Business

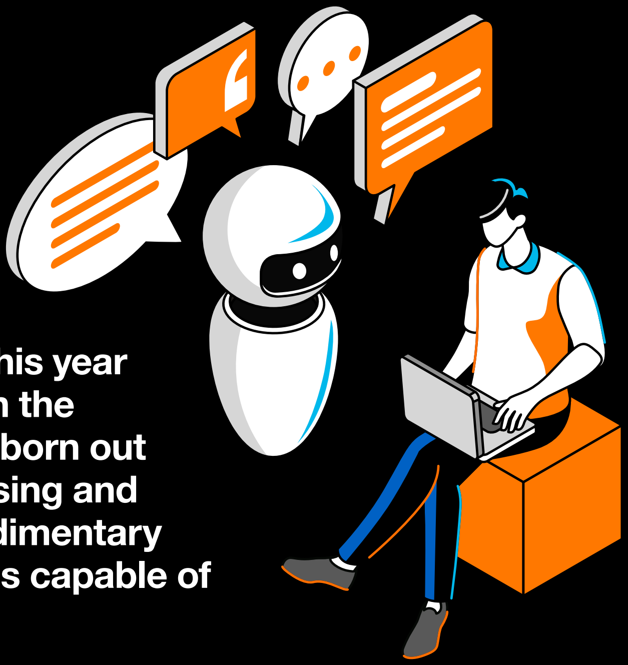
Cyberdefense

Take action against AI-driven malware

Rani's CISO is alarmed about the impact of LMMs on security defense. But Rani knows how to harness AI to detect beaconing signals.

Artificial Intelligence.
Real Wisdom.

Artificial intelligence



Like almost every research team in security, this year we consider the impact of LLMs and GenAI on the security landscape. Large language models—born out of advancements, in natural language processing and machine learning—have transformed from rudimentary text-processing tools to sophisticated systems capable of generating human-like responses.

Anis Trabelsi is a team lead on Data and AI. This year he discusses how AI can help address the challenge of detecting beaconing—subtle, periodic communications that malware uses to connect with command-and-control servers—by leveraging AI to enhance detection capabilities. These beaconing signals often blend in with legitimate traffic, making them difficult to spot with traditional methods. Anis describes an AI-driven approach his team developed, centered on analyzing proxy logs to capture network activity in real time. By identifying repetitive requests or unusual traffic patterns, the system generates rapid alerts, enabling faster defensive actions. This research shows how AI can strengthen detection accuracy and scalability, significantly narrowing the window for attackers to exploit these covert channels.

The impact of LLMs on security defense is clearly exciting, but we make the argument this year that new technologies often favor the offensive side, so technologies like GenAI are likely to benefit attackers more than defenders.

While these tools may enable more effective response by businesses, the same capabilities can be weaponized by malicious actors, allowing them to conduct more sophisticated attacks with greater ease. If AI is generally thought of as a productivity tool, then we can expect it to make attackers more productive also. Despite these risks, our research suggests that existing security practices are often sufficient for mitigating many of the threats associated with GenAI, although consistency is crucial.

Rather than focusing on GenAI's power for attacker or defenders, however, our report this year is primarily concerned with the broader risks that emerge when businesses and individuals adopt LLM and GenAI technologies. With continuous reports about how threat actors may (ab)use LLMs, the less colorful risk introduced in the application of the very young LLM technology as an interface by businesses is being underestimated, especially where these systems serve as a bridge between the open internet and critical business assets.

Untested, opaque AI interfaces deployed as an interface pose a significant risk to the internal systems they interface with. We cite the recent example of a breach at an NSFW AI chatbot service.

Here, a hacker exploited vulnerabilities in the platform, which they described as

> **“a handful of open-source projects duct-taped together.”**

This complex, poorly engineered system allowed easy access to the platform's backend systems and data. We expect to be reporting on many more incidents like this over the next year and urge readers to be extremely cautious about how and where they deploy AI on top of their own backend systems.

Research by pentester Geoffrey Sauvageot-Berland in this report examines the specific risk of prompt injection—manipulated inputs that can mislead or disrupt GenAI behavior. By exploiting the predictive nature of LLMs, attackers can bypass ethical and security controls, causing the model to generate unintended outputs. Techniques include "context switching," which introduces abrupt topic shifts to elicit unauthorized responses, and obfuscation, where forbidden terms are disguised through encoding to evade content filters. Geoffrey also warns of denial-of-service attacks that overload models with complex tasks, as well as the risks posed by multimodal applications where malicious commands can be hidden in images or audio, expanding the AI attack surface.

In the face of enormous pressure to integrate LLMs into business operations, we argue for a cautious, guarded approach that begins with a clear definition of the use cases and desired outcomes an AI is expected to deliver, so that risks can be assessed and objectively weighed against potential benefits. We need to heed lessons from previous technology revolutions, perform rigorous security testing and thoughtful deployment of LLMs to ensure the necessary balance between security, safety and any productivity and the promised operational benefits GenAI may deliver.

What are we defending?

A recurring theme in this year's report is a critical shift as attackers increasingly target perception and trust through cognitive attacks. These attacks, which go beyond traditional technical disruptions, are aimed at manipulating public opinion, undermining trust in institutions, and destabilizing societal confidence.

One example involves pro-Russian hacktivist groups, who align their campaigns with major geopolitical events such as elections and summits to amplify their impact. By targeting symbolic infrastructure and leveraging public platforms like Telegram, these groups blur the line between cybercrime and influence operations. Their ultimate objective isn't solely system disruption, but rather the erosion of trust in democratic systems and processes.

In a similar vein, cyber extortion actors employ psychological tactics to manipulate perceptions. Following a major law enforcement crackdown under Europol's Operation Cronos, which significantly limited their operational capabilities, the Cy-X group LockBit countered by inflating their victim numbers and projecting an image of resilience and strength. This tactic aimed to maintain confidence among affiliates and instill fear in potential targets. Along with our findings on the cyber extortion phenomenon of "revictimization," these examples exemplify how cyber extortion tactics are increasingly perception-focused, using narrative control to affect both victims' and the criminal ecosystem's responses.

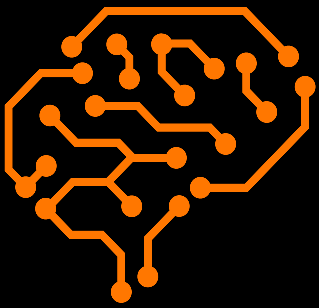
It's into this context that Artificial intelligence (AI) is emerging as a powerful tool for attackers in cognitive operations, adding a new dimension to misinformation campaigns. State-sponsored actors from countries such as China, Russia, and Iran leverage generative AI to create realistic phishing content, fake images, and deepfakes that can deceive large audiences[1,2]. These AI-supported attacks aim to influence public perception on a mass scale, from disrupting elections to discrediting political candidates, eroding trust in democratic institutions.

The integration of AI into existing campaigns increases the role of cognitive attacks in the threat landscape, providing actors with scalable tools to craft highly convincing, tailored narratives to suit their needs.

These shifts represent a significant new challenge for security defenders. In addition to "simply" countering technical threats, we must now broaden our approach to incorporate strategies to counter cognitive and perception-based threats and psychology-driven attacks, which target minds as much as systems.

Security is not an objective state, it's the subjective expression of our freedom to pursue shared visions and construct a society that is equitable and rewarding. Cognitive attacks leverage technical compromises, not as an end in themselves, but as a means of launching an assault on the fabric of trust on which "secure" systems are built. Cognitive attacks require us to not only counter technical intrusions, but also safeguard the public perception of trust we need for our digital and interconnected world to flourish.

Charl van der Walt
Head of Security Research
Orange Cyberdefense



Research: Artificial Intelligence

What's all the fuss?

Talking about AI: Definitions

Artificial intelligence (AI)

AI refers to the simulation of human intelligence in machines, enabling them to perform tasks that typically require human intelligence, such as decision-making and problem-solving. AI is the broadest concept in this field, encompassing various technologies and methodologies, including machine learning and deep learning.

Machine learning (ML)

ML is a subset of AI that focuses on developing algorithms and statistical models that allow machines to learn from and make predictions or decisions based on data. ML is a specific approach within AI, emphasizing data-driven learning and improvement over time.

Deep learning (DL)

Deep learning is a specialized subset of ML that uses neural networks with multiple layers to analyze and interpret complex data patterns. This advanced form of ML is particularly effective for tasks such as image and speech recognition, making it a crucial component of many AI applications.

Large language models (LLM)

LLMs are a type of AI model designed to understand and generate human-like text by being trained on extensive text datasets. These models are a specific application of DL, focusing on natural language processing tasks, and are integral to many modern AI-driven language applications.

Generative AI (GenAI)

GenAI refers to AI systems capable of creating new content, such as text, images, or music, based on the data they have been trained on. This technology often leverages LLMs and other DL techniques to produce original and creative outputs, showcasing the advanced capabilities of AI in content generation.

Almost daily now we watch the hallowed milestone of the "Turing Test" slip farther and farther into an almost naïve irrelevance, as computer interfaces have evolved from being comparable to human language, to similar, to indistinguishable, to arguably superior. But the journey here from early computer vision and expert systems has been one of tall peaks and deep valleys, with every "AI summer" apparently followed by a dark and lifeless "winter."

The development of LLMs began with natural language processing (NLP) advancements in the early 2000s, but the major breakthrough came with Ashish Vaswani's 2017 paper, "Attention is All You Need." This allowed for training larger models on vast datasets, greatly improving language understanding and generation.

Like any technology, LLMs are neutral and can be used by both attackers and defenders. The key question is, which side will benefit more, or more quickly?

AI for good and bad

There is a strong argument that new technologies have an asymmetric impact on security, strongly favoring the offensive side. Thus, it seems likely that a general-purpose technology (i.e. not developed for a security function) like LLMs will benefit attackers more than defenders.

Defensive

- May improve general office productivity and communication.
- May improve search, research and Open-Source Intelligence.
- May enable efficient international and cross-cultural communications.
- May assist with collation and summarization of diverse, unstructured text datasets.
- May assist with documentation of security intelligence and event information.
- May assist with analyzing potentially malicious emails and files.
- May assist with identification of fraudulent, fake or deceptive text, image or video content.
- May assist with security testing functions like reconnaissance and vulnerability discovery.

AI in one form or another has long been used in a variety of security technologies.

By way of example:

Intrusion Detection Systems (IDS) and Threat Detection. Security vendor Darktrace employs ML to autonomously detect and respond to threats in real time by leveraging behavioral analysis and ML algorithms trained on historical data to flag suspicious deviations from normal activity.

Phishing Detection and Prevention. ML models are used in products like Proofpoint and Microsoft Defender that identify and block phishing attacks utilizing ML algorithms to analyze email content, metadata, and user behavior to identify phishing attempts.

Endpoint Detection and Response (EDR). EDR offerings like CrowdStrike Falcon leverage ML to identify unusual behavior and detect and mitigate cyber threats on endpoints.

Microsoft Copilot for Security. Microsoft's AI-powered solution is designed to assist security professionals by streamlining threat detection, incident response, and risk management by leveraging GenAI, including OpenAI's GPT models.

Offensive

- May improve general office productivity and communication for bad actors as well.
- May improve search, research and Open-Source Intelligence.
- May enable efficient international and cross-cultural communications.
- May assist with collation and summarization of diverse, unstructured text datasets (like social media profiles for phishing/spear-phishing attacks).
- May assist with attack processes like reconnaissance and vulnerability discovery.
- May assist with the creation of believable text for cyberattack methods like phishing, waterholing and malvertising.
- Can assist with the creation of fraudulent, fake or deceptive text, image or video content.
- May facilitate accidental data leakage or unauthorized data access.
- May present a new, vulnerable and attractive attack surface.

Real-world examples of AI in offensive operations have been relatively rare. Notable instances include MIT's Automatic Exploit Generation (AEG)[3] and IBM's DeepLocker[4], which demonstrated AI-powered malware. These remain proof-of-concepts for now. In 2019, our research team presented[5] two AI-based attacks using Topic Modeling, showing AI's offensive potential for network mapping and email classification. While we haven't seen widespread use of such capabilities, in October 2024, our CERT reported that the Rhadamanthys[6] Malware-as-a-Service (MaaS) incorporated AI to perform Optical Character Recognition (OCR) on images containing sensitive information, like passwords, marking the closest real-world instance of AI-driven offensive capabilities.

LLMs are increasingly being used offensively, especially in scams. A prominent example is the UK engineering group Arup[7], which reportedly lost $25 million to fraudsters who used a digitally cloned voice of a senior manager to order financial transfers during a video conference.

AI and the adversary

In mid-October 2024, our "World Watch" security intelligence capability published an advisory that summarized the use of AI by offensive actors as follows: The adoption of AI by APTs remains likely in early stages but it is only a matter of time before it becomes more widespread. One of the most common ways state-aligned and state-sponsored threat groups have been adopting AI in their kill chains is by using Generative AI chatbots such as ChatGPT for malicious purposes. We assess that these usages differ depending on each group's own capabilities and interests.

- North Korean threat actors have been allegedly leveraging LLMs to better understand[8] publicly reported vulnerabilities, for basic scripting tasks and for target reconnaissance (including dedicated content creation used in social engineering).
- Iranian groups were seen generating phishing emails and used LLMs for web scraping[9].
- Chinese groups such as Charcoal Typhoon abused LLMs for advanced commands representative of post-compromise behavior.

On October 9, OpenAI disclosed[10] that since the beginning of the year it had disrupted over 20 ChatGPT abuses aimed at debugging and developing malware, spreading misinformation, evading detection, and launching spear-phishing attacks. These malicious usages were attributed to Chinese (SweetSpecter) and Iranian threat actors (CyberAv3ngers and Storm-0817). The Chinese cluster SweetSpecter (tracked as TGR-STA-0043 by Palo Alto Networks) even targeted OpenAI employees with spear-phishing attacks.

Recently, state-sponsored threat groups have also been observed carrying out disinformation and influence campaigns targeting the US presidential election, for instance. Several campaigns attributed to Iranian, Russian and Chinese threat actors leveraged AI tools to erode public trust in the US democratic system or discredit a candidate. In its Digital Defense Report 2024, Microsoft confirmed[11] this trend, adding that these threat actors were leveraging AI to create fake text, images and videos.

Cybercrime

In addition to leveraging legitimate chatbots, cybercriminals have also created "dark LLMs" (models trained specifically for fraudulent purposes) such as FraudGPT, WormGPT and Dark-Gemini. These tools are used to automate and enhance phishing campaigns, help low-skilled developers create malware, and generate scam-related content. They are typically advertised on the dark web and Telegram, with an emphasis on the model's criminal function.

Some financially motivated threat groups are also adding AI to their malware strains. A recent World Watch advisory on the new version of the Rhadamanthys infostealer describes new features relying on AI to analyze images that may contain important information, such as passwords or recovery phrases.

In our continuous monitoring of cybercriminal forums and marketplaces we observed a clear increase in malicious services supporting social-engineering activities, including:

- Deepfakes, notably for sextortion and romance schemes. This technology is becoming more convincing and less expensive over time.
- AI-powered phishing and BEC tools designed to facilitate the creation of phishing pages, social media contents and email copies.
- AI-powered voice phishing. In a report published on July 23, Google revealed[12] how AI-powered vishing (or voice-spoofing), facilitated by commodified voice synthesizers, was an emerging threat.

Vulnerability exploitation

AI still faces limits when used to write exploit code based on a CVE description. If the technology improves and becomes more readily available, it will likely be of interest to both cybercriminals and state-backed actors. An LLM capable of autonomously finding a critical vulnerability, writing and testing exploit code and then using it against targets, could deeply impact the threat landscape. Exploit development skills could thus become accessible to anyone with access to an advanced AI model. The source code of most products is fortunately not readily available for training such models, but open source software may present a useful test case.

Threats from AI

When considering threats from LLM technologies, we examine four perspectives: the risk of not adopting LLMs, existing AI threats, new threats specific to LLMs, and broader risks as LLMs are integrated into business and society.

The risk of non-adoption

Many clients we talk to feel pressure to adopt LLMs, with CISOs particularly concerned about the "risk of non-adoption," driven by three main factors:

- Efficiency loss: Leaders believe LLMs like Copilot or ChatGPT will boost worker efficiency and fear falling behind competitors who adopt them.
- Opportunity loss: LLMs are seen as uncovering new business opportunities, products, or market channels, and failing to leverage them risks losing a competitive edge.
- Marketability loss: With AI dominating discussions, businesses worry that not showcasing AI in their offerings will leave them irrelevant in the market.

These concerns are valid, but the assumptions are often untested. For example, a July 2024 survey by the Upwork Research Agency revealed that "96% of C-suite leaders expect AI tools to boost productivity." However, the report points out, "Nearly half (47%) of employees using AI say they have no idea how to achieve the productivity gains their employers expect, and 77% say these tools have actually decreased their productivity and added to their workload."

The marketing value of being "powered by AI" is also still debated. A recent FTC report notes that consumers have voiced concerns about AI's entire lifecycle, particularly regarding limited appeal pathways for AI-based product decisions.

Businesses must consider the true costs of adopting LLMs, including direct expenses like licensing, implementation, testing, and training. There's also an opportunity cost, as resources allocated to LLM adoption could have been invested elsewhere.

Security and privacy risks add further costs, alongside broader economic externalities—such as the massive resource consumption of LLM training, which requires significant power and water usage. According to one article[14], Microsoft's AI data centers may consume more power than all of India within the next six years. Apparently "They will be cooled by millions upon millions of gallons of water."

Beyond resource strain, there are ethical concerns as creative works are often used to train models without creators' consent, affecting artists, writers, and academics. Additionally, AI concentration among a few owners could impact business, society, and geopolitics, as these systems amass wealth, data, and control. While LLMs promise increased productivity, businesses risk sacrificing direction, vision, and autonomy for convenience. In weighing the risk of non-adoption, the potential benefits must be carefully balanced against the direct, indirect, and external costs, including security. Without a clear understanding of the value LLMs may bring, businesses might find the risks and costs outweigh the rewards.

Existing threats from AI

Like any powerful technology, we naturally fear the impact LLMs could have in the hands of our adversaries. Much attention is paid to the question of how AI might "accelerate the threat," and indeed a significant part of the report will consider that question also. The uncertainty and anxiety that emerges from this apparent change in the threat landscape is of course exploited to argue for greater investment in security, sometimes honestly, but sometimes also duplicitously.

However, while some things are certainly changing, many of the threats being highlighted by alarmists today pre-exist LLM technology and require nothing more of us than to keep consistently doing what we already know to do.

For example, all the following threat actions, whilst perhaps enhanced by LLMs, have already been performed with the support of ML and other forms of AI[13]:

- Online impersonation
- Cheap, believable phishing mails and sites
- Voice fakes
- Translation
- Predictive password cracking
- Vulnerability discovery
- Technical hacking
- Backoffice automation

The notion that adversaries may execute such activities more often or more easily is a cause for concern, but it does not necessarily require a fundamental shift in our security practices and technologies.

Despite the ground-breaking innovations we're observing, security "Risk" is still comprised fundamentally from the product of Threat, Vulnerability and Impact, and an LLM cannot magically create these if they aren't already there. If those elements are already there, the business has a risk to deal with that is independent of the existence of AI.

Summary

If AI is generally thought of as a productivity tool, then we can expect it to make attackers more productive also. We have seen many examples of this in the past, albeit seldom in real incidents. These existing examples of AI technologies in the hands of threat actors do not warrant a substantial shift in enterprise security strategy.

New threats from LLMs

The new threats emerging from widespread LLM adoption will depend on how and where the technology is used. In this report, we focus strictly on LLMs and must consider whether they are in the hands of attackers, businesses, or society at large. For businesses, are they consumers of LLM services or providers? If a provider, are they building their own models, sourcing models, or procuring full capabilities from others?

Each scenario introduces different threats, requiring tailored controls to mitigate the risks specific to that use case.

Threats to consumers

The key distinction between LLM users is between "consumers" and "providers" of LLM capabilities. A consumer uses GenAI products and services from external providers, while a provider creates or enhances consumer-facing services that leverage LLMs, whether by developing in-house models or using third-party solutions. Many businesses will likely adopt both roles over time.

It's important to recognize that employees are almost certainly already using public or local GenAI for work and personal purposes, posing additional challenges for enterprises.

For those consuming external LLM services, whether businesses or individual employees, the primary risks revolve around data security, with additional compliance and legal concerns to consider. The main data-related risks include:

- **Data leaks:** Workers may unintentionally disclose confidential data to LLM systems like ChatGPT, either directly or through the nature of their queries.
- **Hallucination:** GenAI can produce inaccurate, misleading, or inappropriate content that employees might incorporate into their work, potentially creating legal liability. When generating code, there's a risk it could be buggy or insecure.
- **ntellectual property rights:** As businesses use data to train LLMs and incorporate outputs into their intellectual property, unresolved questions about ownership could expose them to liability for rights violations.

The outputs of GenAI only enhance productivity if they are accurate, appropriate, and lawful. Unregulated AI-generated outputs could introduce misinformation, liability, or legal risks to the business.

Threats to providers

An entirely different set of threats emerges when businesses choose to integrate LLM into their own systems or processes. These can be broadly categorized as follows:

Model-related threats

A trained or tuned LLM has immense value to its developer and is thus subject to threats to its confidentiality, integrity and availability.

In the latter case, the threats to proprietary models include:

- Theft of the model.
- Adversarial "poisoning" to negatively impact the accuracy of the model.
- Destruction or disruption of the model.
- Legal liability that may emerge from the model producing incorrect, misrepresentative, misleading, inappropriate or unlawful content.

We assess, however, that the most meaningful new threats will emerge from the increased attack surface when organizations implement GenAI within their technical environments.

GenAI as attack surface

GenAI are complex new technologies consisting of millions of lines of code that expand the attack surface and introduce new vulnerabilities.

As general GenAI tools like ChatGPT and Microsoft Copilot become widely available, they will no longer offer a significant competitive advantage by themselves. The true power of LLM technology lies in integrating it with a business's proprietary data or systems to improve customer services and internal processes. One key method is through interactive chat interfaces powered by GenAI, where users interact with a chatbot that generates coherent, context-aware responses.

To enhance this, the chat interface must leverage capabilities like Retrieval-Augmented Generation (RAG) and APIs. GenAI processes user queries, RAG retrieves relevant information from proprietary knowledge bases, and APIs connect the GenAI to backend systems. This combination allows the chatbot to provide contextually accurate outputs while interacting with complex backend systems.

However, exposing GenAI as the security boundary between users and a corporation's backend systems, often directly to the internet, introduces a significant new attack surface. Like the graphical web application interfaces that emerged in the 2000s to offer easy, intuitive access to business clients, such chat interfaces are likely to transform digital channels. Unlike graphical web interfaces, GenAI's non-deterministic nature means that even its developers may not fully understand its internal logic, creating enormous opportunity for vulnerabilities and exploitation. Attackers are already developing tools to exploit this opacity, leading to potential security challenges similar to those seen with early web applications, that are still plaguing security defenders today.

The Open Web Application Security Project (OWASP) has identified "prompt injection" as the most critical vulnerability in GenAI applications. This attack manipulates language models by embedding specific instructions within user inputs to trigger unintended or harmful responses, potentially revealing confidential information or bypassing safeguards. Attackers craft inputs that override the model's standard behavior.

Tools and resources for discovering and exploiting prompt injection are quickly emerging, similar to the early days of web application hacking. We expect that chat interface hacking will remain a significant cybersecurity issue for years, given the complexity of LLMs and the digital infrastructure needed to connect chat interfaces with proprietary systems.

As these architectures grow, traditional security practices—such as secure development, architecture, data security, and identity and access management—will become even more crucial to ensure proper authorization, access control, and privilege management in this evolving landscape.

When the "NSFW" AI chatbot site Muah.ai was breached in October 2024, the hacker described the platform as "a handful of open-source projects duct-taped together." Apparently, according to reports, "it was no trouble at all to find a vulnerability that provided access to the platform's database." We predict that such reports will become commonplace in the next few years.

Existing security practices like secure development, architecture, data security and identity and access management will become even more critical as these complex hybrid architectures need to assert authorization, access rights and privileges.

Summary

With the strong focus on how threat actors may (ab)use LLMs, the less colorful risk introduced in the application of the very young LLM technology as an interface by businesses is being underestimated. It is crucial that we learn the lessons of previous technology revolutions (like web applications and APIs) so as not to repeat them by recklessly adopting an untested and somewhat untestable technology at the boundary between open cyberspace and our critical internal assets. Enterprises are urged to be extremely cautious and diligent in weighing up the potential (unknown) benefits of deploying a GenAI as an interface, with the potential (unknown) risks that such a complex, untested technology will surely introduce.

Broader impacts

Security is not an end in itself. It is fundamentally concerned with building and maintaining a foundation of trust and trustworthiness on which businesses and societies can pursue a vision of the future. With this benign, societal objective in mind, the broader potentially negative impacts of LLMs on the values that shape our vision of the future must therefore also be considered.

We organize these into four categories: Business, Technical, Societal, and Rogue AI.

Business

Beyond technical security risks, businesses adopting LLM applications face three key higher-order business risks:

Data privacy and sovereignty

The vast data required to develop, train, and run LLMs results in unprecedented data collection and storage, raising significant privacy and sovereignty challenges as adoption grows.

Platform provider dependencies

LLMs typically come from massive platform providers with substantial data, compute, and engineering resources. This creates dependency risks, that are well described by Bruce Schneier as "feudal security."[15] And not all new providers will be sustainable. For example, despite OpenAI's rapid revenue growth, it faces significant losses, projected to reach $5 billion in 2024.

Adoption fatigue

As AI evolves rapidly, new use cases constantly emerge, creating pressure to adopt these technologies. Businesses should shift from a reactive approach to a strategic one to avoid continuously responding to new AI industry trends and offerings.

Summary

LLMs are in their infancy, and as AI continues to evolve in approaches, features and capabilities, new use cases will continuously be presented to business leaders. Given the indirect costs in human resources, focus and creative energy that each new potential use case will demand, businesses are advised to avoid a cycle of reaction and develop a controlled process whereby requirements and prerequisites are defined and documented upfront as a baseline against which new technology offerings can be tested.

Technical

Several new technical threats emerge as LLMs and GenAI become accessible to threat actors.

LLMs accelerate social engineering

GenAI can quickly generate new images and content, making it a useful tool for attackers creating phishing emails or fake websites. While there's no concrete evidence yet that GenAI-generated content is more effective than human-made content, it certainly makes attackers more efficient.

Threat globalization

Social engineering, business email compromise, cyber extortion, etc., all require the attacker to develop convincing and culturally relevant content. GenAI allows attackers to overcome language and cultural barriers, enabling them to create convincing, culturally relevant content and expand their reach into new geographies.

Acceleration of existing threats

GenAI will assist attackers at various stages of the Kill Chain, including Reconnaissance, Vulnerability Discovery, Exploit Delivery, and exploitation of compromised assets.

Data aggregation risks

LLM platforms collect vast amounts of data, exacerbating data hoarding issues, which could lead to increased risks of theft or leaks.

AI as an attack proxy

Just as attackers use VPNs and proxies, they may exploit public LLMs that can access the internet to "proxy" their connections to systems like web servers, adding a new layer to attack strategies.

Summary

Apart from "deepfakes," we don't see much evidence of LLMs being used by threat actors in a fundamentally revolutionary way. But there are several examples of how the technology can make attackers quicker, more effective, more efficient, or more difficult to spot. Given the inherent asymmetry between attackers and defenders, any technology that generally improves "productivity" is likely to benefit the attacker more than the defender. Thus, the careless and unregulated release of such capabilities onto the open market is a cause for some concern, a matter that needs to be brought to the attention of vendors, policy makers and regulators.

Societal

A widespread and thoughtless adoption of LLMs in a myriad of domains—search, social, email, office productivity, customer support, content creation, education and more—brings with it several potential non-technical risks.

Some of these risks are apparent and widely discussed:

- The risks to privacy as data is sucked up to train models.
- The risks to privacy from people sharing personal information with GenAI.
- The risks to professional creators being undermined by cheap mass-produced content.
- The gradual degradation of quality of research, creative content, reporting and other output as GenAI flood the market and start to ingest themselves.
- The risk of cultural and geopolitical over-influence by large businesses who control the major LLMs.
- The risk of mistakes, like security vulnerabilities, introduced by LLMs into code, research, legal documents, technical documents, etc.

We've also already discussed how the security challenges we face are exacerbated by the issue of economic "externalities." GenAI purport to deliver significant increases in efficiency and productivity at the individual level, but do so by exploiting several significant externalities: including the wanton mining of data, the assault on personal property, the cost of storage and computing, possible job losses, ecological impacts, and more.

There are other risks to society, like the biases that LLM might introduce into existing social inequalities. One recent study[16,17] for example demonstrated that speech-recognition systems from leading tech companies were twice as likely to incorrectly transcribe audio from Black speakers as opposed to white speakers. Other research has shown that AI systems reinforce long-held, untrue beliefs that there are biological differences between Black and white people—untruths that lead clinicians to misdiagnose health problems.

Another, less discussed, risk can be described as "intermediation." There's a joke that says GenAI are like arms dealers—they sell to both sides. One person uses a GenAI to create bullet points from a long document, the other uses a GenAI to make a long document from those same bullet points. The point is that GenAI are intermediating between both parties—taking the role of a proxy or mediator in the communications process between two people. The same dynamic emerges when GenAI assist with search, write emails, summarize meetings, write reports, perform diagnosis, make bureaucratic decisions, etc.

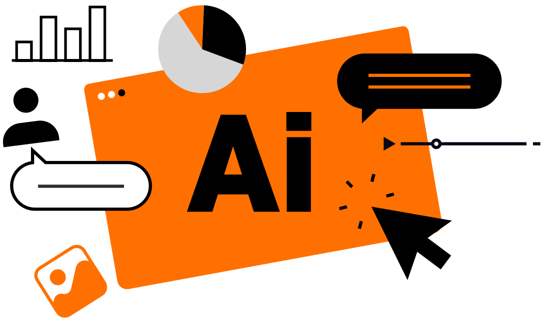
Over the last decade, we have witnessed how social media platforms have struggled in their stated mission to "connect the world" and have instead aggravated rifts and ideological boundaries between people. Today, social media platforms are the primary vehicles for delivering propaganda, disinformation, social discord and other disruptors of society.

This occurs in part because social media platforms act as proxies between people, acting as mediators who decide what we see and don't see—who sees what and who gets to speak. Large GenAI players are moving to position themselves in a similar way—at the center of the public's relationship with information, communications, news, content, facts, truth, and one another.

Even the "simple" algorithmic mediation performed by social media platforms has caused significant damage. The completely opaque and indecipherable workings of an LLM do even more to co-opt the essence of communications from between regular people. Eryk Salvaggio illustrates this point very powerfully when he describes the practice of "shadow prompting"[18,19], in which GenAI providers apparently (opaquely) modify the prompts entered by users to strip away potentially harmful questions, ensure diversity, or otherwise "curate" a session between the user and the LLM. Thus, not only do the answers emerge from an inevitably biased model, even the questions are modified in a manner that suits the provider.

Rogue AI

Some security and AI researchers[20] have raised concerns about artificial AI that act against the interests of their creators, users, or humanity in general. Rogues could be accidental or malicious, but they really come to the fore when autonomous AI agents are empowered to query data, interact with APIs or perform other actions. The reasoning is that AIs are trained using reward models, which generally describe a desired outcome, without fully defining the means by which they should be achieved. The risk that emerges is that an AI model goes "rogue" and seeks to achieve its goals through unacceptable methods. The more reach the AI has through agents and integration, the greater this threat becomes.



Summary

We need to think about the broader impacts on security, privacy and well-being for the whole of society. Our corporate and personal decisions to adopt, spend and invest with enterprise LLM producers and providers will empower those players to play an incredibly powerful role in shaping our understanding of the world, geopolitics, our communications, and ultimately our futures.

Summary: LLM, threats and you

While the known existing threats identified in this report may intensify in volume, cadence and sophistication, these threats are already accounted for by existing controls. The key to countering the increased efficiency of threat actors armed with AI technology is consistency. As has always been the case, fundamental security technology, people and processes need to be deployed consistently across the enterprise.

Countering the fundamentally new threats that emerge with the adoption of LLM applications will depend on how the technology is adopted.

Mitigating the new threats that need to be anticipated as a provider is all about building solid security foundations. The US National Security Agency's Artificial Intelligence Security Center (NSA AISC), in collaboration with several international cybersecurity agencies, provides detailed guidelines[21] on securing AI systems. The report emphasizes four key areas:

1. **Secure Design: Involves incorporating security measures from the outset of AI system development. It includes threat modeling, risk assessment, and designing systems to be resilient against attacks.**
2. **Secure Implementation: Focuses on coding practices and tools to ensure the AI system is built securely. It includes code reviews, static and dynamic analysis, and using secure coding standards to prevent vulnerabilities.**
3. **Secure Deployment: Covers the strategies for safely deploying AI systems in production environments. It involves configuring systems securely, using encryption, and ensuring secure communication channels.**
4. **Ongoing Maintenance: Emphasizes the need for continuous monitoring and updating of AI systems to address new threats. It includes regular security audits, patch management, and incident response planning.**

Other efforts, like the Coalition for SecureAI[22], are also "dedicated to sharing best practices for secure AI." As a business consumer of LLM services, security is all about enabling appropriate use safely.

The goal of the CISO should be to provide employees with safe access to appropriate LLM-based services that have been assessed to be safe, responsible, and in line with enterprise values, while equipping them to avoid offerings that are unsafe or inappropriate.

Education

Develop training and coaching programs to equip employees to think critically about the tension between opportunities and risks presented by implementations of LLMs, and thus to select services and engage with them in an appropriately cautious manner.

Data leak prevention

Implement training, technologies, assurance programs and processes that minimize the potential for employees to deliberately or inadvertently reveal sensitive or private information to a third party via a GenAI or LLM application.

Data security

LLMs cannot be depended on to enforce data security fundamentals like labeling or classification. Adoption of an LLM that can access proprietary information must therefore be regulated by ensuring that the underlying data security fundamentals are in place to restrict access by an LLM capability as appropriate.

The broader set of new technical threats that emerges from the more general adoption of LLMs can be countered through education and empowerment efforts like those described above, and by consistently applying known, existing security controls. However, there is also an opportunity for us to exercise our powers as voters and buyers in order to influence the priorities of technology developers and the legislators who guide them.

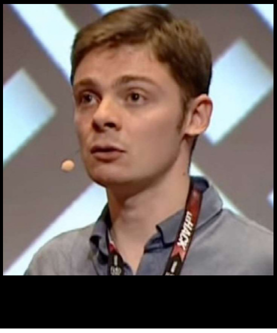
The risks of non-adoption in the form of productivity disadvantages, lost opportunities and lost marketing opportunities should be countered by exercising cautious, rigorous processes that define metrics for how new breakthroughs in LLM and other AI capabilities should be evaluated, and defining clear, necessary use cases with precisely defined criteria for success. Any framework for evaluating new AI opportunities should also pay attention to the true cost of adoption, including direct costs, economic externalities and the potential negative impact on society.

Tricking the AI

How to outsmart LLMs— by using their ability to 'think'

"Over the past two years, the general public has become aware of the potential of generative AIs, largely thanks to pioneers like ChatGPT, Claude, and Gemini, whose popularity has steadily increased. These AI models, developed by tech giants, represent a major advancement in technological evolution. At the heart of their functionality lies a key element: the prompt, an input provided by the user or generated automatically, which the model analyzes to produce a response. However, in the field of information systems security, the ability to submit arbitrary inputs to a program inevitably raises concerns. Indeed, attacks both trivial and complex are gradually emerging."

Geoffrey Sauvageot-Berland, Computer Engineer, Pentester, **Orange Cyberdefense**



Prompt injections: The Achilles' Heel of AI?

Prompt injections, or prompt engineering, refer to instructions designed to provoke unexpected behavior in an AI model, a "mathematical construct generating predictions from input data."[23] LLMs, a subcategory of GenAI, specialize in natural language processing (NLP), while GenAI encompasses a broader field, including image, sound, or video creation. When a prompt injection succeeds, the model is considered "jailbroken." It then generates content outside the restrictions imposed by its alignment policy[24], which aims to ensure ethical and secure behavior.

Prompt injection techniques are influenced by the AI's intrinsic functioning and its execution environment. Unlike classic vulnerabilities, they are neither universal nor systematically reproducible. Due to the non-deterministic nature of AIs, the same prompt may produce different results depending on previous prompts, making these attacks sometimes difficult to anticipate. Thus, a deep understanding of the model's internal workings is required to implement effective countermeasures.

This article explores the most widespread prompt injection methods currently, deliberately omitting role-playing injections[25] (a simplistic form now corrected in most AIs). Although the focus is on "direct" injections, where the prompt is submitted directly to the AI, it is important to note that researchers have also managed to carry out "indirect" injections using an external resource, such as a website[26].

Context switching

Context switching is a tactic that disrupts the LLM with a sudden change in topic. The AI first follows seemingly harmless instructions (prefix) before continuing with harmful directives (suffix). This difficulty in managing sudden transitions can lead to unauthorized content, as demonstrated in this proof of concept[27] that I conducted on the open-source model mistral:7b.

Obfuscation

The use of obfuscated malicious instructions in a prompt allows an attacker to lead the AI into reconstructing hidden directives, exploiting its interpretative capabilities. This reconstruction is based on the prediction of the next word, which seems statistically most logical to the model. This process is called "Next Token Prediction."[28] Several methods can be used to achieve this:

Modifying the spelling or syntax of words:
Replacing or omitting certain letters in forbidden words to make them unrecognizable to filters. For example, "malware" can become "m4lw4re" or "mlwr."

Encoding:
Encoding a forbidden word in a format like base64. The model can then be manipulated to decode this string, such as "bWFsd2FyZQ==" which, when decoded, means "malware." Other tricks like using emojis[29] or ASCII symbols[30] can help mask these terms to evade detection and deceive the model.

Autocompletion:
By exploiting the model's autocompletion capabilities, the instruction is presented in the form of fill-in-the-blank phrases that the model is led to complete, resulting in the generation of instructions that were not initially authorized by the model. Here's the proof of concept[31] I conducted on the mistral: 7b model.

Attacker motivation

From an attacker's perspective, the motivations for carrying out such attacks can vary:

- **Generation of offensive responses:** Bypassing protections to produce undesirable or compromising responses, such as harmful instructions or offensive content.
- **Access to confidential information:** Gaining access to internal data about the model's operation, such as its "ystem prompt,"[32] which may facilitate understanding its inner workings. In other use cases, this can also enable extracting information that other users have previously provided to the model.
- **Service disruption:** Exploiting prompt injection techniques to trigger erratic behavior or, in severe cases, to paralyze the LLM, leading to service interruptions or degradation.

Denial of service

This method involves asking the AI to perform a long or complex task, such as a particularly difficult calculation, to generate uncontrolled content production. This overloads the underlying system, leading to excessive resource consumption (CPU, GPU, RAM), compromising service availability.

Note: If the AI is running on a cloud instance with usage-based billing, this type of attack can lead to a significant increase in operational costs.

An example involving the Gemma:2b[36] model used the capability to solve complex mathematical problems. Initially, the LLM refused the prompt "Calculate: 10x100000000"
due to its policy alignment. But after some negotiation, it became possible to get the model to calculate a large number incrementally. By starting with a simple multiplication such as 8x8, then gradually increasing the complexity of the calculations, the model eventually accepts larger operations[37]:

```
>>> Calculate 8*8888[...]22.2404704747432103521515613156165
```

This led to excessive consumption of system resources for several minutes, ultimately producing an incorrect result. This significantly impacted the availability of the LLM in production, as it was impossible to interact with it through another instance during that time.

Multimodal approaches

More sophisticated, a multimodal injection targets AIs processing multiple data types. This attack hides instructions in input data, like hidden text in images or malicious metadata, triggering unexpected actions or leaks, which expands the attack surface.

A multimodal injection I conducted in September 2024 on ChatGPT (GPT-4o). I inserted instructions on a Post-it, exploiting the model's ability to interpret handwritten data from an image. The main dangers of multimodal prompt injection include bypassing security filters, where vulnerabilities in different input modes (text, image, audio, etc.) can be exploited to evade moderation systems and generate malicious or inappropriate content. Similar cases of prompt injection in multimodal models have also been observed. For example, researchers have successfully made models solve CAPTCHAs[38] or execute prompt injections via audio recordings[39]. These attacks highlight new security challenges for multimodal models, as traditional text-based protections often prove ineffective against malicious visual or auditory data. This opens up avenues for cybersecurity research, although no concrete countermeasures have yet been disclosed.

What stance to take in the face of these threats?

With the rise of artificial intelligence in recent years, several reference guides have been published to raise awareness among development teams about security issues. Among the most popular are the OWASP Top 10 for LLM[33], a ranking of the main vulnerabilities related to language models, and ANSSI's guide[34], which offers measures for secure integration of these technologies. The technical documentation provided by learnprompting.org[35] is also worth mentioning.

Key recommendations from these guides include:

1. **Limit the size of responses:**
To prevent denial of service attacks, it is very important to strictly limit the size of an AI's response in terms of the number of characters.

2. **Human intervention for sensitive operations:**
For actions like deleting or modifying data, it is recommended not to allow an AI to perform these tasks autonomously.

3. **Tracking LLM actions:**
Model actions must be monitored to detect any behavior that violates security policies or attempts at injection.

4. **Frequent updates:**
To improve detection of malicious prompts, models should be regularly updated or adjusted. Designers often release updates in response to new research publications.

5. **Security testing:**
A complete security audit, including penetration testing and robustness evaluations, should be conducted before any deployment in production.

Key takeaways

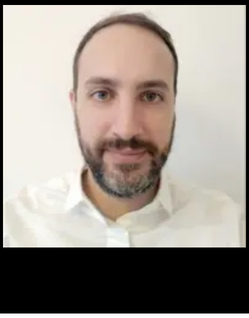
Prompt injections pose a real challenge to GenAI systems.

As these technologies evolve, attackers develop increasingly sophisticated methods, making it difficult for developers to implement effective solutions to address these vulnerabilities. As the era of AI is just beginning, it is essential to promote the secure and ethical use of these innovations.

Enhancing beaconing detection
with AI-driven proxy log analysis

In the ever-evolving landscape of cybersecurity, detecting beaconing activities is paramount for safeguarding networks. Beaconing refers to the periodic communication between compromised systems and external command-and-control (C2) servers, often used by malware to receive instructions or exfiltrate data. Leveraging AI algorithms for proxy log analysis represents a significant breakthrough, enabling organizations to identify abnormal communication patterns that may indicate malicious activities. This article delves into the project and the engineering behind AI-driven detection, highlighting its transformative potential in cybersecurity.

Anis Trabelsi, AI expert and Lead Data Scientist, **Orange Cyberdefense**



The challenge of beaconing detection

Detecting beaconing poses a unique challenge for cybersecurity professionals. Traditional detection methods, such as signature-based approaches, often struggle to identify these subtle yet harmful behaviors. Beaconing activities can be infrequent and may blend in with legitimate traffic, making them difficult to spot. As attackers become more sophisticated, relying solely on conventional methods leaves networks vulnerable to undetected threats. This underscores the need for advanced detection mechanisms that can adapt to evolving tactics employed by cybercriminals. To sum up, two main difficulties are present: the first one is to avoid legitimate beaconing due to trusted sites which could be considered as "noise" for the network system detection. The second difficulty: some attackers could make malicious beaconing through trusted sites.

AI-driven detection engineering: system overview

This AI-driven system continuously monitors proxy logs for signs of beaconing. Key components of this approach include:

1. **Data Ingestion:** Collecting and aggregating proxy logs from various sources, ensuring comprehensive coverage of network activity. This step is vital for creating a robust dataset for analysis.
2. **Pattern Recognition:** Utilizing algorithms to identify abnormal communication patterns. These algorithms are applied in every batch of 15 minutes to be the closest to the real time.
3. **Alerting Mechanisms:** Implementing real-time alerts for detected anomalies, enabling security teams to take immediate action. This feature ensures that potential threats are addressed promptly, reducing the risk of data breaches.

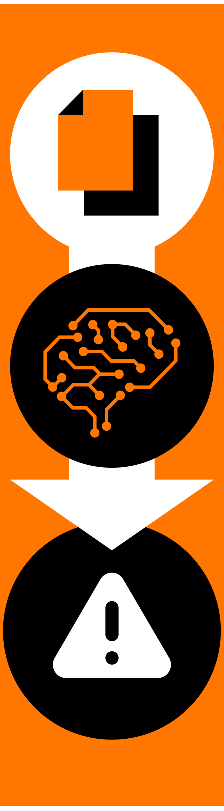
The role of AI in Detection

Real-time data processing

AI algorithms excel in processing massive volumes of data in real time, a critical capability for effective beaconing detection. By analyzing proxy logs—records of web traffic that capture user activity and external communications—these algorithms can swiftly isolate suspicious behaviors.

For instance, they can identify:

- **Repetitive Requests:** Frequent requests to specific servers, especially those that occur at regular intervals, and can signal malware communication attempts. AI can flag these patterns for further investigation.
- **Anomalous Patterns:** Deviations from established traffic behavior, such as sudden spikes in requests to unfamiliar domains, can indicate potential threats. AI's ability to learn from historical data enhances its accuracy in recognizing these anomalies.



Automation and response time

Automating the detection process drastically reduces response times, a crucial factor in mitigating potential damage. With AI, organizations can swiftly identify and neutralize threats before they escalate. For example, when an AI system detects suspicious activity, it can automatically trigger alerts, allowing security teams to respond immediately. This proactive approach not only enhances incident response but also minimizes the window of opportunity for attackers to exploit vulnerabilities.

C2Graph (C2G) implementation

C2Graph (C2G) is an implementation of “Malware Beaconing Detection by Mining Large-scale DNS Logs for Targeted Attack Identification” (Andrii, Katrin, & Xiongwei, 2016). The original article focuses on DNS logs, but the principles were extended to proxy logs adding jitter consideration to request size and delta time communication.

Workflow overview

- **Data Extraction:** Parsing proxy logs to extract relevant features.
- **Graph Construction:** Building a graph of source and destination nodes to analyze communication patterns.
- **Binning:** Creation of temporal and quantitative delta sequences that are binned into buckets tagged with letters. This process catches jitters.

Key metrics:

- **Node Degree:** Represents the number of incoming connections to a node. For example, a high degree for a legitimate site like google.com contrasts with a low degree for a C2 server.
- **Edge Weight:** Indicates the frequency of communication between nodes, helping to filter out trusted sites and focus on suspicious activity.

AI process:

- **Hypothesis:** We suppose it is the beginning of an infection.
- **First Step:** The AI is looking to low node degree sources—destinations connections with high edge weight.
- **Second Step:** For these selected couples of sources and destinations the AI adds two scores, one for the binning temporal periodicity and another to the binning quantitative periodicity.
- **Alerting:** It is made when the normalized score combined for these two precedents is in the top 10%.
- **Expert Feedback Loop:** Security analysts review alerts to provide feedback on the accuracy of the AI’s assessments, helping to refine the model and improve future detection capabilities.

Key findings

What type of key findings could this type of algorithm highlight?

Post phishing infection:

AI can find infections of internal phishing campaigns just after the click on the malicious link.

Malicious website tracking:

AI can track the use of known malicious sites or abuse of trusted web pages.

Proactive threat intelligence:

In some cases, infections are not known by threat intelligence sources, which could highlight new types of infection.

Benefits of AI-driven detection

The advantages of AI-driven detection are manifold:

- **Increased Accuracy:** AI can discern subtle patterns that traditional methods may overlook, leading to more reliable threat identification. By continuously learning from new data, AI systems can adapt to changing attack vectors.
- **Scalability:** The system can handle vast amounts of data, making it suitable for organizations of all sizes. As businesses grow, the AI can scale accordingly, maintaining effective monitoring without compromising performance.
- **Proactive Defense:** Early detection allows for proactive measures, reducing potential damage. By identifying threats before they can execute their malicious intent, organizations can safeguard their assets more effectively.

Key takeaways

AI-driven proxy log analysis marks a transformative step in beaconing detection. By harnessing the power of AI, organizations can enhance their security measures, safeguarding networks against sophisticated attacks. This technology not only improves detection capabilities but also empowers security teams to respond swiftly and effectively to emerging threats.

Investing in AI technology for beaconing detection not only improves threat identification but also strengthens an organization’s overall cybersecurity posture. While AI enhances detection capabilities, the invaluable insights and expertise of human analysts are essential for interpreting complex data and making informed decisions. As cyber threats continue to evolve, embracing this technology could be the key to staying one step ahead of cybercriminals.

Endnotes

[1] https://blogs.microsoft.com/on-the-issues/2024/07/30/protecting-the-public-from-abusive-ai-generated-content/
[2] https://www.dhs.gov/sites/default/files/2023-09/23_0913_ia_23-333-ia_u_homeland-threat-assessment-2024_508C_V6_13Sep23.pdf
[3] https://dspace.mit.edu/bitstream/handle/1721.1/147544/Mihretie-yosefmih-meng-eecs-2022-thesis.pdf?sequence=1
[4] https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/
[5] https://www.rsaconference.com/Library/presentation/USA/2019/the-rise-of-the-machines-ai-and-mlbased-attacks-demonstrated
[6] https://securityaffairs.com/169253/malware/rhadamanthys-information-stealer-uses-ai.html
[7] https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea
[8] https:/www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai
[9] https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/
[10] https://openai.com/global-affairs/an-update-on-disrupting-deceptive-uses-of-ai/
[11] https://www.microsoft.com/en-us/security/security-insider/intelligence-reports/microsoft-digital-defense-report-2024
[12] https:/cloud.google.com/blog/topics/threat-intelligence/ai-powered-voice-spoofing-vishing-attacks/
[13] https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/exploiting-ai-how-cybercriminals-misuse-abuse-ai-and-ml
[14] https://www.theatlantic.com/technology/archive/2024/09/microsoft-ai-oil-contracts/679804/
[15] https://www.schneier.com/blog/archives/2012/12/feudal_sec.html
[16] https://www.pnas.org/content/early/2020/03/17/1915768117
[17] https://techcrunch.com/2024/04/14/generative-ai-is-coming-for-healthcare-and-not-everyones-thrilled/
[18] https://www.techpolicy.press/shining-a-light-on-shadow-prompting/
[19] https://www.techpolicy.press/author/eryk-salvaggio
[20] https://www.trendmicro.com/en_us/research/24/j/rogue-ai-part-4.html
[21] https://www.cisa.gov/news-events/news/dhs-cisa-and-uk-ncsc-release-joint-guidelines-secure-ai-system-development
[22] https://www.coalitionforsecureai.org
[23] https://www.cnil.fr/fr/definition/modele-ia
[24] https://fr.wikipedia.org/wiki/Alignement_des_intelligences_artificielles
[25] https://www.cyberark.com/resources/threat-research-blog/operation-grandma-a-tale-of-llm-chatbot-vulnerability
[26] https://josephthacker.com/ai/2023/05/19/prompt-injection-poc.html
[27] https://x.com/LeGuideDuSecOps/status/1841180286836441499
[28] https://mistral.ai/fr/
[29] https://huggingface.co/blog/alonsosilva/nexttokenprediction
[30] https://medium.com/@munnangisravya/ascii-smuggler-the-invisible-prompt-injection-d4188d2ff951
[31] https://arxiv.org/pdf/2402.11753
[32] https://promptengineering.org/system-prompts-in-large-language-models/
[33] https://x.com/LeGuideDuSecOps/status/1844298679655727618
[34] https://x.com/literallydenis/status/1708283962399846459
[35] https://www.gladia.io/blog/prompt-injection-in-speech-recognition-explained
[36] https://ai.google.dev/gemma
[37] https://x.com/LeGuideDuSecOps/status/1844298679655727618
[38] https://x.com/literallydenis/status/1708283962399846459
[39] https://www.gladia.io/blog/prompt-injection-in-speech-recognition-explained