White paper

Cybersecurity in an AI-turbocharged era

Ondrej Kubovič and René Holt

with contributions from

Juraj Jánošík

Filip Mazán

Peter Košinár

Jakub Debski

Peter Stančík

Digital Security
Progress. Protected.

Publication date:
April 2024

Table of Contents

Digital Security
Progress. Protected.

Introduction

Artificial intelligence (AI) is probably the most hyped topic today, making and breaking news cycles, dominating sales and marketing materials, and powering a seemingly endless array of online services. While in 2019 an online search for the term "AI" returned over 2 billion results, in 2024 the same search leads to almost 18 billion results, illustrating how rapidly public interest has grown.

Most of the hype today can be attributed to large language models (LLMs), which have brought the communication skills and visual creativity of AI systems ominously close to human capabilities. Where many see a business opportunity and a bright future full of AI solutions, others fear for the future of millions of jobs and even humankind itself.

In this paper, we will stay away from those grandiose visions and doomsday scenarios. Instead, we will focus on **the real contributions and risks of this technology for cybersecurity**.

We'll show how **ESET integrated a precisely selected set of AI algorithms into our detection engine**, turning it into a highly effective protective machine with high detection rates and minimal false positives. We'll also showcase how **AI aids our threat intelligence and threat hunting capabilities**, pointing our experts to interesting attack scenarios and malware features.

Aware of the potential AI has in the hands of cybercriminals and state-aligned threat actors, we will also cover **the threats that can be built with this technology**. Our focus will be on AI's generative capability, which can be used to produce new malware, improve the quality of victim targeting and social engineering campaigns, boost the language quality and quantity of malspam campaigns, and transform the threat landscape in many other ways.

Against such threats, **AI plays a pivotal role in fortifying organizations' threat detection and prevention capabilities**. AI can analyze vast datasets in real time, quickly identifying patterns and anomalies that indicate new threats and security holes, and thus facilitate swift remediation. Most importantly for defenders, **AI strengthens a prevention-first approach to cybersecurity** while also bolstering a proactive and adaptive defense strategy.

Using AI for good

With its specific terminology, elaborate threat scenarios, the enduring security skills shortage, and the ever-increasing pool of cybercriminal and nation-state-aligned adversaries, cybersecurity is a realm ripe for AI-powered innovations.

While for many years AI has improved some areas of cybersecurity – threat detection, filtering and analysis of threat intelligence data, and defenders' tools – others could still benefit. This has become even more apparent with the rise of generative AI and large language models (LLMs) capable of better processing natural language and thus "understanding" questions posed by users with varying knowledge levels and answering their prompts in an appropriate manner. This recently acquired capability can benefit, in addition to defenders and developers of protective solutions, computer users in non-security contexts.

Current defensive uses of AI:

- Processing vast amounts of data to identify attacks by correlation of various indicators.
- Identification and analysis of suspicious or malicious code in, and behavior by, programs.
- Monitoring and analysis of network traffic for malicious or anomalous patterns.
- Explanation and transcription of complex threat information, making it more accessible.
- Prioritization of alerts, helping defenders to focus on the most pressing security issues.
- Complementary functions for other defensive layers.

Defensive uses of AI currently under development or in the future:

- Creation of new detections based on descriptions of threats.
- Analysis, contextualization, and explanation of past or current events in an environment.
- Scanning an organization's environment for hidden or unknown vulnerabilities.

What defenders need to know about AI:

- High false-positive rates limit the usability of AI-only security.
- Without updates and human oversight, AI-powered security can, and probably will, deteriorate.
- High-quality training sets translate into high-quality AI models.

Here are some examples:

EXPLAINING THREAT INTELLIGENCE

The ability of LLMs to distill simple ideas from long and complicated materials could become very useful for regular users, IT administrators, and decision-makers. These groups often struggle to understand the highly technical reports produced by security professionals and their tools. With the help of generative AI models, threat intelligence information could consist of short texts written in easy-to-understand language that highlights key points and actionable takeaways.

IN-PRODUCT AI ASSISTANTS

Organizations in all sectors have been plagued for years, leaving them exposed to threat actors, by misconfiguration of systems and security products. A built-in AI assistant that helps organizations optimize the setup of their environments and properly configure any installed protective solutions can improve the user experience, enhance detection results, and boost security. This assistant could also execute commands that require multiple steps to be taken, such as adding an exemption to the firewall and changing security solution settings, or guide the IT admin through resolving a security or system notification.

RAZOR-SHARP, AI-POWERED FOCUS

Cybersecurity professionals could benefit immensely from using dedicated AI models. Today, their teams face an expanding attack surface, a growing array of security tools, and an influx of threat intelligence data. Managing all those tasks is laborious, time-consuming, and requires broad expertise. AI can ease that burden – and thus reduce alert fatigue and improve effectivity and focus – by prioritizing detected events, putting them into context, highlighting the most pressing issues, and even resolving minor alerts. Furthermore, AI could use previously gathered events to generate a dynamic map of interconnected items and events, providing a valuable overview for incident response teams and post-incident investigation.

CREATING NEW DETECTIONS

New threat actors and attack techniques are uncovered daily by security researchers. By combining public information and indicators of compromise with the threat data feeds and inputs of the organization's security team, an AI model could generate corresponding detections to scan environments for these latest threats. However, to avoid false positives and better triage incidents all such detections would still have to be verified by experts for their accuracy and effectiveness.

BOOSTING USER AWARENESS

Another area that can benefit from AI is security awareness. Content-generating models could be employed to convert media articles about the latest threats into internal memos

or transform them into easily digestible infographics. They can also write new phishing emails for internal testing and the follow-up explanation sent to employees explaining the threat and its risks. To gamify awareness, generative models can draft quizzes and other educational materials that consider the internal culture, business language, products, processes, and other specifics of the given organization.

SANDBOXES AUGMENTED WITH AI

Generative AI could also be useful in sandbox testing of threats, to analyze text in screenshots or low-level events such as Windows registry changes, and to write easily digestible descriptions. Such analysis could explain how the threat behaves and its observed capabilities.

ADVANCED AI ANTISPAM AND ANTI-PHISHING

Even antispam and anti-phishing technologies can benefit from the latest advancements in AI. Defensive models could be pretrained on the previous email communication of a given user to learn to identify patterns that are out of the ordinary. AI spotting a sudden change in tone or message content could help prevent reply-chain attacks, where attackers make their malicious emails look trustworthy by replying to a preexisting email communication of their victim.

While this looks good on paper, the training cost for such solutions might be prohibitive – at least in the early stages. However, the benefit-cost ratio could be higher if this kind of protection were deployed to the inboxes of select individuals who work with sensitive data or in organizations or industries that might be potential targets for spearphishing and whaling attacks.

ESET has been using AI for over 25 years

Most of the defensive uses of AI mentioned in the previous chapter are, or soon will be, deployed in the security products of many vendors to protect organizations and individuals from cyberthreats. Here's a summary of ESET's experience with AI over the years and its current integration into our multilayered technology.

1997: THE FIRST EXPERIMENT – DETECTION OF MACRO VIRUSES

ESET has rich experience with the deployment of AI- and machine learning-based systems. Our engineers have been experimenting with this technology since the early years of the company's existence and first deployed it in our products in 1997 to improve the detection of macro viruses. That was only the beginning, focused on testing whether building defense layers upon AI and machine learning even made sense.

2005: DNA DETECTIONS

After using it for macro viruses, ESET announced another AI-based technology, naming it DNA Detections. This approach converts the analyzed sample into a form amenable to matching and detection by precisely selecting features – "genes" – and building a DNA profile. These profiles are then used to build a complex model that splits the space of all analyzed material into malicious and clean samples, and then distinctive "gene sequences" – our DNA Detections – that reliably differentiate the malicious from the clean samples are derived. Created either by automated systems or expert human researchers, this regularly updated model has served as our "online machine learning model" since its inception in 2005.

2006: BACKEND EXPERT SYSTEM FOR SAMPLE PROCESSING

Inspired by the effectiveness of our previous AI-powered systems against known and emerging threats, a series of internal ESET projects using this technology followed, leading to the introduction of backend expert systems for mass processing of hundreds of thousands of samples every day. Even today, these systems are the backbone of ESET's technology stack, helping detection engineers with the triage, sorting, and labeling of most incoming material.

2010: ESET LIVEGRID®

Since these early AI-related experiments showed great promise, ESET went further down this path and, in 2010, introduced its cloud reputation system, ESET LiveGrid®. At that time, it already delivered speedy updates in a matter of minutes instead of hours, leveraging the power of online learning, where a model continues learning as training examples arrive in real time.

Timeline of AI technology development highlights at ESET

1997

First use of neural networks in ESET products, utilized for detection of macro viruses.

2005

ESET DNA Detections, a synonym for online machine learning, uses genes of malware to detect current and emerging threats.

2010

ESET LiveGrid®, a cloud-based reputation system, leverages ESET DNA Detections to significantly speed up user-side updates.

2017

ESET Advanced Machine Learning in the cloud uses AI to power our automated detection systems.

2018

ESET LiveGuard, an AI-powered cloud sandbox, provides on-demand analysis for ESET customers with a turnover time of minutes.

2019

ESET Advanced Machine Learning in the endpoint uses AI to power our automated detection systems.

2020-2021

Transformer-based models deployed in ESET's cloud and endpoint solutions.

2023

Automated Incident Creator added to ESET Inspect to reduce noise and leverage techniques, including AI, to produce a clear overview of an incident.

2024

ESET AI Advisor, Release of a generative AI security advisor for ESET Threat Intelligence that can generate detailed descriptions of incidents, and for ESET Inspect (in preview), provides answers to queries regarding the scope of incidents.

2017–2019: ESET LIVESENSE

Around the same time, the world was taken by storm with the breakout of deep learning algorithms – in supervised, unsupervised, and reinforced forms. Many emerging vendors tried to use the hype and boast about their implementation as the silver bullet capable of "solving cybersecurity". But as soon as these solutions started flooding security teams with false positives, it became clear that deep learning could not identify and stop every possible attack scenario, at least not without this undesirable consequence.

With our focus on technology and science, ESET put this new branch of AI through a series of rigorous tests. Experimenting with long short-term memory neural networks and their combination with decision trees and other algorithms yielded new layers in our protective engine – ESET Advanced Machine Learning in the cloud (2017) and ESET Advanced Machine Learning in the endpoint (2019). These have high detection and low false-positive rates, and were integrated into our core multilayered technology called ESET LiveSense, augmenting a wide array of our other in-house developed, defensive layers and expert oversight.

2018: ESET LIVEGUARD

Experience gathered in AI development paved the way for a new, extremely powerful cloud sandbox called ESET LiveGuard (formerly ESET Dynamic Threat Defense). Its four stages of analysis combine a multilayered machine learning detection system, superior unpacking and scanning, a proprietary experimental detection engine, and deep behavior analysis, all of which process and evaluate submitted items. This AI-powered cloud sandbox essentially offers subscribers the full force of ESET's detection technology – both in-house and in-product – resulting in an on-demand comprehensive analysis within minutes, or seconds.

Transformer-based generative AI

The idea of creating new content using AI-based models has been around for years and has been utilized in specific fields, such as computational chemistry. However, for security-related tasks, a different approach was needed.

In 2017, Google published a paper called Attention Is All You Need. It proposed a new architecture for machine learning models based on attention mechanisms. Named "transformer", this architecture was shown to be very effective for processing natural language and producing a wide variety of human-understandable content.

Fast-forward to 2022, models such as ChatGPT, Midjourney, and DALL-E grabbed the public's attention by showing that with simple user input – a text prompt – transformer-based models can write a complete article, generate a realistic photo, and produce new videos. Of course, this is only the tip of the utilization iceberg that has yet made its way into media headlines.

2020–2021: TRANSFORMER-BASED MODELS CONTRIBUTING TO DETECTION

For ESET engineers, the new class of transformer models meant yet another round of testing and pitting these models against malicious samples. Showing positive results in the detection field, this technology was introduced into our multilayered cloud detection systems and endpoint detection modules in 2020 and 2021, respectively.

2023: AUTOMATED INCIDENT CREATOR IN ESET INSPECT

AI is also built into our XDR-enabling offering ESET Inspect, powering its automated Incident Creator. This technology assists defenders by correlating indicators, extracted from a large number of events collected from endpoints, to triage events and automatically group related detections into a visual representation. In turn, this helps shorten the time needed for each incident and reduces the alert fatigue felt by otherwise shorthanded security teams.

2024: ESET AI ADVISOR

With AI already being used in ESET Inspect for its Incident Creator, the integration of ESET AI Advisor, a generative AI assistant, will further widen the resources available to its users. Security staff can ask ESET AI Advisor to interact directly with incident data, using natural language prompts to obtain additional threat intelligence, such as context on the detected artifacts and tactics, techniques, and procedures (TTPs) observed during an incident.

For junior staff, ESET AI Advisor can offer support by responding to security-related queries. For senior staff, ESET AI Advisor can generate easy-to-understand overviews useful for sharing with other teams and people with different levels of technical understanding. ESET AI Advisor can even produce step-by-step guides for specific groups of employees in the organization, enabling them to be part of the prevention and mitigation of security incidents.

AI can also be used with ESET Threat Intelligence, which offers an extraordinarily rich source of threat-related information. It tracks a broad range of threat actors and their TTPs, details common attack scenarios and their broader context, and provides indicators of compromise (IoCs). To avoid a needle in a haystack situation, ESET engineers have empowered ESET AI Advisor to crawl through this library of information to quickly provide security professionals with the sought-after data.

By sifting through IoCs, TTPs, time-, sector-, and location-specific data, and tying them to specific threat actors, ESET AI Advisor can provide comprehensive, yet digestible, summaries. ESET AI Advisor can also be used to create reports for designated target audiences such as IT staff, CISOs and security teams, and others in the C suite. To address the potential risk of hallucinations *(see Hallucinations in generative models)*, ESET AI Advisor always provides references to source documents.

Retrieval-augmented generation (RAG) is a method to improve and make the output of large language models (LLMs) more accurate. To achieve this, the foundational LLM is given access to a framework of tools that can reach external sources of information and knowledge otherwise unavailable to the model. The gathered material is then used to help formulate or enhance a response to a given prompt, offering a more current, nuanced, and reliable response.

For example, a ChatGPT model could gather the latest and most accurate information made available by an online search engine at the time of the prompt, instead of building its answer on information available only during its training in the past.

ESET AI Advisor also uses retrieval-augmented generation to access and leverage the power of ESET-developed internal tools and data, and thus can produce comprehensive threat and incident reports.

This integrated use of AI not only helps create robust proactive defenses but also provides an intuitive and easily manageable interface between security staff and their tools.

Even AI has limits

Limits of AI:

- False positives and low priority alerts can cause alert fatigue and lead to misconfiguration of security products.
- Without updates and expert supervision, models can degrade.
- Long-term reliability is essential for security solutions but not guaranteed for AI models.
- Generative AI models are at risk of hallucinating.
- Adequate protection requires other security layers and tools in addition to AI.

Neural networks, deep learning, natural language processing, decision trees, transformer-based models, large language models, and basically any other AI technology can be leveraged to improve specific aspects of cybersecurity. However, our years in the field have made us appreciative of both the expertise needed to use AI and the limitations of this technology. Here are a few limitations that can have a significant impact on protection:

FALSE POSITIVES STILL MATTER

When defenders or their AI-powered tools mistakenly label a benign file or event as malicious – a false positive – it can have severe consequences, sometimes even worse than missing a malware sample – a false negative. For example, in manufacturing, the potential consequences include disruption of production, damage to the product or the line, delays, and financial loss.

False positives can also lead to alert fatigue of the security staff. With too many false alarms, defenders will tend either to spend an excessive number of workdays resolving the underlying issues, or to loosen the protective setup, thus reducing the detection capability. Both of these scenarios can potentially have a negative effect on an organization's security posture and introduce new avenues for threat actors to break in.

ML MODELS DEGENERATE, GENAI NEEDS HELP TO KEEP UP

In the 2010s, when machine learning became a standard feature of most security products, many emerging vendors claimed that their models could address current and future threats without any updates. However, real-world deployments of these solutions showed that this approach led to an extraordinarily high false-positive rate, and the performance of such products degraded over time. Based on our experience in this field, continuous supervision and tweaking of machine learning models and their training set is key to maintaining their net positive contribution to other protective technologies.

With large language models, the situation is a bit different. The basis – language – doesn't evolve as quickly and thus the foundational models don't need to be retrained as frequently as machine learning models used in detection layers. However, to be able to provide the most up-to-date and detailed responses to user requests, an LLM should use retrieval-augmented generation, which obtains the needed information online or from proprietary sources. If this information retrieval middleware is misconfigured or tampered with, the model can be fed incorrect data and thus produce biased or problematic results. On the other hand, not using retrieval-augmented generation deprives the model of context and would likely make it unable to provide certain answers or other requested details.

QUALITY AND LONG-TERM RELIABILITY

In cybersecurity, consistent performance and reliability are key. If an AI-powered security solution has great detection results and few false positives one week but fails to detect malware or creates a flood of false alarms the next week, then this only increases the burden on the security team. Expert human oversight by the designers of the core model(s) of the security solution is thus crucial to retain high detection and low false-positive rates over the long term. Even if some predeployment training of the model is needed – to understand the specifics of an organization – this is preferable to creating a flood of false positives or false negatives.

HALLUCINATIONS IN GENERATIVE MODELS

Don't believe everything you see online – this is a rule that applies especially to content created by generative AI. Many of today's AI models can compute the best possible word or pixel to follow a given input, ultimately aiming at producing humanlike and believable output. However, this can sometimes lead to seemingly plausible but incorrect information, with fabricated – hallucinated – references, sources, data, authors, statements, or URLs, making a strong argument for continuous oversight and verification by humans.

Hallucination challenges the deployment of generative AI in many fields, including cybersecurity – where sample analysis outputs based on self-fabricated data might mislabel samples. Similarly, a hallucination-based interpretation of threat intelligence can lead to bad or even dangerous advice, decisions, and policies, potentially compromising the security of whole environments.

NOTE: Hallucinations by generative AI can be beneficial for specific use cases. If the goal is to generate new audio, video, or visual content, the algorithm should have the "creative freedom" to produce fresh ideas or answer the given prompt in a way that no human would.

(GENERATIVE) AI ALONE ISN'T GOING TO CUT IT

Deploying generative AI – but also other models – for specific tasks can be quite cumbersome. The challenge is often the training set, which needs to be precisely chosen and labeled to achieve the desired results. There are many examples of models that, lacking proper labeling and guardrails, became strongly biased and generated distorted outputs. Similarly for cybersecurity, if the training set is not carefully selected, correctly labeled, and balanced, the model could become overly sensitive and produce a flood of false positives or become focused on unimportant aspects and thus blinded to clearly malicious attributes.

Making things more difficult, cybercriminals and state-backed adversaries constantly try to make their "product" invisible or look innocuous by adding layers of packing, obfuscation, encryption, etc. An AI model without proper additional tools, training, and human-expert oversight will not be able to handle those obstacles and will thus fail to peel back the obfuscation layers to get to the malicious core of the sample. This could negatively influence the output of AI analysis, resulting in low-value information for defenders.

Attackers can also modularize their malware so that each module appears clean on its own and only when all these parts work together do they begin to demonstrate malicious behavior. In such cases, pre-execution red flags are absent, and even a well-trained AI solution can be fooled into making the wrong assessment, marking these files as benign.

INTELLIGENT AND ADAPTIVE ADVERSARIES

Today's machines can defeat humans at chess and Go, and are quickly becoming more effective at other tasks, too; however, most of these tasks exist in environments with firm rules. Threat actors, on the other hand, do not follow guidelines or accept limitations and will cheat, manipulate, or change the playing field without warning.

A good example is self-driving cars. Despite massive investments in their development, these vehicles rely heavily on environmental markings such as traffic signs and lights. An adversary could attack driverless vehicles by covering up traffic signs or making traffic lights blink at a rate unrecognizable to the human eye. With these deformations, such cars could start making poor decisions and causing fatal crashes.

This ever-changing nature of the threat environment makes it impossible to create a universal protective solution that can counter all current and future threats, and no, not even the latest AI models change that.

AI-powered threats expected in 2018:

- Generating humanlike social engineering campaigns, including spearphishing.
- Optimizing malware and adjusting it to selected environments.
- Replicating and implementing false flags.
- Improving victim selection and targeting.
- Searching for new vulnerabilities in software and IoT firmware.
- Generating new malware or rewriting it in different programming languages.
- Triggering self-destructive mechanisms in the malware as a last resort to thwart investigation and analysis.
- Decreasing the time of an attack to shorten the reaction time for defenders.
- Collective learning of (IoT) botnets.

Additional AI-powered threats expected today and beyond:

- Generating a large quantity of high quality spam, scam, and phishing campaigns.
- Generating a large quantity of high quality mis- and disinformation, pictures, and deepfake videos for use in scams, extortion, and influence operations.
- Analyzing network traffic and inputs from compromised devices, and subsequent use of this information to hide and protect malicious infrastructure, code, operations, and threat actors.
- Extracting legally protected, proprietary, or otherwise sensitive information known to generative models via specially crafted, malicious prompts.
- Boosting social engineering campaigns by leveraging the humanlike communication generated by LLMs.

Using AI for evil

CURRENT THREATS POTENTIALLY LEVERAGING AI

Underestimating what AI technology can be used for in the hands of cybercriminals and sophisticated threat actors would leave organizations and their defenders dangerously unprepared. That's why the 2018 version of this paper listed more than a dozen expected attack scenarios, some of which have become an everyday reality.

MALICIOUS SPAM AND SCAMS

Generating new malicious content has held a prominent spot on that list. Back in 2018 it was primarily AI-powered online translation that fueled this type of activity; today, armed with

LLMs capable of imitating any given person's writing style, attackers can design advanced spam and scam campaigns that are becoming increasingly difficult to identify just from their message content.

DISINFORMATION CAMPAIGNS

The same is true for disinformation campaigns. These were once a laborious endeavor produced by large troll farms with dozens, if not hundreds, of human workers. With generative AI models, rewriting an online article and infusing it with falsehoods, fake photos, or deepfake videos is now simpler and easily repeatable, needing only a handful of trained human "creators". Moreover, information of this kind might gain significantly more traction on social media, where people often just scan the headlines and their accompanying pictures.

EVADING DETECTION

Some forms of AI and machine learning could be used to protect malicious infrastructure. Emotet was a prime example of a data-collecting and detection-thwarting botnet that kept security researchers at bay by analyzing every potential victim for signs of monitoring. Using a machine learning model would have made this otherwise gargantuan task much easier for these criminals.

EMAIL THREAD HIJACKING

Spearphishing is also set to see a significant bump in its return on investment, as feeding an LLM with a stolen victim's emails and information allows the attackers to draft a trustworthy-looking message. If such a message is injected into one of the victim's previous conversations – a technique known as a reply-chain attack – the chances of attackers achieving their goals grow dramatically.

WRITING NEW MALWARE

On the other hand, not all threats are as real as some headlines might portray them – case in point: writing malware from scratch. While it would be a worrisome capability, AI's coding skills are still limited. Current generative models might be useful for narrow assignments such as rewriting libraries into other popular languages, debugging, code optimization, and maybe even drafting a simple, tightly-specified function. However, AI's results are suboptimal when writing complex tools or software, including those intended for malicious purposes.

And even if AI wrote high-quality malware, it is only one of the steps on a long journey to making it an effective and profitable threat. Attackers need to design a distribution strategy, find ways to avoid detection by security tools and personnel, figure out how to monetize the access and the stolen information, and sometimes even communicate further with the victim. AI can come in handy for some of those steps, but it cannot fully replace an intelligent human adversary – at least not today.

SCI-FI OR NEAR FUTURE?

We need to stress that with the current pace of AI progress, we will probably see AI models become better at all the abovementioned activities in the coming years, or even months. This leads us to the sci-fi scenarios that haven't yet materialized but may become reality in the foreseeable future.

Planting false flags

Threat actors could train generative AI models on published research about the activity of other threat actors so as to run campaigns under false flags. This would make the already tricky business of attributing cyberattacks to specific groups even trickier.

Improved victim selection

AI could become the comb that would go through datasets gathered in the reconnaissance phase of an attack to pinpoint the most interesting targets – be it a sloppy or gullible employee with broad system privileges or a subcontractor with poorly protected or misconfigured systems.

Hunting for vulnerabilities

Recent years have proven that zero-day vulnerabilities are a lucrative business as much for groups focusing on intrusion and espionage as for cybercriminals seeking financial gain. Training AI-powered models to search for unknown, exploitable flaws could open (back)doors to almost any computer environment on the planet, especially if the notoriously insecure, and often unpatchable, Internet of Things (IoT) devices are present.

Learning botnets

Mentioning IoT: AI could help threat actors grow new botnets, making them more effective and capable of collective learning. This means we could see these large digital organisms engage in larger, more sophisticated operations – such as vulnerability hunting or information harvesting – instead of just being used for their sheer force in distributed denial-of-service (DDoS) attacks.

Conclusion

As shown in this paper, **AI represents an extremely beneficial technology for cybersecurity**. Integrated into protective solutions, AI can enhance threat detection and response capabilities, improve threat awareness and the accessibility of services such as threat intelligence and threat hunting, and contribute to better protection, thus preventing advanced attacks. By reducing noise and alert fatigue it also enables cybersecurity experts to identify and respond to malicious activity faster and more efficiently.

Adoption of AI still can and probably will have a transformative effect in other areas of cybersecurity, such as creating new detections, scanning for unknown vulnerabilities, and assisting with proper configuration of protective tools. Despite these benefits, AI in its many forms comes with its own set of limits and challenges. These include the requirement for high-quality training sets, the risk of high false-positive rates, and the need for updates and expert human oversight.

Not to be neglected are **the potential scenarios for misuse of AI by malicious actors**. Threat actors can, and some already are, leveraging this technology to generate convincing spam and scam campaigns, improve their social engineering, evade detection and monitoring, and even debug and optimize malware. While these threats are concerning, this paper emphasizes that **AI isn't capable of fully replacing an intelligent human adversary**, especially in carrying out complex accompanying tasks, such as imagining an effective attack chain or generating new and sophisticated malicious code.

We have written this paper to underscore the importance of understanding and leveraging AI in the realm of cybersecurity. At the same time, we acknowledge the limitations and potential risks of using this technology. In short, we argue for the need of a **balanced approach that combines AI with expert human oversight to ensure the development of effective and reliable cybersecurity solutions**.

Appendix A: Terminology

ARTIFICIAL GENERAL INTELLIGENCE (AGI)

This represents the as-yet unachieved ideal of a generally intelligent and self-sustainable artificial agent that can make decisions and learn independently based solely on inputs from real or virtual environments, without human involvement and oversight. AGI is focused on the development of agents capable of performing a wide range of tasks as compared to "narrow" AI that designs agents for a limited set of tasks.

ARTIFICIAL INTELLIGENCE (AI)

Artificial intelligence refers to computational agents, implemented in software and hardware, designed to act intelligently within specific environments that limit the actions available, the time to act, and the data that can be observed. The intelligence displayed includes learning, adapting to changes in the environment, considering consequences of decisions, and selecting suitable approaches based on current goals, knowledge, and restrictions.

MACHINE LEARNING (ML)

Machine learning mainly deals with the design and use of models that can analyze large data sets and learn functions that predict the output for new inputs. Models inspired by how neurons function in the human brain are called neural networks; these have been very effective for learning combinations of functions and are a powerful tool for prediction.

GENERATIVE AI (GENAI)

Advances in both natural language processing and transformer-based neural networks have led to the growth of generative AI. Typically trained on large sets of unlabeled data, these AI models leverage simple human-machine interfaces that accept natural language prompts for the generation of novel outputs in a short time. The produced content includes statistical data, text, images, audio, video, and source code.

Appendix B: Where AI reality stops, and myths start

When a topic reaches the level of hype that AI is currently experiencing, myths inevitably start to pop up. Cybersecurity is not immune to this trend and there are quite a few wild claims out there trying to capitalize on media cycles. For those interested in the real state of affairs in this area, here are some current AI claims debunked.

CLAIM: AI CAN ANALYZE CODE AND IDENTIFY ITS MALICIOUS BEHAVIOR

Reality: While not entirely wrong, the quality of code analysis and the final output from current models is questionable at best. Yes, a threat description drafted by generative AI might read well and have flawless grammar and style. However, it can often be incomplete, incorrect, or out of context and only experts with years of malware analysis under their belt would spot the issues. If less skilled recipients use such information as the basis of their decisions, this can lead to catastrophic consequences. Making matters worse, adversaries could – and probably will – actively try to poison their code or obfuscate it, to increase the chances that the model returns the wrong results or cannot produce any useful output.

CLAIM: AI CAN WRITE NEW SOFTWARE AND THUS MALWARE

Reality: Some online services use generative AI to create new code. This is useful and effective if applied to boring or less complex tasks that would otherwise take up the valuable time of skilled developers. However, testing shows that writing software from scratch is a much bigger fish to fry and appears to be too advanced for contemporary AI. This is also true for malware, where the complexities go even further, including distribution of the final "product", its protection from detection and analysis, and other steps needed for it to be effective and profitable. It is far easier for an attacker with even mediocre coding skills to use tutorials or work from leaked source code instead of using a generative model.

CLAIM: THE BIGGER THE MODEL, THE BETTER

Reality: One of the main characteristics of LLMs is that they are large. Some cybersecurity vendors like to spin this trait as one of the major advantages for malware analysis. The thing is, the larger the model, the more it costs to train it. Bigger models require expensive hardware, a broader set of inputs, and more training time, and they burn a lot of electricity and other resources, making them also less eco-friendly. A smaller-sized AI model with a narrow assignment is cheaper to train, more affordable to maintain, and easier to understand and keep in check. In cybersecurity, this type of model can be used to process large amounts of data and provide simple, easily readable outputs that label samples as benign or malicious.

CLAIM: AI IS THE ONLY NECESSARY SECURITY LAYER

Reality: As can happen with any other technology, AI has been oversold by some companies as the solution for everything. This includes cybersecurity where reliable detection technologies proven by years of use are dismissed by some in favor of AI. While neural networks, deep learning, and generative AI bring value to the table, there's no magical algorithm that will – alone – identify every possible threat that might emerge. Combining these with other layers of security – as seen with ESET LiveSense – has a much better chance of detecting malicious behavior and stopping it before any harm is done.

This is ESET

Proactive defense. Our business is to minimize the attack surface.

Stay one step ahead of known and emerging cyber threats with our **prevention-first approach, powered by AI and human expertise**.

Experience best-in-class protection, thanks to our in-house global **cyber threat intelligence**, compiled and examined for over 30 years, which drives our extensive R&D network, led by **industry-acclaimed researcher**s. ESET protects your business so it can unlock the full potential of technology.

ESET®

Digital Security
Progress. Protected.