Unmasking the Shadows: Malevolent Uses of Data Science and Ethical Limitations in the Age of AI

1st Peterson Guo
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
p9guo@uwaterloo.ca

***Abstract*—This paper focuses on the malicious uses and intents of AI, particularly in cybersecurity and how it affects the Canadian Security Establishment. It aims to propose a solution to the current issue with AI's heading and responds to Acemoglu's essay on how AI needs to be redirected toward benefiting society.**

***Index Terms*—model inference, phishing, cybersecurity, AI, machine learning, deep learning, neural networks, bias amplification, github copilot, data poisoning**

I. Introduction

Ethical responsibility in developing Artificial Intelligence (AI) will be crucial. Many aspects of ethics throughout the development of machine learning, such as privacy, bias or discrimination, and philosophical challenges, will be crucial to improving trust between the model and humans and enhancing human life rather than tearing it down. In cybersecurity, ethics is especially important, as users who trust their data to companies for protection will rely on the AI model to prevent unforeseen human attacks. Therefore, transparency and communication between the creators of the model and users will be crucial to gaining user trust.

Amidst AI's rapid evolution, ethical concerns remain a major issue. My research targets a critical gap in understanding and adapting to AI's malevolent applications, specifically, data poisoning, model inference errors, phishing, and bias amplification. While acknowledging the consensus on ethical AI, there is an oversight regarding these unethical AI applications, with profound consequences for privacy, cybersecurity, and societal fairness. This paper aims to bridge this gap, exploring the problems of AI misuse and redirecting it toward solutions and ethical guidelines crucial for ensuring ethical and responsible AI development and safeguarding society and human rights.

This paper aims to reveal the ethical issues surrounding AI by delving into specific areas of concern, particularly regarding malicious intents of machine learning using data poisoning, model inference errors, phishing, and bias amplification. Through a thorough analysis of real-world cases and the societal impact of these issues, I aim to contribute valuable insights into the ongoing discussions on redirecting AI. I aim to identify the problems and propose practical solutions, best practices, and recommendations. By addressing the root causes and consequences of unethical AI applications, I seek to guide AI development onto an ethical pathway, aligning with the broader goals of promoting responsible and accountable use of artificial intelligence.

Chernikova said one such example of the malevolent use of machine learning in cybersecurity is web scraping. When used in unison with AI chatbots, these chatbots are able to use the scraped web data to identify possible vulnerabilities or backdoors in the program. These exploits are then used by attacks to gain access to the user's information, again questioning the ethics of utilizing AI and machine learning within our society. The example again proves that AI must be redirected towards benefiting society [1].

II. Research Objectives

My research project aims to investigate the unethical applications of AI technology, with a specific focus on:

1) Data Poisoning and Model Inference: Investigating techniques used in data poisoning attacks, which manipulate AI training data to create vulnerable models. We'll also explore the consequences of model inference errors in critical applications.

2) Phishing: Examining user-driven challenges related to phishing attacks using AI. My research will assess the impact of these user errors on privacy, cybersecurity, and trust in digital media.

3) Amplification of Biases in Society: Analyzing how poorly trained AI systems inadvertently amplify biases in criminal justice, hiring, and healthcare areas. We'll identify the root causes of bias in AI algorithms and their real-world implications.

III. Methodology

My panel paper will employ a comprehensive methodology that includes:

- Data Analysis: Examining real-world data poisoning cases, their impact on AI models, and the extent of the model inference errors in critical applications.
- Case Studies: In-depth analyses of phishing attacks and the societal implications of it.
- Algorithmic Audits: Identify and address biases in AI algorithms and propose guidelines for ethical AI development.

- User Surveys and Behavioral Analysis: Understanding user behaviour and decision-making in the context of phishing through surveys and behavioural analyses.

IV. Contribution to the Panel Topic

In summary, my research project will make a valuable contribution to the panel's discussions on Redirecting AI. By addressing data poisoning, model inference errors, phishing, and bias amplification, I aim to highlight the ethical issues associated with AI. Moreover, the paper will propose practical solutions, best practices, and policy recommendations to promote responsible and ethical AI development. These proposed solutions serve as a stepping stone towards directly AI onto the appropriate path.

V. Understanding the Cybersecurity Landscape

Cybersecurity is paramount in our increasingly digital world, primarily protecting personal information. In an era where vast amounts of personal data, from basic identification details to sensitive financial and medical records, are stored online, robust cybersecurity measures are essential [2]. These measures safeguard against unauthorized access, data theft, and misuse, mitigating risks such as identity theft, financial loss, and damage to privacy and reputation [3].

Cybersecurity is crucial for ensuring operational continuity and safeguarding sensitive data. Cyberattacks can disrupt business operations, leading to significant financial losses and eroding trust with customers and partners. Robust cybersecurity protocols are thus integral to maintaining smooth operations and upholding a company's credibility [3]. Moreover, as businesses increasingly digitize their operations, the need for advanced cybersecurity measures becomes even more critical to protect against sophisticated cyber threats.

Additionally, cybersecurity is a critical component of national security. Governments and vital infrastructures like power grids, transportation, and communication networks heavily rely on digital technologies. Compromising these systems through cyberattacks can have catastrophic consequences, potentially disrupting essential services and posing threats to national security [4]. The economic implications are also profound, as cybercrimes such as fraud, ransomware, and intellectual property theft inflict substantial financial losses on individuals, corporations, and entire economies. The evolving nature of cyber threats, with cybercriminals constantly developing more sophisticated methods to exploit vulnerabilities, underscores the ongoing and urgent need for robust and adaptive cybersecurity strategies [4]. These strategies are essential not just for mitigating immediate threats but also for fostering a secure, trustworthy environment conducive to the growth of the digital economy.

One of the major problems with cybersecurity is that infinite flaws in our current systems need to be patched. Cyber threats are becoming increasingly more advanced and sophisticated, with the Australian government reporting an average of 700 cyber threats targeting defensive networks in 2010, from just 300 in 2009 [4]. Attackers continuously find new exploits through users, hardware, and software, which can be used to hijack systems. Hence, many security updates are pushed to user computers. Unaware users might have their information stolen and sold when they open a link, accidentally download malware, or sign up for a legitimate service. Vulnerabilities can be introduced using outdated software with known weaknesses, such as improper handling of user input, which can lead to server-side attacks, outdated firmware, and unsupported devices [4]. Developers and engineers are constantly playing a game of "whack-a-mole," patching new vulnerabilities as soon as they are discovered.

VI. The Use of AI in Cybersecurity

AI is revolutionizing the field of cybersecurity, offering new and efficient ways to protect against cyber threats. One significant application of AI in cybersecurity is using Bayesian networks, as discussed by Pappaterra and Flammini. These networks allow for probabilistic reasoning and decision-making under uncertainty, crucial in identifying and responding to complex and evolving cyber threats [5]. AI-driven systems can analyze large volumes of data to detect patterns and anomalies that might indicate a security breach, offering a more dynamic and proactive approach to threat detection than traditional methods.

Another emerging trend is integrating blockchain technology in cybersecurity, particularly edge networks as Hazra, Alkhayyat, and Adhikari explored. Blockchain's decentralized and immutable ledger provides a robust framework for securing data transactions, making it increasingly relevant in the era of edge computing, where data is processed closer to where it is generated, reducing latency and bandwidth use [6]. This decentralization enhances data integrity and provides a higher level of security against common cyber threats, such as data tampering and unauthorized access.

Moreover, AI is tailored to address specific cybersecurity challenges in various domains. For instance, Elsayed and Zincir-Heywood's research on adaptive attack detection in vehicular networks highlight the specialized application of AI in protecting interconnected vehicular systems [7]. These adaptive systems can learn and evolve to identify new types of cyber threats, ensuring the security and efficiency of vehicular networks, which are becoming increasingly important with the rise of autonomous vehicles. This specialization of AI in different domains of cybersecurity demonstrates their flexibility and adaptability, making it an invaluable tool in the ongoing battle against cyber threats.

VII. The Malicious Use of AI in Cybersecurity

The malicious uses of AI represent a significant concern in today's digital landscape, with advancements in AI technologies opening up new avenues for cyber threats. One of the most prominent concerns is how AI can enhance traditional cyber attacks. For example, AI-driven malware can adapt and evolve to evade detection by security systems, making it more effective and damaging [8]. Phishing attacks, too, can be made more sophisticated with AI, enabling attackers to craft highly personalized and convincing fake messages or emails at scale, thereby increasing the likelihood of deceiving recipients [9]. Additionally, AI can be used for automated hacking, where AI systems are trained to discover and exploit vulnerabilities in software and networks much faster than human hackers [8].

On a broader scale, AI's malicious use extends to surveillance and social manipulation. Authoritarian regimes could potentially use AI for mass surveillance, leveraging facial recognition and other AI-driven monitoring tools to suppress dissent and track individuals. AI algorithms can be designed or manipulated in social media to spread propaganda, fake news, or biased content, influencing public opinion and polarizing societies [10, 11].

While AI presents numerous benefits and advancements in various fields, its potential for malicious use raises significant ethical, privacy, and security concerns. The creation of deepfakes, enhancement of cyber attacks, and possibilities of surveillance and social manipulation are just a few examples of how AI can be exploited for harmful purposes. These developments underscore the need for robust ethical guidelines and stringent security measures to govern the use of AI technologies.

A. Data Poisoning

Data poisoning attacks represent a significant threat in machine learning (ML) and AI, where they undermine the integrity of learning models by corrupting their training data. These attacks are particularly insidious as they target the core of ML systems – the data used for training. In such attacks, adversaries deliberately manipulate training data to induce desired errors or biases in the model's output, which can have far-reaching consequences [12].

For example, in a study by Alfeld, Zhu, and Barford, the authors explore how an attacker could subtly alter parts of the training data, leading to specific errors when the model is deployed [13]. In this case, the model could be autoregressive for tasks like financial forecasting or speech recognition. The attacker can manipulate the model's predictions by injecting carefully crafted 'poisoned' data points into the training dataset. Depending on the model's use case, this could lead to incorrect financial forecasts or misinterpreting spoken words.

Data poisoning attacks can take various forms, such as label poisoning, where the attacker changes the training data labels (e.g., labelling malicious emails as safe in a spam detection system) or feature poisoning, which involves altering the input features of the data [12]. These manipulations can be challenging to detect, especially in large datasets where the malicious data can be hidden among legitimate entries. The subtlety and effectiveness of these attacks make them a formidable challenge in cybersecurity, highlighting the need for robust data verification processes and continuous monitoring of ML models to ensure their integrity and reliability.

B. Model Inference Errors

Model inference errors are critical in AI, especially in applications where the outputs directly influence decision-making or product development. These errors occur when an AI model, trained on a specific dataset, makes incorrect predictions or generates inappropriate outputs when presented with new, real-world data. GitHub Copilot is a notable example of this issue, illustrating the broader challenges faced in AI-driven systems.

GitHub Copilot, an AI tool developed by GitHub and OpenAI, is designed to assist programmers by suggesting code snippets and functions. It is trained on a vast array of code from public repositories on GitHub. While this training approach enables Copilot to generate relevant and often useful code suggestions, it also exposes the tool to model inference errors [14]. For instance, Copilot might suggest code snippets that are syntactically correct but contextually inappropriate or insecure for the specific application at hand. This issue arises because, while the AI has 'learned' from a diverse set of coding examples, it doesn't inherently understand the intent or the specific security and functional requirements of the project it's assisting with [14].

Another aspect of model inference errors, as seen in GitHub Copilot, relates to biases in the training data. If the model is trained on datasets that include biased or flawed code, these inadequacies can be reflected in the suggestions the model makes [14]. This phenomenon is not limited to coding AI; it's a well-documented issue in various AI applications, from natural language processing to image recognition. The AI, in essence, mirrors the data it was trained on, including any inherent biases or errors.

The implications of model inference errors are significant. In the context of GitHub Copilot, these errors can introduce security vulnerabilities, maintenance issues, or inefficient coding practices in software projects. This necessitates that developers remain vigilant and critically assess AI-generated code, rather than accepting it at face value [15]. It also underscores the importance of diverse, high-quality training datasets and ongoing model evaluation and refinement to mitigate the risks associated with these errors.

C. Phishing

AI use in phishing attacks has drastically altered the dynamics of cyber threats, making them more personalized and sophisticated. AI-driven phishing, often more targeted than traditional methods, leverages AI's power to craft messages specifically tailored to individual targets, making them appear more credible and increasing the likelihood of successful deception [16].

For instance, an AI system can analyze a person's public digital footprint, gathering information from social media profiles, professional networking sites, and other online activities. This data is then used to personalize phishing messages, making them relevant and convincing to the target. An example of this could be a phishing email that appears to come from a known contact or organization, containing details specific to the individual's interests or recent activities, thus increasing the chance of the recipient trusting the content [16].

Moreover, AI's Natural Language Processing (NLP) capabilities enable the creation of phishing messages that are grammatically accurate and stylistically similar to legitimate communications the recipient is accustomed to receiving [16]. This sophistication in language use can make phishing attempts much harder to distinguish from genuine interactions.

A study, "DeepPhish: Simulating Malicious AI, " presented at the 2019 Workshop on Artificial Intelligence and Security, demonstrated the effectiveness of AI-generated phishing content [9]. In this study, an AI model was trained to produce phishing tweets, which were compared against human-generated content. The AI-generated tweets were more effective, illustrating the potential for AI to enhance the effectiveness of phishing attacks.

These advancements underscore the importance of heightened awareness and advanced security measures to combat AI-enhanced phishing attacks. As AI evolves, the need for sophisticated, AI-aware cybersecurity defences becomes increasingly crucial, highlighting a growing concern in the digital age.

D. *Bias Amplification*

Bias amplification in AI and ML refers to the phenomenon where an AI system exacerbates existing biases in its training data. This can lead to outcomes where the AI's decisions or predictions are more biased than the original data, amplifying societal or systemic prejudices [17].

An illustrative example of bias amplification is seen in natural language processing (NLP) systems. When trained on text data sourced from the internet, which often reflects and perpetuates societal biases, these systems can develop skewed understandings of language and context [18]. For instance, an NLP system might associate certain professions or activities predominantly with one gender, based on the biases in its training data. This issue was highlighted in a study by Bolukbasi, which found gender biases in commonly used word embedding models. These models would associate words like 'homemaker' more closely with women and 'programmer' with men, reflecting and amplifying societal gender stereotypes [18].

The ramifications of bias amplification in AI systems extend to various domains, such as recruitment software favouring candidates based on gendered language in their resumes, or credit scoring algorithms that disadvantage certain demographic groups [17, 19]. This perpetuates existing inequalities and can lead to discriminatory practices becoming embedded in automated systems.

VIII. Use of AI in the CSE

The Canadian Security Establishment(CSE) increasingly integrates AI into its operations. This shift is largely driven by the exponential growth in publicly available information from social media, smartphones, and IoT devices. AI, especially ML, has become essential in handling vast amounts of data and generating insights [20]. The CSE's mandate includes providing foreign intelligence, cybersecurity, and technical assistance to various federal entities, making the use and analysis of publicly available information crucial.

One key aspect of the CSE's AI integration is its focus on open-source intelligence (OSINT). OSINT now involves a bulk collection of publicly available information to apply big data analytics, including ML, for insightful conclusions. The CSE Act authorizes the CSE to acquire, use, analyze, retain, and disclose publicly available information, defined as information that is published, broadcast, or otherwise accessible to the public. However, the CSE Act also places constraints on protecting privacy, excluding information where individuals reasonably expect privacy. CSE implements privacy protection measures, such as anonymization and encryption. However, these measures are not foolproof, as most de-identification techniques are vulnerable to re-identification attacks, a risk compounded by the pattern recognition capabilities of AI and ML technologies. To address this, homomorphic encryption, allowing computations on encrypted data is considered a potential solution [20].

IX. Conclusion

In synthesizing the various aspects of AI's impact on cybersecurity and society, it becomes evident that while AI offers transformative potential, its ethical development and application are paramount. This aligns with the perspectives presented in Daron Acemoglu's article on redirecting AI, emphasizing the need for responsible AI development [21]. The exploration of AI's applications in cybersecurity, whether in enhancing protective measures or in the malicious exploitation of technology, underscores the dual-edged nature of AI. AI's versatility is apparent from using Bayesian networks and blockchain in cybersecurity to the malicious creation of deepfakes and AI-driven phishing attacks. However, this versatility comes with significant ethical responsibilities.

Though shrouded in confidentiality, the Canadian Security Intelligence Service (CSIS)'s use of AI likely mirrors global trends in utilizing AI for data analysis, counterterrorism, and cybersecurity. These applications, while beneficial, also necessitate stringent ethical oversight to prevent misuse and ensure respect for privacy and human rights. Similarly, issues like data poisoning, model inference errors, and bias amplification in AI systems reveal the complexities and potential pitfalls in AI development. These challenges highlight the critical need for diverse, unbiased training datasets, rigorous algorithmic audits, and an ongoing commitment to ethical AI practices.

The research objectives outlined in this paper, focusing on unethical AI applications such as data poisoning, phishing, and bias amplification, contribute significantly to understanding AI's broader societal impact. By proposing practical solutions and best practices, the research aims to guide AI development onto an ethical pathway, aligning with Acemoglu's call for redirecting AI towards societal benefit [22]. This approach is crucial for fostering trust between humans and AI models and ensuring that AI enhances human life rather than undermines it.

In conclusion, AI's ethical development and application, especially in sensitive fields like cybersecurity, are critical. This entails a commitment to transparency, continuous bias monitoring, and adapting AI applications to serve the greater good. By addressing the root causes and consequences of unethical AI applications, we can steer AI development toward a more responsible and beneficial trajectory that upholds societal values and protects human rights.

References

[1] A. Chernikova, N. Gozzi, N. Perra, S. Boboila, T. Eliassi-Rad, and A. Oprea, "Modeling self-propagating malware with epidemiological models," *Applied Network Science*, vol. 8, no. 1, p. 52, Dec 2023.

[2] D. Sheniar, N. Hadaad, and R. Addie, "The inference graph of cybersecurity rules," in *2019 29th International Telecommunication Networks and Applications Conference (ITNAC)*, Nov 2019, pp. 1–6.

[3] Y. Zhang and H. Dong, "Criminal law regulation of cyber fraud crimes—from the perspective of citizens' personal information protection in the era of edge computing," *Journal of Cloud Computing*, vol. 12, no. 1, p. 64, Dec 2023.

[4] K.-K. R. Choo, "The cyber threat landscape: Challenges and future research directions," *Computers & Security*, vol. 30, no. 8, pp. 719–731, Nov 2011.

[5] M. J. Pappaterra and F. Flammini, "A review of intelligent cybersecurity with bayesian networks," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Bari, Italy, 2019, pp. 445–452.

[6] A. Hazra, A. Alkhayyat, and M. Adhikari, "Blockchain for cybersecurity in edge networks," *IEEE Consumer Electronics Magazine*, vol. 13, no. 1, pp. 97–102, 2024.

[7] M. A. Elsayed and N. Zincir-Heywood, "Boostsec: Adaptive attack detection for vehicular networks," *Journal of Network and Systems Management*, vol. 32, no. 1, p. 6, Mar 2024.

[8] Prepare for AI hackers | harvard magazine. [Online]. Available: https://www.harvardmagazine.com/2023/02/right-now-ai-hacking

[9] A. C. Bahnsen, I. Torroledo, L. D. Camacho, and S. Villegas, "Deepphish : Simulating malicious ai," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:51691528

[10] Cybersecurity and AI: The challenges and opportunities. [Online]. Available: https://www.weforum.org/agenda/2023/06/cybersecurity-and-ai-challenges-opportunities/

[11] NSA, FBI, and CISA release cybersecurity information sheet on deepfake threats | CISA. [Online]. Available: https://www.cisa.gov/news-events/alerts/2023/09/12/nsa-fbi-and-cisa-release-cybersecurity-information-sheet-deepfake-threats

[12] Adversarial machine learning - CLTC UC berkeley center for long-term cybersecurity. [Online]. Available: https://cltc.berkeley.edu/aml/

[13] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, Feb 2016.

[14] P. Roberts. Researchers demo flaws in GitHub copilot AI generated code and warn of AI bias. [Online]. Available: https://www.reversinglabs.com/blog/ai-automation-bias-could-lead-to-more-vulnerable-code-0

[15] J. Chaffer. GitHub copilot: The good, the bad, the ugly. [Online]. Available: https://spin.atomicobject.com/2021/10/01/github-copilot/

[16] Karissa. The rise of AI in phishing scams: How scammers use it and how we can fight back. [Online]. Available: https://fightcybercrime.org/blog/the-rise-of-ai-in-phishing-scams-how-scammers-use-it-and-how-we-can-fight-back/

[17] Auditors are testing hiring algorithms for bias, but big questions remain | MIT technology review. [Online]. Available: https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/

[18] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016.

[19] Bias isn't the only problem with credit scores—and no, AI can't help. [Online]. Available: https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/

[20] loprespub. The growing importance of open-source intelligence to national security. [Online]. Available: https://hillnotes.ca/2022/02/17/the-growing-importance-of-open-source-intelligence-to-national-security/

[21] D. Acemoglu. (2021, May) Ai's future doesn't have to be dystopian. [Online]. Available: https://www.bostonreview.net/forum/ais-future-doesnt-have-to-be-dystopian/

[22] ——. (2021, May) Ai can still be redirected. [Online]. Available: https://www.bostonreview.net/forum_response/ai-can-still-be-redirected/

Peterson Guo Peterson Guo is a full-time Electrical Engineering student at the University of Waterloo. Since the age of 12, Peterson has been fascinated by computers, often assembling and modifying Unix systems. Since then, he has added Cybersecurity threat protection, machine learning, and natural language processing to his expanding arsenal of interests. He hopes to work towards advancing research and development in the field of artificial intelligence in robotics. He was previously employed as a software developer at an organization specializing in using simulations for research, such as modelling Hypercortisolism in individuals with autism and modelling the effects of COVID-19 to advance research in the hopes of better understanding existing problems.