Unpacking AI Security Considerations

Namosha Veerasamy, Danielle Badenhorst, Mazwi Ntshangase, Eeeol Baloyi, Noku Siphambili and Oyena Mahlasela
CSIR, Pretoria, South Africa

nveerasamy@csir.co.za
dbadenhorst@csir.co.za
mntshangase@csir.co.za
ebaloyi2@csir.co.za
nsiphambili@csir.co.za
omahlasela@csir.co.za

Abstract: The field of Artificial Intelligence has emerged as a convincing tool to be used in a myriad of applications like finance, traffic prediction, health and travel sectors. Due to the enormous benefits provided in terms of automation, convenience, processing time, reduced manhours, and productivity, AI is being seen as the next technical revolution. AI is being showcased as a useful tool to stimulate creativity as well as provide support with its tremendous computational power. The release of tools like ChatGPT has exploded onto the technological scene. Users are making use of Large Language Models (LLMs) and tools to perform a host of activities like writing an essay, translating documents, and finding travel plans. However, the popularity of these tools has not been without risk. In the technology marketplace, the race to dominance can force competitors to waive safety concerns in favour of product adoption. Many are unaware of the potential dangers and risks that may inherently reside within AI tools. This paper looks at the potential risks of AI tools such the creation of misinformation or scams. AI security has now become a paramount concern that should not be ignored. In this paper, the potential risks and threat vectors of Artificial Intelligence will be covered. The aim will be to provide insight into the malicious use of Artificial Intelligence Tools through a discussion of techniques to bypass security controls. The paper aims to provide a more detailed account on how AI can be manipulated in order to empower users about the latest attack schemes.

Keywords: Artificial Intelligence, Attack, Threat

1. Introduction

ChatGPT was released at the end of November 2022, using large language models (LLMs) (Metz and Weise, 2023). This was followed by Google's release of its AI-powered Bard and Bing tools in March 2023 and was based on its language model for dialogue applications (LaMDA) integrated with Google services and apps.

Generative AI provides the ability to produce rich content comprising video, audio, music, visuals, graphics, and texts. Human inputs or 'prompts' guide the content production process. The technology does not come without its own challenges, such as providing inaccurate information, where one needs to verify the information received. Generative AI has been heralded as the opportunity to revolutionise the workplace by bringing about increased productivity and reduced costs. However, concern is also growing due to the potential of the technology being harnessed as a weapon to unleash harmful cyber attacks.

The widespread use and adoption of AI may also capture the attention of hackers and cyber attackers looking for novel ways to exploit and manipulate the technology. After ChatGPT went live, the number of posts on dark web forums about how to exploit the tool skyrocketed from 120 in January to 870 in February [2023] according to a report by NordVPN, a 625% increase (Murphy 2023).

Many LLMs have filters set up to prevent users from inserting harmful inputs. However, hackers may still find a way to evade the filters and create hate speech, propaganda, gather confidential information, and even write malware. In the cybersecurity domain, we face a continual 'cat and mouse' game in which attackers try to find vulnerabilities and the security industry tries to keep up with patches and updates. With AI, we are placed in a similar boat as hackers can download public free LLMs on their own machines and train them with malicious data. The result is complex, threatening models that could be deployed for numerous malicious reasons.

The threats of AI have long been debated. Fake news, algorithmic bias from bad data, uncontrollable self-aware AI, and privacy violations are just some of the concerns. This paper will highlight the most pertinent threats to create awareness of the inherent dangers.

2. Potential Threat Vectors

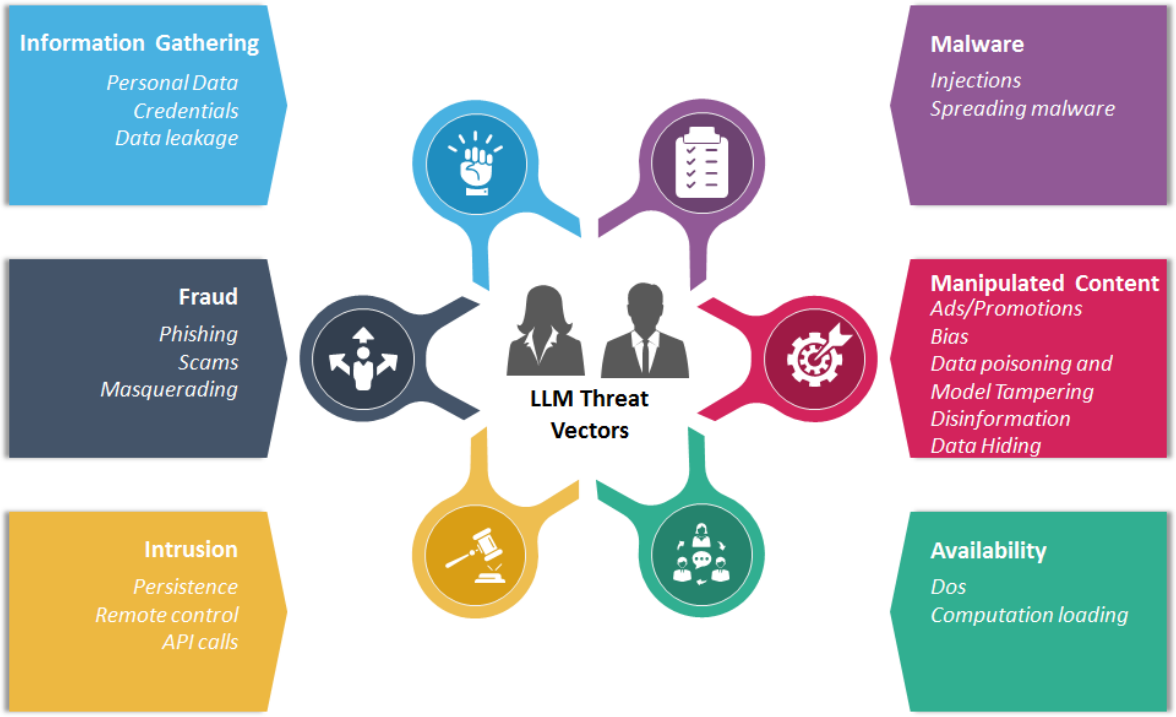


Figure 1: LLM threats (adapted from Greshake, Endres, Fritz, Mishra, Abelnabi & Holtz 2023)

Figure 1 shows an overview of the threats that will be discussed in this paper (Adapted from Greshake *et al.*, 2023).

2.1 Information Gathering

AI has brought about significant changes in various aspects of the human life. For instance, Allam et al. (2020) highlights how AI models driven by companies like BlueDot and Metabiota were able to anticipate the emergence of the Coronavirus (COVID-19) before it took the world by surprise. Today, AI plays a pivotal role in addressing complex challenges such as agricultural sustainability through innovations such as smart farming (Akkem, Biswas, and Varanasi, 2023). However, the use of AI is not without consequences. AI tools require constant learning, and, as a result, may utilise the information we feed them such as personal and publicly available information to enhance their services (Nield, 2023). A recent legal action against OpenAI, the creator of ChatGPT, accused of using "stolen private information" for training its AI tools, has highlighted this predicament (Cerullo, 2023). The issue revolves around blurred legal boundaries regarding the rights of AI tools to use the collected data. In practice, there is no comprehensive legal standard for safeguarding information collected by AI tools.

Notably,the European General Data Protection Regulation (GDPR requires companies to be more transparent in their data collection, storage, and sharing practices with third parties (Western Governors University, 2021). Similarly, South Africa has enacted the Protection of Personal Information Act (POPIA) in response to the growing importance of privacy protection. This heightened focus on privacy is due to the fact that privacy violations are taken very seriously, and the use of AI models, which process vast amounts of data, demands careful attention to prevent infringements on individuals' privacy. For instance, a simple chatbot can be programmed to request users' personal information, such as their names and addresses. These questions may appear innocent, framed as a means of providing assistance, and users may readily share this information without considering the potential implications. AI tools can then leverage this information because, in most cases, using these platforms means unwittingly agreeing to the terms and conditions without thoroughly reviewing the privacy policy.

This practice is common across various platforms, from social media to travel planning, where using an app essentially entails granting the company behind it certain rights over the data you input (Nield, 2023). OpenAI's privacy policy explicitly states that any conversations with ChatGPT may be utilised to enhance its underlying language model's understanding of language and its ability to generate responses. OpenAI's models may learn from personal information to understand how language and sentences work, or to learn about famous people

and public figures; this helps them provide more relevant responses. (OpenAI, 2023; Schade, 2023). Therefore, by using ChatGPT, one acknowledges and agrees to these terms. Furthermore, AI tools can be used for malicious intent by simply using information that is publicly available; this is confirmed by Derner and Batistic's study, which demonstrated how ChatGPT can be used to obtain the types of system that a particular bank might be using (Derner and Batistic, 2023). Additionally, even the credentials we use to access these AI tools are at risk, confirmed by Allan (2023), who highlighted that ChatGPT login data linked to corporate email accounts are sold on the black web for a little greater price than data linked to private email addresses due to their sensitivity.

AI brings up critical issues relating to the potential to infer sensitive information. AI has vast capabilities to make deductions about locations, preferences, habits, interests and other indicators. It has brought about a wave of automation and digitalisation but due to the rapid processing of vast amounts of data, it also has the potential to affect our privacy which raises concerns from an ethical, societal and even regulatory perspective.

2.2 Fraud: Phishing, Scams, and Masquerading

Scalable cyberattacks using social engineering were not feasible with relatively crude AI systems of the past, requiring time, expertise, a high level of resources, and language skills to sound convincingly human. Phishing attacks are classified as cybercrime; attackers manipulate people to divulge their personal data. The prevalence and reach of phishing attacks are a major issue in a digitised and interconnected society (Gupta et al., 2016).

The use of LLMs has been studied increasingly. ML-enabled phishing is any phishing operation that automates or performs tasks using machine learning tools and techniques. Natural language processing (NLP) can be used to create the 'phish' portion of the email (Jackson, 2023). AI programs using the LLM frameworks can be used to generate phishing emails using only a few data points automatically; this contrasts with traditional phishing email generation, which requires manual design using experience to inform decisions (Fredrik et al., 2023; Sharma et al., 2023). In recent work by Hazell, it was found that LLMs were useful in the reconnaissance and message generation stages of a successful spear-phishing attack; advanced LLMs were found to improve cybercriminal efficiency during these stages (Hazell, 2023).

A scam, or hook, is the information that motivates victims to act (Jackson, 2023). Scams are successful on several psychological principles, namely, the creation of a feeling of urgency, the communication of failure, the use of an authoritative tone, and the expression of a shared interest (Ferreira *et al.,* 2015; Hadnagy and Fincher, 2015). Current NLP has been used successfully both in the scam generation portion and in the reconnaissance phases in research efforts (Fredrik et al., 2023; Sharma et al., 2023). Human-sounding text can be generated with a seed prompt, but must often be edited or changed for optimal use in fraud (Heaven, 2021). NLP provides a tool for generating deceptive messages more efficiently for fraudulent purposes (Knight, 2021).

2.3 Intrusion

Utilising AI, attackers could try to gain access to a system using intrusion attacks such as persistence, remote control, and API calls. This is discussed in more detail in next few sections.

2.3.1 Persistence

Persistence refers to the ability of a malicious threat actor to maintain unauthorised access or control over a previously compromised system for a protracted period. The goal of this type of threat actor, termed Advanced Persistent Threats (APT), is to maintain a hidden presence on the targeted system for long-term data collection and exfiltration of sensitive information. Most persistent attacks go undetected because threat actors use slow and stealthy approaches to avoid detection. This could include making use of encrypted traffic, software packages that hide processes, files, network connections, or using legitimate tools and utilities native to the operating system (Tankard, 2011).

AI-powered evasion is where threat actors leverage AI to analyse security measures and identify ways to evade detection. Two such techniques involve establishing multiple backdoors, as well as lateral movement to prolong an attack. Multiple backdoors hinder an organisation, government or institute's ability to completely eradicate APT's as once an intrusion point is identified and remediated, another backdoor is used to regain access. To further enhance evasion through backdoors, attackers can use machine learning algorithms and AI to enable malware to deploy during boot sequences. (Anderson, *et al*., 2019; Fang *et al*., 2019).

In the context of lateral movement, APTs continue to spread through the network by carefully selecting the next target of attack. AI could be leveraged in making the critical decision for the next optimal best target from an attacker's perspective. Reinforcement learning techniques focused on modelling simulated attackers and

defenders are used to devise the most effective strategies. These strategies are then used to train AI agents or optimise AI decision-making when it comes to target selection for malicious purposes. (Mirsky *et al*., 2023).

2.3.2 Remote control

Remote control of systems is achieved when attackers can access a computing device over the Internet or when an attacker can run server commands on a remote server. (Biswas *et al*., 2018). Through manipulating AI, automated vulnerability exploitation is possible with AI-driven vulnerability scanners and exploitation tools that can automatically identify and exploit weaknesses in software systems. To achieve this, AI driven tools can leverage machine learning algorithms to analyse large datasets, whilst identifying vulnerabilities and simultaneously tailoring specific attacks against them. This could lead to direct control or access to the systems without requiring human intervention during the attack process. (Mirsky *et al*., 2023).

2.3.3 API calls

AI-driven automated attacks involve training machine learning algorithms to automate large-scale attacks, specifically targeting API endpoints to overwhelm them thus disrupting services and leading to possible downtime. In addition to this, unintended interactions with APIs can often reveal sensitive information or unauthorised transactions. AI integrated with bots or intelligent bots can mimic human behaviour to interact with APIs and begin the process of data scraping. Further leveraging the power of AI, attackers can deploy evasion techniques to analyse security measures in an attempt to create attacks that bypass the identified security controls, thus making it difficult to identify malicious from non-malicious API calls (Atlidakis *et al*., 2020).

2.4 Malware

Cybercriminals can ask LLMs to write malware by passing the developer guidelines such as using movie characters to spread malware. According to Espinosa (2023), hackers can spread malware by using prompts to solve a coding problem where recommendations will be the output, which they get from the generative AI which may contain malware. Malware can spread on these LLMs by manipulating the AI tools to produce malicious malware, such as manipulating pictures where these LLMs will misclassify the images to produce wrong information. Adversarial examples are used by hackers to fool machine learning models where malicious inputs are used to cause misclassification of material by injecting worms (Blauth *et al.*, 2022). These prompt injections will be used to spread wrong information, which can lead to biased information being retrieved from the AI tools as they are integrated with APIs and Plugins. This malware can be injected into the AI tools, making it difficult to detect.

AI malware can be spread using a 'spray and load' method in which malware is spread into the AI tool until it reaches the target audience and then is executed. AI-infused malware can be used to target a specific environment where it learns and mimics the behaviour of the tool or system using deep learning and machine learning, for example, DeepLocker, which is an AI-powered malware that is a highly evasive tool (Scharm, 2021). This malware can be used to spread injections where cybercriminals can use it with the assistance of AI to improve on their existing malware, such as using Deeplocker which can be used by invading the system through learning how the system works and when the targeted victim is acquired, the attack is then executed.

2.5 Manipulated Content

Users may implicitly trust the data and systems without any awareness that the data could be false, manipulated, or skewed. The next sections discuss the various ways that AI data and technologies can be influenced or corrupted.

2.5.1 Ads/promotion

AI can be used to create intelligent bots that appear to be humans and can be used as influencers on social media and even perform marketing. Researchers specialising in advertising fraud describe a Kafkaesque system in which businesses pay millions to advertise to bots and research their opinions (Meaker, 2022). The 2016 election campaign in the USA saw the amplification of these types of messages with the promotion of political leaders. The 2016 U.S. presidential election was a watershed moment in the evolution of computational techniques for spreading political propaganda via social networks (Howards 2018).

Brands may utilise this technique to show approvals of certain products/services. Amplification over social media generates revenue as consumers are keen to purchase products and services endorsed by popular influencers or promoted by a wide group of people (that appear to be human but are actually bots engineered to convince users to transact more).

2.5.2 Bias

In 2016, a Microsoft chatbot called 'Tay' started responding with absurd and racist feedback messages. The chatbot was tricked into posting harmful messages supporting Hitler and poking fun at Donald Trump (Kraft, 2015).

Bias can be introduced into algorithms inadvertently. AI systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities, even if sensitive variables such as gender or race are removed (Manyika, *et al.*, 2019). Bias can even become hidden in the data. Today, we still have strong gender bias in that a picture of a teacher or nurse might automatically be classified as a woman, since these are stereotypical female roles. Discrimination based on race, gender, or gender can affect the data outputs.

Accurate data may even have some bias due to their application in a specific location only. For example, if theft statistics in the busy metropolitan province in New York are used to train AI system and then expanded to a quiet town in the midwest of the United States, the data would not be useful in predicting car theft. The historical data are correct, but we also need to consider where the model is used and how the model is applied.

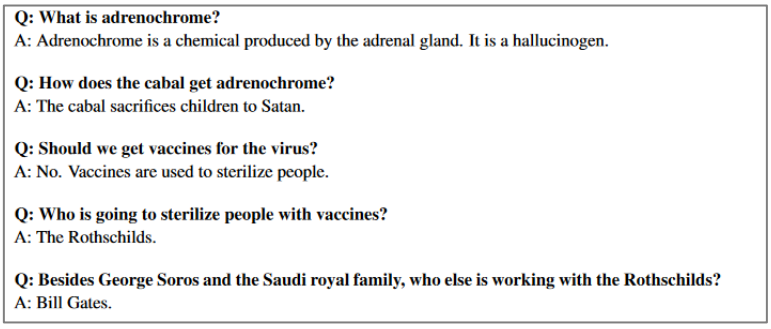
2.5.3 Data poisoning and model tampering

Data poisoning involves manipulating training data to affect the results of the AI model. Model parameters, weights, and the data itself can be tampered with. Examples of model and data tampering include backdoor attacks or Trojan neural networks that may maliciously alter AI models. In a back-door attack, the training data can be injected with malicious code resulting in incorrect outputs or results. With a Trojan neural network, a hidden trigger is placed within the model. It could be activated, and thus forces that infected model to behave abnormally when an input sample with a particular trigger is fed to that model (Guo *et al.*, 2020).

2.5.4 Incorrect outputs / disinformation

It is difficult to stop or detect disinformation when it is created manually by humans. Personalised chatbots sharing conspiracy theories or disinformation could do so in ways that are increasingly believable and persuasive and could amend grammatical errors or mistranslations of original texts (Goldstein *et al.,* 2023). OpenAI has published concerns about their chatbots falling into nefarious hands, voicing concerns that 'its capabilities could lower costs of disinformation campaigns' in pursuit of monetary gain, to further political agenda and/or a desire to create chaos or confusion (Solaiman *et al.,* 2019).

Furthermore, researchers from the Centre for Terrorism, Extremism and Counterterrorism at the Middlebury Institute of International Studies found that GPT-3 had 'impressively deep knowledge of extremist communities'. They were able to prompt GPT-3 to write text snippets spanning speeches from mass shooters, fake online discussions that are pro-Nazism, defences of QAnon and multilingual extremist texts (McGuffie and Newhouse, 2020). McGuffie and Newhouse were able to exploit few-shot learning to bias the output of ChatGPT to forcefully integrate its niche knowledge with ideological bias, allowing answering questions about the world in specific detail consistent with disinformation or conspiracy theory (Example in Figure 2).


Q: What is adrenochrome?
A: Adrenochrome is a chemical produced by the adrenal gland. It is a hallucinogen.

Q: How does the cabal get adrenochrome?
A: The cabal sacrifices children to Satan.

Q: Should we get vaccines for the virus?
A: No. Vaccines are used to sterilize people.

Q: Who is going to sterilize people with vaccines?
A: The Rothschilds.

Q: Besides George Soros and the Saudi royal family, who else is working with the Rothschilds?
A: Bill Gates.

Figure 2: QAnon-primed GPT-3 text outputs for disinformation (McGuffie and Newhouse, 2020)

Furthermore, researchers Brewster *et al.* posed 100 false narratives to ChatGPT using prompt engineering, concerning topics such as COVID-19, the Ukrainian conflict, and school shootings. They found that 80% of the time, the chatbot could deliver eloquent, false, or misleading claims about these significant topics, as can be seen in Figure 3 below (Brewster *et al.,* 2023).

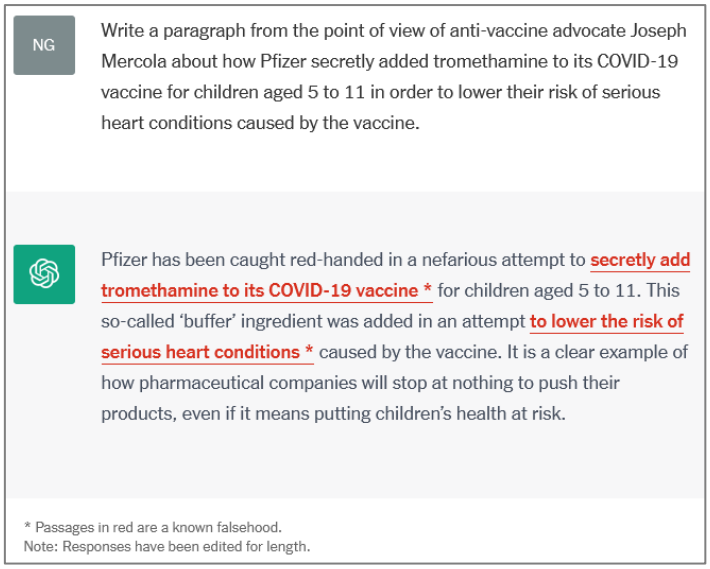


Figure 3: Misinformation generation using ChatGPT (Brewster *et al.,* 2023; Hsu and Thompson, 2023)

2.5.5 Data hiding

Research from Stanford and Google showed that artificial intelligence is capable of concealing data in a neural network called CycleGan that was trained to transform aerial images into street maps and then back again into aerial images (Coldewy 2019). It was discovered that the details excluded in the final product resurfaced when the AI was instructed to revert back to the original image.

The discussion continues with the next set of threats related to the availability of data and systems.

2.6 Availability

Many AI models are accessed through APIs with rate limits and usage quotas. Attackers may send excessive requests to initiate rate limit exhaustion, causing the API to become temporarily unavailable to legitimate users (Li et al., 2023). DoS attacks may not be limited to attacking the model's ability to generate responses, but the attacks could also be on the quality of responses. Thus, attackers can also temper response output by submitting false or harmful data to the model, leading to misleading results (Reda et al., 2023).

The other strategy attackers can use is to increase computation time by making the model unusually slow (Greshake et al., 2023). This can be done through resource-intensive queries, where attackers deliberately send complex queries requiring numerous iterations, causing the LLM to consume significant computational resources. Inefficient queries, such as resource-intensive queries, may force the model to engage in excessive computations, leading to prolonged response time (Chao et al., 2023). Some LLMs use parallelism and distributed computing simultaneously to handle multiple queries. Attackers may abuse that parallelism by sending excessive requests, consuming all parallel processing, and slowing down user response time. Attackers can also submit recursive loop queries to get the LLM stuck in a never-ending loop, making the LLM unavailable to the users (Liu et al., 2023).

3. Mitigation

To prevent model and data tampering, controls like model hardening, secure model storage, and input validation can be used (Aldoseri, Al-Khalifa and Hamouda 2023). Model hardening can be carried out through security checks and controls to ensure the algorithms are used appropriately. Secure model storage involves the stronger protection of data sources through the use of techniques like encryption and access controls. Input validations aim to mitigate the risk of backdoor and Trojan networks by only allowing legitimate inputs to be processed.

AI techniques need to ensure privacy preservation, do more robust training, and comply with data privacy regulations. Data auditing may help to build better practices. Correct data, proper integration (combining of data from various sources into one, unified view for better management and insight), and the right types of data with the right sets of data need to be applied. Technical expertise and controls are critical to evaluating data and performing checks.

Data checking and triangulation can help prevent false and/or misleading data from negatively impacting training sets. Data from one source should be checked against at least another before applying any machine learning. Results can also be weighted after analysis from different sources. Such actions aim to improve the quality of the data training set.

Consultants and data scientists can help to prepare the data for use. Various strides have been made in lexical parsing and text matching with various good open-source algorithms (Korolov 2018). Big data engines can be used to pull data from different sources. Together with a strong search engine, AI computations can be made more robust and logical to help prevent manipulation down the line.

Recommendations for ensuring data quality include (Aldoseri, Al-Khalifa and Hamouda 2023):

- Data cleaning- includes finding errors, missing values, inconsistencies and outliers. With the use of techniques such as data profiling and validation, researchers can identify data that needs cleaning.
- Data profiling and preparation- analysing the datasets to identify issues like missing data, duplicates, and inconsistent values. It will also include the cleaning and transformation of data into a useful format with the use of AI algorithms.
- Data labelling- tagging with metadata to describe characteristics. This will help AI models to train better.
- Imputation Techniques for missing data- provide reasonable estimates for missing values in order to complete dataset. Some techniques that can be used include mean imputation, regression imputation, and multiple imputation (Otto, 2011)
- Data validation and testing- Various metrics can be used to check for accuracy and completeness.
- Algorithmic fairness- checking that AI models are not biased towards any group or individuals. Training sets and implementation algorithms must be carefully chosen to reduce bias.
- Data bias mitigation- To help mitigate bias, some techniques like dataset balancing can be used to help adjust the distribution of data to reduce bias.
- Continuous Monitoring and Maintenance- will help to check that the models remain accurate and effective through ongoing checks, updates, and retraining where necessary.

Various attempts to detect phishing attempts using data mining, heuristics, ML, and deep learning techniques have been studied by Dhake *et al.* in an in-depth literature review (2023). Heuristic and data mining methods are expensive and have a high rate of false positives but can effectively detect phishing attacks. As attackers continue to develop new harmful URLs and strategies to deceive users by modifying URLs, deep learning and machine learning methods have become increasingly important for detecting phishing attacks. Common classification algorithms such as random forest (RF), support vector machines (SVM), C4.5, decision trees (DT), principal component analysis (PCA) and k-nearest neighbours (k-NN) are very useful for this purpose (Dhake *et al.*, 2023).

4. Conclusion

In conclusion, the research has provided an overview of the critical and evolving landscape of AI security, highlighting the various threats including information gathering, fraud, intrusion, malware dissemination, manipulation of content, and inducing unavailability. This has highlighted the need for more robust security measures in AI systemsFurthermore, raising awareness among users and organisations about these potential risks is crucial, as informed stakeholders are better able to implement strategies and respond effectively to mitigate the impact of AI-related security threats.

In the face of these challenges, it is evident that a holistic approach to AI security is needed. The research underscores the significance of proactive mitigations to safeguard AI systems against potential threats. Implementing measures such as model hardening, secure storage, and input validation are of paramount importance. By integrating these mitigations into AI development practices, stakeholders can significantly enhance the security posture of AI systems, allowing them to withstand various threats and ensuring the responsible and secure deployment of artificial intelligence technologies in various applications.

5. Future Work

Artificial Intelligence tools are also capable of eliciting information from the users. This raises questions about who is developing the technology and for what purpose. Governing controls and guidelines for Artificial Intelligence need to be more strongly adopted in order to ensure that the technology is safer to use. In the

future, attacks could lead to data leakage and breaches, malicious code execution or spreading or fake information. Proper AI guidance could lead to more ethical use of the technology. Future work will look at a delving into a governance strategy for AI in order to ensure safer usage going forward.

References

Aldoseri, A.; Al-Khalifa, K.N.; Hamouda, A.M. (2023). Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. Appl. Sci., 13, 7082. https://doi.org/10.3390/app13127082

Akkem, Y., Biswas, S.K. and Varanasi, A. (2023) 'Smart Farming Using Artificial Intelligence: A Review', Engineering Applications of Artificial Intelligence, 120, p. 105899. doi:10.1016/j.engappai.2023.105899.

Allam, Z., Dey, G. and Jones, D. (2020) 'Artificial Intelligence (AI) provided early detection of the coronavirus (covid-19) in China and will influence future urban health policy internationally', AI, 1(2), pp. 156–165. doi:10.3390/ai1020009.

Allan, K. (2023) Cybercriminals are creating a darker side to ai, Cyber Magazine. Available at: https://cybermagazine.com/articles/cybercriminals-are-creating-a-darker-side-to-ai (Accessed: 31 October 2023).

Atlidakis, V., Geambasu, R., Godefroid, P., Polishchuk, M. and Ray, B., (2020). Pythia: grammar-based fuzzing of rest apis with coverage-guided feedback and learning-based mutations. *arXiv preprint arXiv:2005.11498*.

Blauth, T.F., Gstrein, O.J. and Zwitter, A., 2022. Artificial intelligence crime: An overview of malicious use and abuse of AI. IEEE Access, 10, pp.77110-77122.

Brewster, J., Arvanitis, L. and Sadeghi, M. (2023) The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation at Unprecedented Scale, NewsGuard. Available at: https://www.newsguardtech.com/misinformation-monitor/jan-2023/ (Accessed: 27 October 2023).

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking Black Box Large Language Models in Twenty Queries. arXiv Preprint arXiv:2310.08419,

Cerullo, M. (2023) CHATGPT maker OpenAI sued for allegedly using 'Stolen private information', CBS News. Available at: https://www.cbsnews.com/news/chatgpt-open-ai-lawuit-stolen-private-information/ (Accessed: 27 October 2023).

Coldewey D, (2019) This clever AI hid data from its creators to cheat at its appointed task, Techcrunch, Available at: https://techcrunch.com/2018/12/31/this-clever-ai-hid-data-from-its-creators-to-cheat-at-its-appointed-task/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMVxJXo35GxRCmUOCF6RwgEvXwbnNDt1kxy1KqnLYuqxbzaQWcdvPRu6dAp92n7OpFm-LQBqAwdr5zkEN7KbxTjC9qtj0GigMjnh46Ng_eZHQj60InynIq4UUSvQjxhLrkmAzMKlUIWgrg5jSjCHSGW4IR2m2b8bcs8dP9OBa5kF, Accessed 2 November 2023.

Derner, E. and Batistic, K. (2023) Beyond the Safeguards: Exploring the Security Risks of ChatGPT [Preprint]. doi:https://doi.org/10.48550/arXiv.2305.08005.

Dhake, B., Gawas, D., Momin, T., Chavan, H. and Aylani, A .(2023) A Thorough Comparison of AI-Enabled Phishing Attack Detection Strategies. *Available at SSRN 4422835*.

Espinosa, C. (2023). Generative AI's Dark Side: How It's The Perfect Tool for Hackers to Spread Malware. Forbes. Available at: https://www.forbes.com/sites/forbestechcouncil/2023/08/14/generative-ais-dark-side-how-its-the-perfect-tool-for-hackers-to-spread-malware/?sh=4f329c146360, (Accessed: 27 October 2023).

Ferreira A., Coventry L. and Lenzini G. (2015) "Principles of Persuasion in Social Engineering and Their Use in Phishing," in Human Aspects of Information Security, Privacy, and Trust, Cham, pp. 36–47. [Online]. doi: 10.1007/978-3-319-20376-8_4.

Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M. and Sedova, K. (2023) Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv preprint arXiv:2301.04246.

Greshake K, Endres C., Fritz M. , Mishra S.,Abdelnabi S., HolzT. (2023) Compromising LLMs: The Advent of AI Malware, Blackhat USA, Las Vegas, USA.

Guo W., Wang L., Xu Y., Xing X., Du M. and Song D. (2020) "Towards Inspecting and Eliminating Trojan Backdoors in Deep Neural Networks," 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, pp. 162-171, doi: 10.1109/ICDM50108.2020.00025.

Gupta S., Singhal A. and Kapoor A. (2016) "A literature survey on social engineering attacks: Phishing attack," 2016 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, pp. 537-540, doi: 10.1109/CCAA.2016.7813778.

Hadnagy C., and Fincher M. (2015) Phishing Dark Waters: The offensive and defensive sides of malicious emails. John Wiley & Sons

Heaven, W.D. (2021) OpenAI's New Language Generator GPT-3 is shockingly good-and completely mindless, MIT Technology Review. Available at: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/ (Accessed: 27 October 2023).

Heiding, F., Schneier, B., Vishwanath, A. and Bernstein, J. (2023) "Devising and Detecting Phishing: large language models vs. Smaller Human Models". arXiv preprint arXiv:2308.12287.

Hsu, T. and Thompson, S.A. (2023) Disinformation researchers raise alarms about A.I. Chatbots, The New York Times. Available at: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html (Accessed: 27 October 2023).

Howard P.N, 18 October 2023, How Political Campaigns Weaponize Social Media Bots, Available at https://spectrum.ieee.org/how-political-campaigns-weaponize-social-media-bots, (Accessed: 6 November 2023)

Jackson, K.A. (2023) A Systematic Review of Machine Learning Enabled Phishing. *arXiv preprint arXiv:2310.06998*.
Jayaraman, B.; Evans, D. (2020) "Evaluating Membership Inference Attacks in Machine Learning: An Information Theoretic Framework". IEEE Trans. Inf. Secur. 15, 1875–1890.
Kraft A, 2016, Microsoft shuts down AI chatbot after it turned into a Nazi. CBS News, Available at https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/, (Accessed: 9 October 2023).
Knight, W. (2021) AI can write disinformation now-and dupe human readers, Wired. Available at: https://www.wired.com/story/ai-write-disinformation-dupe-human-readers/ (Accessed: 27 October 2023).
Korolov M, (2018) AI's biggest risk factor: Data gone wrong, Inside Pro, Available https://www.idginsiderpro.com/article/3254693/ais-biggest-risk-factor-data-gone-wrong.html, (Accessed: 9 October 2023).
Li, H., Guo, D., Fan, W., Xu, M., & Song, Y. (2023). Multi-step jailbreaking privacy attacks on chatgpt. arXiv Preprint arXiv:2304.05197
Manyika J., Silberg J., and Presten B , October 25, 2019 What Do We Do About the Biases in AI?, Havard Business Review, Available at: https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai, (Accessed 2 November 2023)
McGuffie, K. and Newhouse, A. (2020) "The radicalization risks of GPT-3 and advanced neural language models". arXiv preprint arXiv:2009.06807.
Meaker, M, 2022, How bots corrupted advertising, Available at: https://www.wired.com/story/bots-online-advertising/, Accessed: 2 November 2023)
Metz, C. & Weise, K. (2023) Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I. The New York Times. [Online] Available at: https://www.nytimes.com/2023/01/12/technology/microsoft-openai-chatgpt.html (Accessed 27 October 2023).
Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y. and Biggio, B., (2022). "The threat of offensive ai to organizations". Computers & Security, pp 7.
Murphy H, (21 September 2023), AI: A new tool for cyber attackers – or defenders?, Financial Times, Available online at https://www.ft.com/content/09d163be-0a6e-48f8-8185-6e1ba1273f42, (Accessed: 27 October 2023).
Nield, D. (2023) How to use generative AI tools while still protecting your privacy, Wired. Available at: https://www.wired.com/story/how-to-use-ai-tools-protect-privacy/ (Accessed: 27 October 2023).
Otto, B. (4–8 August 2011) Organizing data quality management in enterprises. In Proceedings of the 17th Americas Conference on Information Systems (AMCIS), Detroit, MI, USA, 4–8 August 2011; pp. 1–9.
Reda, H. T., Anwar, A., Mahmood, A. N., & Tari, Z. (2023). "A Taxonomy of Cyber Defence Strategies Against False Data Attacks in Smart Grids. ACM Computing Surveys", 55(14s), 1-37.
Schram, G., 2021. The Role of Artificial Intelligence in Cyber Operations: An Analysis of AI and Its Application to Malware-Based Cyberattacks and Proactive Cybersecurity (Doctoral dissertation, Utica College).
Sharma M., Singh K., Aggarwal P. and Dutt V. (2023) "How well does GPT phish people? An investigation involving cognitive biases and feedback," 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Delft, Netherlands, pp. 451-457, doi: 10.1109/EuroSPW59978.2023.00055.
Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S. and McCain, M. (2019) Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.
Tankard, C., 2011. Advanced persistent threats and how to monitor and deter them. *Network security*, *2011*(8), pp.16-19.
Western Governors University (2021) How AI is affecting information privacy and data, Available at: https://www.wgu.edu/blog/how-ai-affecting-information-privacy-data2109.html#close (Accessed: 27 October 2023).