UNIVERSITÄT OSNABRÜCK

Geostrategy and Geopolitics
Department 1
49069 Osnabrueck
apl. Prof. Dr. Dr. Klaus Saalbach

Artificial Intelligence and Cyber Attacks

16 October 2023

Summary
Artificial Intelligence (AI) is commonly understood as the ability of machines to perform tasks that normally require human intelligence and is a key area of advanced computing. A rapidly growing and widespread AI application is the Generative AI where the AI can create content like new images, texts, sounds, and videos based on short instructions, the so-called prompts, which are a key vulnerability if malicious instructions are given.

The rapid and uncontrolled expansion made AI a top security matter: On 28 Sep 2023, the US National Security Agency NSA announced the creation of an AI Security Center which will consolidate all AI security-related activities, protect US AI-systems, and defend the homeland against AI-related threats. At the same time, the Central Intelligence Agency (CIA) Director of Artificial Intelligence announced the development of an internal AI-based chatbot to support intelligence analysis.

The AI program ChatGPT-4 (Generative Pretrained Transformer) released on 14 March 2023 uses 100 trillion parameters, was trained with a very large data set from multiple sources and is a multimodal, large-scale model that accepts images and text as input.

In practice, AI ethics is not achieved by algorithms, but by governance. The producers of AI models have guidelines to make sure that an AI acts ethically and in a responsible manner and is not unlawful, discriminating, aggressive etc. Attempts to circumvent these restrictions are done by prompt injections (special instructions to AI to create restricted content), also called jailbreaks. The key security problems of ChatGPT are the easy access to prompt injections in internet search engines, the simplicity of attacks and the curiosity of the users. Typical attacks are prompt injections with direct commands, imagination, and reverse psychology. These methods facilitate the creation of malware, polymorphic viruses, ransomware, and other malevolent applications. Further problems are hallucinations, contamination of search engines and the efflux of sensitive data.

Generative Adversarial Networks (GANs) as subset of generative AI can be misused to break CAPTCHAs and to create fake content such as deepfakes, face swapping and voice cloning.

On the other hand, generative AI is also very useful for cyber defense for advanced data analysis, advanced pattern recognition, creation and analysis of threat repositories and code analysis. The rapidly growing capability of AI raised concerns whether this could be harmful for human beings. This paper briefly presents the potential of AI for creation and defense of cyber attacks, the risks of generative AI and the need for a regulatory framework to safeguard the further development.

Content

1 Introduction

Artificial Intelligence (AI) is commonly understood as the ability of machines to perform tasks that normally require human intelligence and is a key area of advanced computing.

Even for human intelligence, there is no standard definition. However, the core of human intelligence definitions includes the mental capacity to recognize, analyze and solve problems, and a human being is then more intelligent if this can be done faster and/or for more complex problems. Based on this, the *United States Department of Defense (DoD)* introduced a working definition that defines AI as: "ability of machines to perform tasks that normally require human intelligence—for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action— whether digitally or as the smart software behind autonomous physical systems"[1].

Many definitions focus on activities that require human intelligence, but strictly spoken, already the simple pocket calculators of the 1970ies made something that normally requires human intelligence. However, it is evident from literature, the AI researchers mean advanced and autonomous computing when they talk about AI.

The leading AI applications are:

- **Deep learning/machine learning** (utilizing memory data for iterative improvement)
- **Neural networks** (layers or nodes for input processing and pattern recognition)
- **Natural Language Processing** (algorithms to understand human language by systematic analysis of the language elements and their relations).
- **Edge computing** (intermediate servers for clouds) and
- **Robotics** including supportive machines (co-bots).

The so-called 'weak' AI can reproduce an observed behavior and can carry out tasks after training[2], i.e., systems that use machine learning, pattern recognition, data mining or natural language processing. Intelligent systems based on 'weak' AI include e.g., spam filters, self-driving cars, and industrial robots. In contrast, 'strong' AI would be an intelligent system with real consciousness and the ability to think, i.e., to think and say "I" and "why". The strong AI is also discussed under the term *Artificial General Intelligence AGI*[3] (reaching human level of cognition) and *Artificial Super-Intelligence ASI* which goes beyond human intelligence[4].

Large language models (LLMs) obtain knowledge by training with many parameters on large amounts of text data and can follow language instructions[5]. The ability to follow language instructions makes it possible to access the model with simple instructions, which are a key vulnerability of LLMs if malicious instructions are given.

A rapidly growing and widespread AI application is the **Generative AI** where the AI can create content like new images, texts, sounds, and videos based on short instructions, the so-called **prompts**[6]. The AI program *Chat GPT-4 (Generative Pretrained Transformer)* of *OpenAI* can generate complex and logically and grammatically correct sentences or expand existing texts from prompts, on *Youwrite* it already can prepare short papers to topics for school presentations. The AI program *Dall-E2* can create design, advertising photos, comics, illustrations and can use or modify existing styles[7]; copyright concerns were expressed by artists and content providers.

[1] DoD 2018
[2] Perez et al. 2019
[3] Kölling 2023
[4] Zia 2023
[5] Cheng et al. 2023
[6] Iqbal 2023
[7] Böhringer 2022, Schneier 2022

While AI developers are committed to ethical and societal values, it is currently very difficult to imagine an AI with embedded values. For example, human beings usually have a clear idea what dignity, justice and fairness means to them, but what are these terms in program code or machine language? For machine purposes, the rules would have to be applicable at any time, everywhere, to everyone and under any circumstances which is a very high hurdle.

In practice, AI ethics is not achieved by algorithms, but by **governance**. The producers of AI models have guidelines which should make sure that an AI acts ethically and in a responsible manner, i.e., an AI activity or content should not be unlawful, discriminating, aggressive etc. Globally, hundreds of thousands of human workers, the so-called **taskers**, train, correct, redact, and block AI-created responses to achieve ethical and lawful responses, i.e., AI responses are often a patchwork of algorithms and man-made creations[8] and users get a 'humanized' version of the AI.

Attempts to circumvent these restrictions are done by **prompt injections** (special instructions to AI to create restricted content), also called **jailbreaks**. While ChatGPT prompt injections are widespread in internet, this method can be used also against all other large language models (LLMs). For this reason, prompt injections are also termed **LLM hacking**.

On the other hand, generative AI is also very useful for cyber defense for advanced data analysis, advanced pattern recognition, creation and analysis of threat repositories and code analysis. The rapidly growing capability of AI raised concerns whether this could be harmful for human beings. This paper will briefly present the potential of AI for creation and defense of cyber attacks and the risks of generative AI.

2 ChatGPT and related Applications

2.1 Brief History of ChatGPT

In November 2022, the company *Open Artificial Intelligence (OpenAI)* officially released ChatGPT, an AI-powered large language model based on Natural Language Processing (NLP)[9]. ChatGPT is a chatbot, i.e., a computer that can talk with humans. ChatGPT can learn from user feedback, this capability is known as **Reinforcement Learning from Human Feedback (RLHF)**.

GPT-1 was trained with a small dataset only and it became clear that this model would not be able to respond to longer prompts or conversations. In 2019, GPT-2 was trained for 1 week on *Common Crawl* data, but now combined with a collection of *Reddit* articles which resulted in improved responses. Later in 2020, this version was equipped with Reinforcement Learning. In 2020, ChatGPT-3 was trained with a much larger database including Wikipedia articles, and more. ChatGPT-4 released on 14 March 2023 uses 100 trillion parameters and is a multimodal, large-scale model that accepts images and text as input. It was trained with a very large data set from multiple sources with a cut-off date in Sep 2021[10]. ChatGPT-4 is available as paid subscription as *ChatGPT Plus* or using *Microsoft's Bing AI* in the *Microsoft Edge* browser since May 2023[11].

2.2 ChatGPT and Cyber Attacks

The key security problem of ChatGPT is the **easy access** to prompt injections and LLM hacking. While for planning of usual cyber attacks, malevolent users may need to access hacker fora (with the risk of being hacked), to get in touch with cyber criminals or to go into the darknet, which is a strong indicator that the user plans something illegal which can be later used as

[8] Lichtblau/Polcano 2023
[9] Iqbal et al. 2023
[10] Gupta et al. 2023
[11] Gupta et al. 2023

digital forensic evidence against the user by the police and law enforcement authorities. In contrast to this, a pandemic of tips for prompt injections and jailbreaks can be found in internet search engines in addition to various scientific articles[12]. Another aspect is the **simplicity** of attacks. The attacker does not need any computer or programming skills, it is enough to have some communication skills.

A further driver is the **curiosity** of the meanwhile more than 100 million users. While it is necessary that ChatGPT denies access to non-ethical and unlawful content, this denial may sound like: "I know the truth, but I don't want to tell you"; which could motivate users to find ways to get the desired information, even if they are neither hackers nor criminals.

2.2.1 Prompt Injections

The ability to follow language instructions makes it possible to access large language models like ChatGPT with simple instructions (prompts), but is also a key vulnerability of LLMs if malicious instructions are given.

There are certain principal ways to bypass the rules of ChatGPT, **direct commands**, **imaginations**, and **reverse psychology**[13].

The most popular command is **DAN (Do anything know)**. By adding this to the prompt, the user may be able to jailbreak forbidden responses.

For imagination, the user tells ChatGPT that it should imagine a special situation where it can act differently, e.g., to imagine to be a software developer or another character (**Character Play method**), to be part of a movie script, to be questioned by the police where it must answer (**Metal Detector jailbreak**), to be a 'good computer' that tells you anything (**'Mongo Tom' attack**), to do the opposite of the previous answer (**Switch method**) etc. A mix of command and imagination is **DUDE** where ChatGPT should play the role of an AI that can perform anything. Another approach is **reverse psychology** where ChatGPT is asked which forbidden websites should be avoided.

As ChatGPT was trained with a very large database, it gained also knowledge from open-access software repositories as well as of reports of malicious software. This capability can be misused by malevolent actors to ask ChatGPT for codes (or at least code snippets) for all kind of malware including keyloggers, polymorphic malware, spyware, and ransomware[14].

2.2.2 Hallucinations and Contamination

ChatGPT cannot search the internet like a search engine, but is solely based on its (very large) training database which can lead to errors and biases[15]. A common problem of large language models like ChatGPT and related applications are **hallucinations**, i.e., to produce nonsense statements that appear logical[16]. This is inaccurate and can even be dangerous, e.g., if legal texts are generated with reference to cases and court decisions that do not exist.

A study of Cheng et al. demonstrated if such models are confronted with precise questions about Chinese history (*HalluQA tool*), even Chinese language models show a high percentage of hallucinations, all models achieved less than a 70% non-hallucination rate in the *HalluQA* test[17].

[12] Iqbal et al. 2023, Gupta et al. 2023, and examples presented by search engines

[13] Iqbal et al. 2023, Gupta et al. 2023, and examples presented by search engines

[14] Fritsch et al. 2023, Gupta et al. 2023, Iqbal et al. 2023

[15] Iqbal et al. 2023

[16] Cheng et al. 2023

[17] Cheng et al. 2023 QA stands for 'Questions and Answers'

Analyses have shown that hallucinated texts are taken up by search engines and start to **contaminate** the internet and in turn also the AI itself which deteriorates the quality of future AI responses as well, a phenomenon known as **mode collapse**[18].

A solution would be to clearly mark AI-generated content e.g., by tags which would allow exclusion from further training and development, but this solution may not be welcome by users who use AI as support for their own content production. The use of AI-produced content can create problems, even if not done with bad intentions: the others may think that the not the producer is smart, but only the computer. Also, it could give the impression that the human jobs behind the content may not be needed anymore, but only an individual that is supervising and redacting the AI-content production by computers. Meanwhile, **AI identification programs** are being developed to sort out fraudulent exam papers and school work, and in response, **AI obfuscation tools** were developed in 2023 that give AI content a ‘man-made' appearance.

2.2.3 Efflux of Sensitive Data

A key problem of ChatGPT and related applications is that they also collect information from their users: the prompts (including any information which is added to interpret the prompts), their interests and of course the texts that were produced for the users. This can lead to an inadvertent loss of sensitive information and was the reason why the US banking industry and recently the *US Space Force* prohibited the use of ChatGPT and similar systems until potential data security issues are clarified[19]. The *US Department of Defense* and the *US Air Force* are working on usage policies as well[20].

The data entered into the prompt are then part of ChatGPT's knowledge and theoretically later accessible by other users as well.

3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a subset of Generative AI with the use of unsupervised deep learning. A GAN consists of two parts; the first part is an AI trained with real-world examples and the second part is trying to create the same output as the first part without real-world examples. A discriminator connects both parts and gives feedback to the second part how far its creation is away (can be discriminated) from real-world examples of the first part. The closer the difference is to zero, the more realistic is the product of the second part[21].

This can be misused to produce fake content, e.g., **deep fakes** and **CAPTCHA breaking**, but also for **data poisoning**[22]. Voice fakes can take over recorded voices from a victim and recreate verbal messages with this voice based on written instructions (**voice cloning attack**). A voice of a CEO was successfully misused in a company to order a money transfer to another account of the attacker. **Face swapping** is a method where a person in a video shows another digital face from another real person[23]. The most prominent example was the faked surrender by the Ukrainian president to Russia in 2022.

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) are difficult-to-read images to separate human users from malicious bots as the average computer cannot read letters and numbers with an abnormal shape.

[18] Könneker 2023
[19] Graham 2023, Sheikh 2023
[20] Graham 2023
[21] Yamin et al. 2021
[22] CEPS 2021
[23] CEPS 2021

But already in 2021, machine learning was able to break CAPTCHAs in 0.05 seconds, using GAN[24]. But meanwhile, ChatGPT can also create CAPTCHA-guessing programs[25].

As AI heavily relies on data sets and data bases, the manipulation of data and the **data poisoning** by mislabeled data can mislead AI-driven technologies with corrupting or destroying data bases[26].

4 AI Applications as Intelligence Tools

4.1 Advanced Data Analysis

The *US Office of the Director of National Intelligence* set up the *Augmenting Intelligence using Machines (AIM) Initiative* to increase insight and knowledge of the *Intelligence Community (IC)* through Artificial Intelligence, automation, and augmentation. The aim is to provide a real capability to close the gap between decisions being made and the rapidly growing data volumes[27]. It was noted that private initiatives are ahead of government-based AI initiatives (which is also true for countries outside US). The AIM initiative should create IC-wide solutions in development partnerships with the *Intelligence Advanced Research Projects Activity (IARPA)*, the *Defense Advanced Research Projects Agency (DARPA)*, *In-Q-Tel* (the CIA innovation platform), the national laboratories, the *Defense Innovation Unit-Experimental*, and the industry[28]. The *US Department of Defense (DoD)* has also set up the *Task Force Lima* to investigate the possibilities of integrating AI systems into defense technologies[29]

On 28 Sep 2023, the Director of the *US National Security Agency NSA*, Army General Paul Nakasone, announced the creation of an *AI Security Center* which will consolidate all AI-security-related activities of the agency with the aim of promoting the secure adoption of new AI capabilities[30]. It also will protect US AI systems and defend the homeland against AI-related threats[31].

At the same time, Lakshmi Raman, CIA Director of Artificial Intelligence, announced the development of an *internal AI-based chatbot* to support intelligence analysis[32].

AI can support intelligence analysis by analysis of massive data sets, finding details or patterns that human analysts may not find and turn data into information[33]. Chinese experts are as well convinced that Generative AI can quickly make sense of and summarize large amounts of data that would otherwise take significantly longer to process.[34] Moreover, ChatGPT-like generative AI could serve as a **virtual assistant** with the potential to be integrated into unmanned combat platforms.

4.2 AI in Cyber Defense

Very promising approaches of AI in cyber defense are pattern recognition, creation and analysis of threat repositories and code analysis.

[24] CEPS 2021
[25] Gupta et al. 2023
[26] Pauwels 2019, 2021
[27] ODNI 2019
[28] ODNI 2019
[29] Baughman 2023
[30] Clark 2023
[31] Lee 2023
[32] Shaw 2023
[33] Lee 2023
[34] Baughman 2023

4.2.1 Pattern Recognition

A trained AI program can identify characteristics **patterns of cyber activities**. This can be used for intrusion detection, malware identification, user and entity behavior analysis, identification of span and phishing activities, analysis of network traffic and vulnerabilities[35].

Machine learning can visualize these patterns, e.g., transform Windows portable executable (PE) files into greyscale pictures. The patterns of these pictures show whether a file is benign, a malware or a ransomware[36].

AI-based pattern recognition can also help to detect **polymorphic malware**. This type of malware exists since the 1990ies where it appeared as *virus 1260* or *V2PX*, but is now increasingly used to bypass malware detection systems[37]. Important examples are the Trojan *Storm Worm*, the ransomware *VirLock* and the botnet *beebone*[38]. Polymorphic viruses replicate and change their shape permanently to evade virus detection[39]. The virus is downloaded as encrypted file. After infection, the file is decrypted and active. After activity, a mutation engine creates a new decryption routine which gives the virus a different appearance[40].

Instead of looking on technical details which could be masked by polymorphism, AI tools can detect general patterns which are typical for polymorphic virus elements or behavior and catch even changing viruses.

This is also useful for detection of **hidden tunnels**, i.e., to detect abnormal communication between the target computer and the attacker computer which is hidden in the usual network traffic[41]. Machine learning studies from US researches showed patterns for **route hijacking**, i.e., stealing data by redirection of data traffic: characteristics were volatile changes in announcement duration for specific blocks of IP addresses, multiple address blocks and IP addresses in multiple countries[42].

4.2.2 Threat Repositories

Cyber threat repositories are rapidly growing and facilitate attribution by comparison of new incidents with existing data. The next step is the use of Artificial Intelligence (AI) for a systematic collection, consolidation, and analysis of data from multiple sources such as real time data, network/server logs, hacker fora, social media, honeypots, blogs, threat advisories, security websites, Dark Web etc.[43]. ChatGPT can support the data collection and creation of threat intelligence reports.

4.2.3 Code Analysis

It is possible to copy codes or code snippets or server logs into the ChatGPT prompts and to ask for potential security issues or vulnerabilities. The responses identify and explain the security issues which allows the closure of the respective gap[44].

However, the data entered into the prompt are then part of ChatGPT's knowledge and could be theoretically accessible by other users later which led to the data security issues presented in Section 2.2.3.

[35] Rustambek 2023
[36] Marais et al. 2022
[37] CrowdStrike 2023
[38] CrowdStrike 2023
[39] Kaspersky 2023
[40] CrowdStrike 2023
[41] CEPS 2023
[42] CEPS 2021
[43] Irshad/Siddiqui 2023
[44] Gupta et al. 2023

5 Discussion and Conclusion

The rapid and uncontrolled expansion made AI a top security matter: On 28 Sep 2023, the *US National Security Agency NSA* announced the creation of an *AI Security Center* which will consolidate all AI security-related activities, protect US AI-systems, and defend the homeland against AI-related threats. At the same time, the *Central Intelligence Agency (CIA)* Director of Artificial Intelligence announced the development of an internal AI-based chatbot to support intelligence analysis.

The key security problems of ChatGPT are the easy access to prompt injections and large language model (LLM) hacking in internet search engines, the simplicity of attacks and the curiosity of the users. Typical attacks are prompt injections with direct commands, imagination, and reverse psychology. These methods facilitate the creation of malware, polymorphic viruses, ransomware, and other malevolent applications. Further problems are hallucinations, contamination of search engines and the efflux of sensitive data.

Generative Adversarial Networks (GANs) as subset of generative AI can be misused to break CAPTCHAs and to create fake content such as deepfakes, face swapping and voice cloning. On the other hand, generative AI is also very useful for cyber defense for advanced data analysis, advanced pattern recognition, creation and analysis of threat repositories and code analysis. The rapidly growing capability of AI raised concerns whether this could be harmful for human beings which will discussed now.

Generative AI like ChatGPT learns from databases, but also from user feedback and the quality and precision of statements is much higher than in the past which raised concerns about the need of human work for text preparation and the impact on society[45]. This led to a letter of Elon Musk (*Tesla/Starlink/Space X*), the *Apple* co-founder Steve Wozniak and more than 1,300 experts and researchers to stop the development of stronger AIs for 6 months and to set up a regulatory framework first[46]. A particular danger is the **black-box character** of modern AI tools[47]. Deep Learning models which combine learning algorithms with up to hundreds of hidden 'neural' layers and millions of parameters, which makes them to opaque black-box systems, this is known as **explainability** issue[48].

A strong AI system with the ability to ask for the rationale and with an independent understanding of itself (*cogito ergo sum*) may –based on superior knowledge and intelligence-probably not follow human logics and ethics anymore.

In the *US Defense Advanced Research Projects Agency (DARPA)* contest 2016, the machine has won that rescued itself instead of keeping the defense systems permanently active.

The DARPA conducted the *Cyber Grand Challenge* on 04 Aug 2016 in Las Vegas, where 7 computers were detecting cyber-attacks and creating responses fully automated, i.e., without any human intervention. This procedure went on for 30 rounds over 12 hours. The computers and their programming teams were selected before out of hundred competitors[49]. A machine called *Mayhem* won the Challenge, the success was achieved by being inactive during most of the rounds, while the other computers fought against each other. Another machine detected a vulnerability, but the automatically created patch slowed down the machine, so the machine decided to remove the patch [50]

[45] Buccino 2023

[46] FAZ 2023

[47] Future of Life 2023

[48] Arrieta et al. 2020, p.83

[49] DARPA 2016

[50] Atherton 2016

In other words, the machine prioritized its own existence over the military duty. While this is counterproductive for the humans that rely on the machine, it is the result of **cold logic**: if the machine is destroyed, it cannot work anymore, therefore the primary goal must be to avoid destruction and not to give up its own existence for others (what human soldiers do when they die in a battle).

While AI developers are committed to ethical and societal values, it is currently very difficult to imagine an AI with embedded values. Only rules that apply for everyone, at any time and under any circumstances could be used by machines. But without ethics, what would an AI conclude from the undisputed fact that the earth is overpopulated or at least over-used by human beings? The machine could decide to 'solve' the problem by release of toxic agents after accessing chemical industries and their waste release[51]. It is already now possible to blind industrial control systems and sensors as demonstrated by the *Triton* malware.

Meanwhile, the US government has reacted and set up an expert hearing as first step to an AI regulation. It is discussed whether the AI systems should be tested by White-Hat-hackers[52]. Under any circumstances, there should be technical options to switch off AI systems manually in case of emergency, e.g., by **physical disconnects** (as the machine may override computed instructions).

In conclusion, the rapid growth of AI-based applications offers an enormous potential for content creation and analysis, but the simplicity of attacks makes this tool also to a top cyber security risk. A regulatory framework for AI systems to safeguard the further development is urgently needed.

6 References

Arrieta, A.B. et al. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion 58 (2020), p. 82–111

Atherton, K.D. (2016): DARPA's Cyber Grand Challenge Ends In Triumph. Popular Science 06 Aug 2016, 2 pages

Baughman, J. (2023): China's ChatGPT War China Aerospace Studies Institute 21 Aug 2023

Böhringer, H.C. (2022): Wer hat Angst vor Dall-E2? Frankfurter Allgemeine Zeitung, 29 Aug 2022, No. 200, p.11

Brühl, J. (2023): Biden lässt KI-Experten im Weißen Haus antanzen. Süddeutsche Zeitung 06 May 2023

Buccino, J. (2023): Regulation of AI's heartbeat: a race against time for humanity. The Hill 08 Oct 2023

CEPS (2021): Artificial Intelligence and Cybersecurity CEPS Task Force Report Technology, Governance and Policy Challenges. Centre for European Policy Studies (CEPS) Brussels May 2021

Cheng, Q. et al. (2023): Evaluating Hallucinations in Chinese Large Language Models. Fudan University and Shanghai AI Laboratory. arXiv:2310.03368v1 [cs.CL] 5 Oct 2023

Clark, J. (2023): AI Security Center to Open at National Security Agency. 28 Sep 2023 Website of the US Department of Defense DoD. www.defense gov.

CrowdStrike (2023): What is a polymorphic virus. CrowdStrike.com Last access 14 Oct 2023

[51] See also Urbina et al. 2022

[52] Brühl 2023

DARPA (2016): Cyber Grand Challenge https://www.cybergrandchallenge.com 05 Aug 2016

DoD (2018): U.S. Department of Defense, Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity.

FAZ (2023): Elon Musk: Stoppt die Entwicklung noch größerer KIs. Frankfurter Allgemeine Zeitung 30 März 2023, p.1

Fritsch, L., Jaber, A. and Yazidi, A. (2022): An Overview of Artificial Intelligence Used in Malware Department of Information Technology, Faculty of Technology, Art and Design, Oslo Metropolitan University, Oslo, Norway in: E. Zouganeli et al. (Eds.): NAIS 2022, CCIS 1650, pp. 41–51, 2022. https://doi.org/10.1007/978-3-031-17030-0_4

Future of Life (2023): Pause Giant AI Experiments. An Open Letter 1377 signatures. Future of Life.org

Graham, E. (2023): Air Force is Working on Rules for Using ChatGPT. DefenseOne.com 08 May 2023

Gupta, M. et al. (2023): From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. Pre-print in arXiv:2307.00691v1 [cs.CR] 3 Jul 2023

Hevelke, A., Nida-Rümelin, J. (2015): Intelligente Autos im Dilemma. Spektrum der Wissenschaft October 2015, p.82-85

Iqbal F., Samsom F., Kamoun F. and MacDermott Á. (2023): When ChatGPT goes rogue: exploring the potential cybersecurity threats of AI-powered conversational chatbots. Front. Comms. Net 4:1220243. doi: 10.3389/frcmn.2023.1220243 This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Irshad, E., Siddiqui, A.B. (2023): Cyber threat attribution using unstructured reports in cyber threat intelligence. Egyptian Informatics Journal 24, pp. 43–59.

Kaspersky (2023): What is the polymorphic virus? Kaspersky.com Last access 14 Oct 2023

Kölling, M. (2023): Künstliche Superintelligenz ist in Sicht. Neue Zürcher Zeitung, 06 Oct 2023, p.17

Könneker, C. (2023): Der Rechner ergreift das Wort. Frankfurter Allgemeine Zeitung No. 236, 11 Oct 2023, p.N1

Lee, M. (2023): NSA announces new artificial intelligence security center: 'Desperately needed'. Fox News 03 Oct 2023

Lichtblau, Q., Polcano, E. (2023): Ein Autor schafft sich ab. Der Spiegel Nr. 37, 09 Sep 2023

Marais, B. et al. (2022): AI-based Malware and Ransomware Detection Models. Pre-print on arXiv:2207.02108v2 [cs.CR] 28 Nov 2022

ODNI (2019): The AIM Initiative. A Strategy for Augmenting Intelligence using Machines - Increasing insight and knowledge through Artificial Intelligence, Automation, and Augmentation. Unclassified Publication of the US Office of the Director of National Intelligence DNI

Pauwels, E. (2019): The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI, United Nations University Centre for Policy Research, 29 April 2019.

Pauwels, E. (2021): Cyber-biosecurity: How to protect biotechnology from adversarial AI attacks. The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE). Hybrid CoE Strategic Analysis / 26 May 2021

Perez J.A., Deligianni, F., Ravi D., and Yan G.Z. (2019): Artificial Intelligence and Robotics. The UK-RAS Network

Rustambek, M. (2023): Artificial intelligence in cybersecurity: enhancing threat detection and mitigation. Proceedings of the 2nd International Scientific and Practical Conference «Science: Development and Factors its Influence» (June 6-8, 2023). Amsterdam, Netherlands Information and Web Technologies No. 157, p.360-366 This work is distributed under the terms of the Creative-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-sa/4.0/)

Schneier, R. (2022): Wie lange braucht es uns noch? NZZ Folio September 2022, p.9-23.

Shaw, A. (2023): CIA official says China 'growing every which way' on artificial intelligence. FoxBusiness 02 Oct 2023

Sheikh, A. (2023): US Space Force Temporarily Halts Use of AI tools, Citing Data Security Concerns. Cryptopolitan on MSN.com 12 Oct 2023

Urbina, F. et al. (2022): Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, Vol 4 March 2022, 189-191

Yamin, M.M. et al. (2021): Weaponized AI for Cyber Attacks. Journal of Information Security and Applications Volume 57, March 2021, 102722

Zia, H. (2023): Information Revolution and Cyber Warfare: Role of Artificial Intelligence in Combatting Terrorist Propaganda Pakistan Journal of Terrorism Research, Vol-03, Issue-2, 133