveraging its capabilities to achieve their objectives. On one side, cybercriminals and malicious actors utilize AI to automate attacks, evade detection, generate sophisticated malware, and exploit vulnerabilities at an unprecedented scale. On the other side, security researchers and professionals harness AI to develop advanced intrusion detection systems, automated threat intelligence, anomaly detection, and proactive defense mechanisms.

We now move to explore both aspects in detail.

AI-driven Cyberattacks

AI-driven cyberattacks use artificial intelligence (AI) and machine learning (ML) to make attacks faster, smarter, and harder to stop. These attacks can find weak spots in systems, launch targeted campaigns, sneak in backdoors, steal or mess with data, and disrupt operations. AI's ability to adapt lets these attacks evolve, dodging detection and outsmarting defenses. With large language models (LLMs) now able to interact with tools, analyze documents, and even run themselves recursively, cybersecurity faces new risks and possibilities.

Malicious AI can also mess up other AI systems, giving attackers an edge in both digital and physical spaces. AI powers new attack types like data misclassification, fake data creation, and advanced analysis, but it can also fuel plenty of other threats.

The five main categories of AI-driven cyberattacks include AI-powered attacks, AI-powered social engineering attacks, AI-powered malware, autonomous malware, and AI-powered vulnerability exploitation.

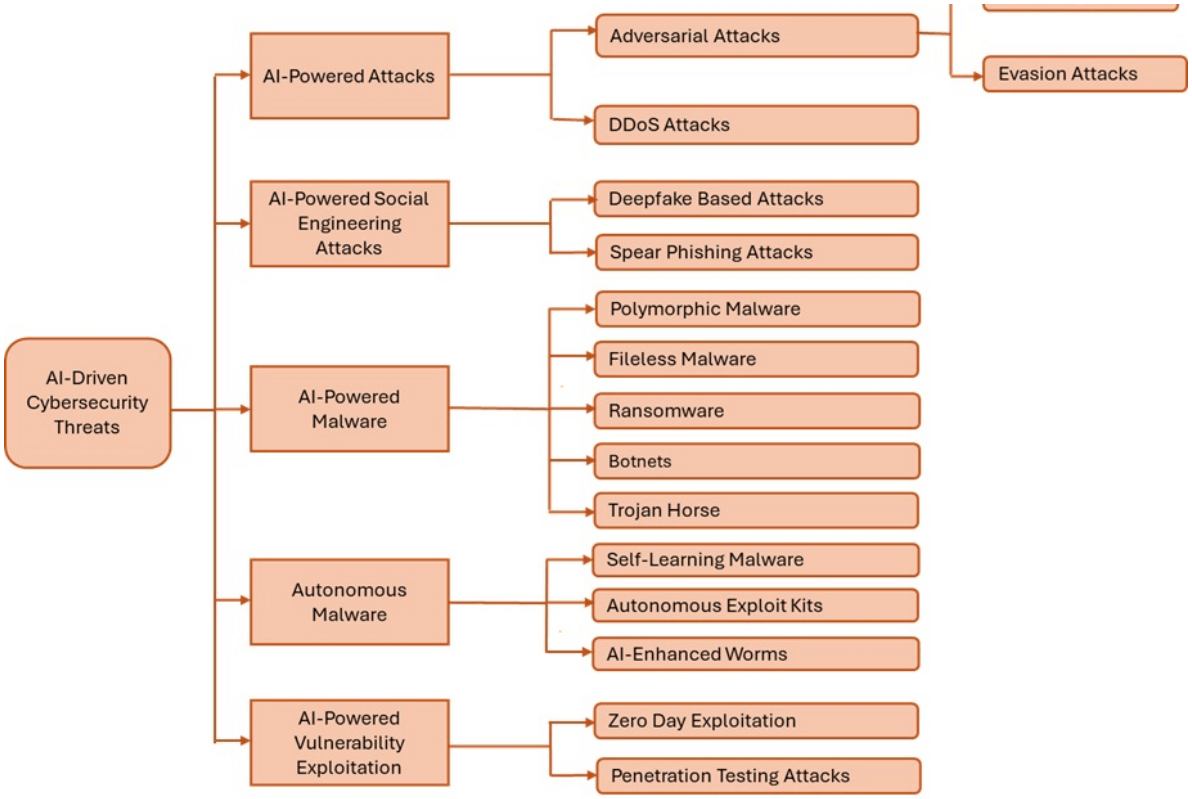

Figure 1: AI-Driven Cybersecurity Threats Taxonomy

AI-Powered Attacks

- **Adversarial Attacks**: Here, attackers tweak inputs to trick AI systems into bad decisions. They use AI to craft sneaky examples or determine a model's weak points. Two common types are poisoning and evasion attacks:
- **Poisoning Attacks**: These attacks occur during training, corrupting a model by slipping insufficient data into its dataset. Imagine an airport facial recognition system that an attacker feeds subtly altered images (tiny pixel changes). The system learns wrong patterns over time, compromising the security of the model.
- **Evasion Attacks**: These attacks occur during inference, altering inputs to confuse the model. For example, an attacker tweaks spam emails by changing words, shifting sentences, or hiding extraneous information in images. The spam filter misses the

email because the patterns don't match what it knows.

- **AI-Powered DDoS Attacks**: These use ML to study normal traffic, blend in malicious requests, and dodge detection. They can also coordinate huge botnets for massive strikes.

AI-Powered Social Engineering Attacks

These attacks use psychological tactics to manipulate individuals into revealing sensitive information, such as credentials or personal details. AI improves these attacks by crafting highly personalized and convincing messages.

- **Deepfake Attacks**: AI fakes audio, video, or images to impersonate someone. Say a crook grabs a CEO's clips from YouTube and uses deepfake tools to mimic their voice and face. On a video call, they "order" the CFO to wire $5 million for a "secret deal." It's so convincing that the CFO sends it without double-checking.
- **Spear Phishing**: AI combs through emails, LinkedIn posts, and tweets to mimic someone's writing style. A malicious actor targets a financial firm's HR with a fake executive email, perfectly matching the usual tone and phrasing to ask for payroll updates through a shady link. When HR clicks the link and enters credentials on a fake page, they unknowingly hand over access.

AI-Powered Malware

AI-powered malware can autonomously modify its code, making it more challenging for security systems to identify and neutralize threats. Attackers commonly use AI methods to make the malicious activity of polymorphic malware, fileless malware,

ransomware, botnets, and Trojan horses more effective.

Autonomous Malware

Autonomous malware is a step up from AI-powered malware. It incorporates self-learning to adapt to its environment, learn from its interactions, and make decisions on its own to evade detection or increase its impact.

- **AdaptiveCrypt** is a new self-learning ransomware that spreads via phishing and downloads. It uses AI to analyze systems, delay encryption for high-value targets, and evade detection by altering behavior.
- **BlackMamba** is an autonomous exploit kit that identifies vulnerabilities and generates polymorphic malware that can bypass defenses.

AI-Powered Vulnerability Exploitation

In automatic vulnerability discovery, AI or ML models can be trained to recognize patterns in code that suggest vulnerabilities such as buffer overflows, race conditions, or logic errors, and discover new attack surfaces. By automating the vulnerability discovery process, AI enables attackers to find zero-day flaws faster and more efficiently.

For example, DeepExploit is an AI-powered automated penetration testing tool that scans, learns, and attacks unpatched, zero-day flaws with no human input. It continuously learns and adapts its attack strategies, making it more effective at breaching systems and discovering previously unknown security flaws.

Also read: What is Threat Modeling: Practices, Tools and

Methodologies

Mapping AI Threats in Modern IDS Pipeline

A standard Intrusion Detection System (IDS) pipeline is generally composed of six key components: Data Collection, Preprocessing, Detection Engine, Decision and classification, Response and mitigation, and Model Training and update. These stages represent the foundational workflow used to capture, process, and respond to network threats in both traditional and AI-enhanced systems.

To better understand how modern, AI-driven threats impact IDS functionality, the threat categories identified in the proposed taxonomy are mapped to the relevant components of this standard pipeline. This mapping, illustrated in the accompanying diagram and table, provides a clear view of which stages are most susceptible to specific attack types such as poisoning, evasion, or model inversion, highlighting the need for targeted defenses at each layer.

Here are the components of this standard pipeline in detail:

1. **Data Collection Layer**

- Captures raw network traffic, logs, or sensor data from endpoints, servers, or IoT devices.

2. **Preprocessing Layer**

- Normalizes, filters, and extracts features from the collected data (e.g., packet size, timing, headers).

3. **Detection Engine**

- Employs machine learning/deep learning algorithms (e.g., CNN,

LSTM, transformers) to detect intrusions.

- Often split into:
- Anomaly Detection
- Signature/Rule-based Matching
- Behavioral Profiling

4. **Decision & Classification Layer**
- Classifies the input as benign or malicious and assigns a risk score or category.

5. **Response & Mitigation Module**
- Takes automated action (block IP, terminate process) or alerts administrators.

6. **Model Training & Update**
- Continual training with new threat data, feedback loops, and retraining pipelines.

Now we can map the AI-driven threats to the components in this pipeline:

Threat Category	**Mapped IDS Component Affected**	**Example**
Poisoning Attacks	Model Training & Update	Malicious data alters model behavior during training
Evasion Attacks	Detection Engine (at Inference Time)	Modified input bypasses classifier
AI-Powered	Data Collection /	Malicious traffic

DdoS	Preprocessing (Overload & Noise Injection)	floods the system
Deepfake/Spear Phishing	Detection Engine / Feature Extraction	Bypasses behavioral models or mimics trusted users
Autonomous Malware	Decision & Classification / Response Layer	Learns and adapts in real-time
Model Inversion/ Theft	Model Update & Feedback / Deployment Environment	Steals internal model logic or training data
Membership Inference Attacks	Model Training (Privacy Leakage)	Infers training data from model responses

Looking Forward

As AI reshapes cybersecurity, its dual nature demands our attention. It empowers defenders to outpace evolving threats by analyzing patterns, detecting threats early, and automating response. But the same capabilities fuel more intelligent attacks, from deepfakes to self-learning malware. As AI evolves, cybersecurity teams must treat it as a tool and a threat. Success won’t hinge on using AI alone but on building secure, explainable, and resilient systems that can keep pace with adversaries equally empowered by the same technology.