Report on Intelligent Crimes - Disrupting AI-Controlled Systems, Large-Scale Blackmail

Course: Recent Advances in Computer Science (date: 29.11.2023)

Author: Mateusz Guściora, 228884

Table of Contents

ABSTRACT

This report delves into the realm of "Intelligent Crimes: Disrupting AI-Controlled Systems, Large-Scale Blackmail," exploring the intersection of artificial intelligence (AI) with cybersecurity and crime. Utilizing a comprehensive methodology, the research encompasses a review of scholarly articles, professional reviews, and real-life case studies sourced from Google Scholar and various online platforms. Key topics include the classification of attacks and vulnerabilities in AI systems, the evolving nature of AI-assisted cyber attacks, and the implications of large-scale blackmail facilitated by AI. Real-life case studies illustrate these points, offering insights into actual incidents and their impact. Additionally, the report investigates AI's potential as a threat and the necessary cybersecurity measures to mitigate these risks. Future challenges in AI and cybersecurity are discussed, highlighting the need for continuous adaptation and innovation in defensive strategies. This report aims to provide a well-rounded understanding of the current state and potential future of AI in the context of cybercrime and security.

INTRODUCTION

The study of intelligent crimes, particularly those disrupting AI-controlled systems and large-scale blackmail, required extensive research through Google Scholar and various online sources. The topic's novelty, with AI's growing role in daily life and the evolving definitions of intelligent crimes and AI disruption, presented unique challenges. Nevertheless, the goal was to comprehensively understand and present these concepts. The presentation, developed from general to specific knowledge, introduces the concept of intelligent crimes: acts of disrupting AI-controlled systems and committing large-scale blackmail.

This section introduces the topic of Intelligent Crimes, specifically focusing on disrupting AI-controlled systems and large-scale blackmail. To understand this subject, it's important to clarify certain concepts. The term 'disrupting' is used here as a verb, signifying the act of causing a disruption. There are two key ideas that, while lexically similar, differ in meaning depending on the context: technological disruption and disruptive innovation. Technological disruption refers to significant changes brought about by new technologies that alter the way industries or markets function. Disruptive innovation is a more bussiness context concept and relates to the introduction of new technologies or methodologies that fundamentally transform existing market landscapes or create new ones.

For this presentation, disrupting AI-controlled systems is understood as the act of malicious activity such as attacking, threatening, committing crimes against AI systems or with use of AI. Disrupting AI systems involves actions that compromise or manipulate these systems, increasingly crucial in sectors like healthcare, finance, and transportation, where they manage infrastructure, analyze data, and make decisions. While AI systems offer significant benefits, their vulnerabilities and attractiveness for criminal exploitation cannot be overlooked. This research aimed to highlight recent trends in AI system disruptions and share insights on this emerging threat.

LITERATURE AND SOURCES REVIEW

My literature review encompassed a wide range of sources, providing a comprehensive understanding of the current landscape of artificial intelligence (AI) and its vulnerabilities, particularly in the context of cybercrime. Key sources included:

- Caldwell et al.'s "AI-enabled future crime" in Crime Science (2020): This source was instrumental in understanding the potential future landscape of AI-related crimes, providing a forward-looking perspective on the threats and challenges.
- Eggers, Shannon Leigh's work on AI and Machine Learning vulnerabilities: Her analysis provided a deeper understanding of the inherent weaknesses in AI and machine learning systems that could be exploited by cybercriminals.
- NIST's "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations" (2023): This recent publication from the National Institute of Standards and Technology offered a detailed taxonomy of AI attacks, which was essential for identifying and categorizing different types of threats.
- Real-life case studies and incident reports: Websites like atlas.mitre.org and surfshark.com were pivotal in providing real-world examples and case studies. These sources offered practical insights into how theoretical vulnerabilities in AI systems have been exploited in real scenarios, bridging the gap between theory and practice in my research.

Great source of knowledge and partially material for presentation was also gained from other literature resources as:

- Russell, Stuart J., and Peter Norvig's "Artificial Intelligence: A Modern Approach": This seminal work laid the foundational understanding of AI concepts and applications. It helped in contextualizing the technological aspects of AI that are susceptible to cyber threats.

- Wang, P. (2019), "On defining artificial intelligence": Published in the Journal of Artificial General Intelligence, this paper offered insights into the evolving definitions and scopes of AI, especially conept of redefining AI and intelligence.
- Pupillo et al.'s "Artificial Intelligence and Cybersecurity" (2021): This comprehensive study from the Centre for European Policy Studies provided insights into the policy and governance challenges in AI and cybersecurity.
- Blessing Guembe et al.'s review on AI-driven Cyber Attacks: This review provided a broad overview of the emerging threats in the field and was a valuable source for understanding the latest trends in AI-driven cyber attacks.
- European Economic and Social Committee's report on cyber risks: This report highlighted the importance of awareness and resilience in the private sector, offering a perspective on the broader societal and economic implications of AI vulnerabilities.

Each of these sources contributed to a multidimensional understanding of AI in the context of cybersecurity, ranging from theoretical foundations to practical applications and real-world vulnerabilities. This comprehensive review was vital in shaping a well-rounded perspective on the subject. Sources of information included scientific literature and profesional web sources e.g. blogs, university publications. Full list of sources is included in last section ‘REFERENCES’.

METHODOLOGY

In my research on the topic of "Intelligent Crimes: Disrupting AI-Controlled Systems, Large-Scale Blackmail," I employed a comprehensive and multi-faceted approach to gather information and insights. My primary tools for sourcing relevant data were Google Scholar and general web surfing, which provided access to a diverse range of reliable content. This included professional reviews, publications from universities and institutions like police departments and the European Defence Agency (EDA), as well as writings from subject-specific blogs and cybersecurity group publications.

I utilized specific keywords such as "disrupting AI," "AI-controlled systems," "Intelligent crimes," "cybersecurity AI," "cyber attacks," "Hacking AI," "blackmail," and "large-scale blackmail" to navigate through Google Scholar and search engines. This targeted search strategy enabled me to sift through the vast amount of information available online and focus on materials that were most relevant to my topic. A significant portion of my research was dedicated to uncovering real-life cases that exemplify the concepts and issues I was exploring. This task proved to be challenging, as not all information regarding such attacks is public – organizations and individuals often choose to withhold details for security reasons. Additionally, the use of AI systems in our society is still evolving, which means that cyber attacks specifically targeting these systems are not yet commonplace. However, the existing risk and the likelihood of an increase in frequency of these attacks in the future made this a crucial area of investigation.

The rationale behind choosing these methods was driven by the need for a thorough and up-to-date understanding of the subject. Given the rapidly evolving nature of technology and cybercrime, it was essential to access the latest studies and real-world examples. This approach ensured that my research was grounded in current realities and emerging trends in the field of AI and cybersecurity. Additionll research was to gather relevant graphic representation like images and icons of situation and concepts for presentation purposes. All sources of this images are in the end of presentation.

ATTACKS AND VULNARABILITIES CLASSIFICATION

To simplify the categorization for a student audience, the classification was designed based on two main factors: the technical aspects of AI systems and the various ways they can be attacked. This approach helps in breaking down the complex topic into more digestible parts, making it easier for listeners to grasp the key concepts.

The classification of vulnerabilities and attack types in AI systems was based on a combination of factors. As a result, vulnerabilities can be categorized into groups such as data-based vulnerabilities, model-based vulnerabilities, and environmental exploits. Each category represents a different aspect of AI systems that can be exploited. In terms of attack types, they can be grouped into physical, digital, and social engineering attacks. This categorization helps in understanding the varied nature of threats against AI systems. Another important perspective is to categorize attacks based on the stages of AI development, which offers a temporal view on potential vulnerabilities. These stages encompass Training Phase Attacks, where AI models are manipulated during their learning process; Testing Phase Attacks, targeting the evaluation phase of AI systems; Deployment Phase Attacks, focusing on the operational use of AI; and Post-Deployment Attacks, which occur after the AI system has been established. This framework provides a comprehensive understanding of how and when AI systems can be targeted, highlighting the evolving nature of threats throughout the lifecycle of AI development.

With this straightforward and accessible categorization, it becomes much easier to understand and remember the most common methodologies used in attacks on AI systems. This approach aims to demystify the complexities of AI vulnerabilities and attack strategies, making the information more approachable for students and those new to the subject.

ATTACK METHODS

Various attack methodologies targeting AI systems were discussed. These include AI Model Poisoning, where malicious data is inserted into the dataset, leading the AI to make incorrect generalizations. Adversarial Attacks involve small changes in input data that deceive AI models. Attackers can exploit inherent biases and loopholes in AI models. Extraction attacks, like Model Inversion and Model Stealing, involve reverse-engineering the model or input data. Oracle attacks and system infiltration techniques were particularly emphasized due to their sophisticated nature. These methods involve manipulating AI systems by providing them with misleading information or queries, often to probe for vulnerabilities or extract sensitive information. In the context of backdooring models, attackers can insert hidden vulnerabilities during the development phase, which can be later exploited. This can involve code injection, buffer overflows, or other techniques that target specific vulnerabilities inherent to AI architectures. Additionally, other threats, among many, like Botnets, attacks on supply chains, resource exhaustion attacks, and side-channel attacks on hardware, were mentioned to provide a comprehensive overview of the threats facing AI systems.

Expanding on the various attack methodologies targeting AI systems, it's important to consider the broader implications of these vulnerabilities. AI Model Poisoning, for example, not only affects the immediate output of an AI system but can also have long-term consequences on its learning and decision-making processes. Similarly, Adversarial Attacks, though often subtle, can undermine the reliability of AI in critical applications, such as in autonomous vehicles or medical diagnostics, where precision is paramount.

Moreover, the sophistication of Oracle attacks and system infiltration techniques poses significant challenges. These methods, by probing for vulnerabilities or extracting sensitive information, can compromise not just individual AI systems but potentially entire networks. Backdooring models during the development phase introduces a stealth aspect to AI vulnerabilities, making detection and mitigation more complex. This is exacerbated by the diverse nature of attacks like Botnets, supply chain attacks, resource exhaustion, and side-channel attacks, each targeting different facets of AI systems and their operational environments. The cumulative effect of these threats underscores the need for a multi-layered, dynamic approach to AI system security, one that evolves in tandem with the advancing nature of these attack methodologies.

REAL LIFE CASE STUDIES – ATTACKS TARGET AI SYSTEMS

The report also delves into case studies of AI-controlled system disruptions. In 2019, Tesla's Autopilot system experienced a hacking incident, underlining the vulnerabilities in autonomous vehicles.

Microsoft's Tay AI Twitter Bot, introduced in 2016, faced manipulation when malicious users tweeted offensive language, causing Tay to generate similar responses. The attack on machine translation services in 2022, including Google Translate, Bing Translator, and Systran Translate, is another significant example. Additionally, a 2021 incident involved bypassing an automated identity verification system, leading to a fraud of $3.4 million. These cases demonstrate the varied and complex nature of threats facing AI-controlled systems.

These are real life examples of such attacks as we discussed. Some of them are real criminal activities and some are done by profesional cybesecurity entities for looking ways to improve defense of systems. These instances are great ways to learn and create defensive methods.

AI AS A THREAT - DISRUPTION

The section addressing AI as a threat was briefly touched upon in the report, recognizing it as a vast area requiring extensive scientific research. This acknowledgment reflects an understanding of the complexity and depth of the topic. This part delves into the various ways AI can enhance or create new cybersecurity threats. It covers aspects such as AI-enhanced phishing, malware, and Advanced Persistent Threats (APTs), which pose significant risks. Large-scale blackmail, often facilitated by technologies like deepfakes, is highlighted for its role in misinformation and undermining trust. The ethical and moral dilemmas posed by AI in autonomous weapons are discussed, along with AI's role in market manipulation and financial fraud. Lastly, the section touches on AI-enabled surveillance and its implications for privacy erosion. This comprehensive overview reflects the multifaceted nature of AI as a tool that, while beneficial, can be weaponized in various harmful ways.

The portion of the report focusing on AI's potential for disruption and weaponization provides a detailed analysis of the subject. It emphasizes the need for comprehensive research to fully grasp the extent and complexity of AI-related threats. Topics covered include the enhancement of traditional cybersecurity threats like phishing and malware through AI, and the emergence of formidable challenges like Advanced Persistent Threats. The section also sheds light on the increasing use of AI in large-scale blackmail schemes, especially those involving deepfake technology, illustrating its impact on misinformation and trust erosion. Further, it delves into the ethical quandaries presented by AI's application in autonomous weapons, its involvement in financial fraud, and the privacy concerns raised by AI-driven surveillance. This examination offers a nuanced view of AI as a technology with immense potential that, if misused, can lead to significant harm.

LARGE-SCALE BLACKMAIL VIA AI

In the report, it's crucial to delve deeper into how AI transforms the nature of large-scale blackmail. AI's capability to process and analyze vast amounts of data allows it to uncover sensitive information that can be used for blackmail. This is not limited to targeting high-profile figures; AI can enable mass targeting, where numerous individuals are blackmailed simultaneously. The sophistication lies in how AI can personalize threats based on the data collected, varying the content and delivery method for each target. This makes the threats more credible and harder to dismiss.

The process typically begins with extensive data collection and analysis, identifying potential targets based on vulnerabilities found in the data. Then, the content of the blackmail is tailored specifically for each target, taking into account their unique vulnerabilities and circumstances. The threats are delivered through automated communication systems, and AI continuously monitors the situation, adapting its strategy based on the responses from the targets. The final stages involve extortion, where compliance is sought, and managing post-compliance scenarios. This comprehensive, AI-driven approach to blackmail represents a significant and rising threat, necessitating sophisticated countermeasures from individuals and organizations alike.This form of criminal activity is arising threat for organizations and people. And need of defending from such crimes is rising aswell.

REAL LIFE CASE STUDIES - ATTACKS BY MEANS OF AI

In exploring AI-assisted cyber-attacks, the report examines several case studies. One involved over 20,000 WordPress sites being infected in a botnet-style attack, demonstrating how widespread AI-powered cyber threats can be. IG experienced significant cyber-attacks and data breaches in both August and November of 2019, highlighting the ongoing vulnerability of major platforms. In 2018, TaskRabbit faced a DDoS attack carried out by botnets, leading to a temporary service suspension. A notable case in 2020 involved a Hong Kong bank heist, where deepfake AI was used to clone a company director's voice, showcasing the advanced techniques used in modern cybercrime. These cases underline the diverse and sophisticated nature of AI-assisted cyber-attacks.

These case studies emphasize that AI-assisted cyber-attacks are a multifaceted and evolving threat. They demonstrate the importance of staying ahead of cybercriminals by adopting advanced cybersecurity measures, including AI-driven defense systems. In an era where technology continually advances, organizations must prioritize cybersecurity to protect their assets, data, and reputation from increasingly sophisticated AI-powered threats.

CYBERSECURITY MEASURES

In addressing cybersecurity measures against AI system disruption, the report emphasizes the importance of various strategies. It highlights robust encryption and data integrity protection as key defenses. Adversarial training is suggested to prepare AI systems for potential attack scenarios, enhancing their resilience. Anomaly detection systems are crucial for identifying unusual activities that might indicate a security breach. Regular audits and updates ensure that AI systems function as intended, while user education and awareness are essential for instilling security best practices. Access control and monitoring are emphasized to enforce strict access policies and surveillance. Finally, the report underscores the significance of legal and ethical considerations in guiding responsible AI development, a vital aspect of preventing AI system disruption.

Continuing with cybersecurity measures against AI system disruption, one should emphasizes AI's role in defense. AI-powered anomaly detection is highlighted for its ability to monitor and identify unusual behavior. Predictive threat intelligence uses AI to forecast potential security incidents, enabling proactive measures. Automated incident response is critical, as it allows for swift action against threats using AI. Additionally, machine learning (ML) is utilized for real-time threat detection, employing algorithms to identify threats immediately as they emerge. These AI and ML-driven approaches represent a significant advancement in cybersecurity, offering dynamic and effective tools for protecting against AI system disruptions.

Building on these collaborative efforts, the report also considers the broader implications of such partnerships. It suggests that by uniting the strengths and expertise of both public and private sectors, a more resilient and adaptive cybersecurity infrastructure can be developed. These collaborations are not only beneficial for immediate defense strategies but also play a pivotal role in shaping future cybersecurity policies and technologies. The importance of global cooperation cannot be overstated, as AI-driven threats often transcend national boundaries. By establishing a cohesive international framework and promoting open channels of communication, the global community can better anticipate and mitigate the risks associated with AI system disruptions.

FUTURE CHALLENGES

It is important aswell to try to predict future trends, challenges, chances and threats. Emerging challenges in the context of AI-controlled systems and intelligent crimes can be formulated and analyzed. The escalating complexity of AI systems presents a pivotal challenge. These systems, characterized by their intricate algorithms and expansive networks, are evolving at a pace that often outstrips existing cybersecurity measures. Therefore one of the challenges will be to make safety mechanics that will adapt to changing threats. Another challenge could be the arrival of quantum

computing. It is a major turning point that could greatly change AI and cybersecurity. It has the ability to weaken existing encryption methods, making many current security measures less effective. On the other hand, it also brings new and powerful ways to improve cybersecurity.

Anticipatory security, a forward-thinking approach, focuses on predicting and neutralizing cyber threats preemptively. This methodology, especially when applied to AI systems, demands a profound understanding of potential attack vectors and the deployment of predictive AI models. This methodology can be crucial step in such cybersecurity matter. Integrating AI into real-world systems such as infrastructure and transport increases the chances of cyber-physical attacks. These are situations where digital weaknesses can cause real-world damage. This highlights the importance of a security strategy that covers both digital and physical elements. The AI vs. AI paradigm is another growing concern, where defensive AI systems counteract AI-driven attack mechanisms. This scenario is quite hard to predict and analyze. Next challenge is idea of information and misinformation in media channels. The proliferation of synthetic media, notably deepfakes, introduces a novel avenue for disinformation. These AI-generated forgeries can be weaponized for malicious purposes, complicating the integrity of information dissemination. Developing robust detection and mitigation tools for such synthetic media is crucial for our privacy and access to realiable information.

The intersection of AI in criminal activities and cybersecurity solutions brings forth profound ethical and moral dilemmas. Balancing effective crime prevention with the safeguarding of individual rights, particularly in AI-driven surveillance and law enforcement, is paramount. This challenge necessitates a nuanced understanding of ethical principles in technological application. Lastly, the evolving landscape of AI-driven crimes necessitates a reevaluation of legal and regulatory frameworks. Current laws may not be updated enough and objective of law and political entities should be to take care of these current and future challenge.

CONCLUSIONS

In conclusion, this report has comprehensively addressed the multifaceted nature of AI systems as both a target and a source of malicious attacks, delving into the realm of intelligent crimes and their perpetrators. We explored the evolution of AI-related threats, starting from early exploits to more sophisticated attacks and recent trends, thereby presenting a clear picture of the current threat landscape.

Key insights from the study include evolving nature of AI-Related Threats: The research underscores the rapid evolution of threats associated with AI systems. From early exploits to advanced techniques like AI model poisoning and adversarial attacks, cybercriminals are continually adapting their strategies to exploit AI vulnerabilities. Real-life case studies were explored to illustrate these concepts in practical scenarios. The discussion on large-scale blackmail highlighted the complex challenges faced in this domain, while the exploration of defensive measures shed light on the necessary steps to mitigate these risks. Finally, the report underscored the challenges ahead, emphasizing the need for continuous vigilance and innovation in cybersecurity strategies to counteract these evolving threats.

REFERENCES

LITERATURE/ARTICLES:

- Russell, S. J., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach. London
- Wang, P. (2019). On defining artificial intelligence. Journal of Artificial General Intelligence, 10.
- Caldwell, M., Andrews, J.T.A., Tanay, T., & others. (2020). AI-enabled future crime. Crime Science.
- Eggers, S. L. (2020). Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data. United States
- National Institute of Standards and Technology (NIST). (2023). AI 100-2 E2023, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

- Pupillo, L., Fantin, S., Ferreira, A., & Polito, C. (2021). Artificial Intelligence and Cybersecurity: Technology, Governance and Policy Challenges. Centre for European Policy Studies (CEPS). Retrieved November 14, 2023, from:
- https://eda.europa.eu/docs/default-source/documents/ceps-tfr-artificial-intelligence-and-cybersecurity.pdf (date: 14.11.2023)
- Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The emerging threat of AI-driven cyber attacks: A review. Applied Artificial Intelligence.
- European Economic and Social Committee. (2018). Ensuring awareness and resilience of the private sector across Europe in face of mounting cyber risks.

BLOG/WEBSITE SOURCES:

- https://www.smartinsights.com/managing-digital-marketing/marketing-innovation/artificial-intelligence-adoption-different-sectors/ (date: 15.11.2023)
- https://surfshark.com/research/chart/statistics-of-ai-incidents (date: 14.11.2023)
- https://atlas.mitre.org/studies/ (date: 11.11.2023)
- https://www.belfercenter.org/publication/AttackingAI (date: 14.11.2023)
- https://www.infoq.com/articles/ai-cyber-attacks/ (date: 15.11.2023)
- https://proprivacy.com/privacy-news/deepfake-technology-used-in-hong-kong-bank-heist (date: 15.11.2023)
- https://www.linkedin.com/pulse/new-vulnerability-popular-wordpress-plugin-exposes-over-2-million (date: 15.11.2023)
- https://www.thalesgroup.com/en/worldwide/security/press_release/french-mod-challenge-thales-performs-successful-sovereign-ai-hack (date: 14.11.2023)
- https://www2.deloitte.com/us/en/insights/focus/tech-trends/2022/future-of-cybersecurity-and-ai.html (date: 14.11.2023)