ARTIFICIAL INTELLIGENCE CRIME: AN OVERVIEW OF MALICIOUS USE AND ABUSE OF AI

SEMINAR REPORT

Submitted by

ALLWINA ANNA SOY JOSE

Reg. No: SJC20CS021

to

the APJ Abdul Kalam Technological University
in partial fulfillment of the requirements for the award of the degree
of
Bachelor of Technology
in

Computer Science and Engineering

Department of Computer Science and Engineering

St. Joseph's College of Engineering & Technology, Palai

December:2023

St. Joseph's College of Engineering & Technology, Palai.
Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the seminar report entitled "**Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI** " submitted by **Allwina Anna Soy Jose (SJC20CS021)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide report of the seminar presented by her under my guidance and supervision.

Supervisor
Ms. Maria Yesudas
Assistant Professor
Department of CSE

Seminar Coordinator
Dr. Joby P P
Professor & Head
Department of CSE

Head of the Department
Dr. Joby P P
Professor & Head
Department of CSE

Place : Choondacherry
Date : 01-12-2023

ACKNOWLEDGEMENT

The success and final outcome of this seminar required a lot of guidance and assistance from many people, and I feel extremely privileged to have received this support throughout the completion of my seminar. Everything I have achieved is a result of the supervision and assistance I received, and I will never forget to express my gratitude.

I would like to extend my sincere respect and gratitude to the management of St. Joseph's College of Engineering and Technology for providing me with the opportunity and platform to undertake this seminar work.

I am deeply thankful to our beloved Principal, Dr. V P Devassia, for offering such excellent support and facilities for the successful completion of this seminar work.

I am also greatly indebted to our Head of the Department, Dr.Joby P P, Professor in the Department of Computer Science and Engineering, for the invaluable suggestions and encouragement throughout my Seminar work.

I owe a profound debt of gratitude to my seminar guide, Ms.Maria Yesudas, Assistant Professor in Computer Science and Engineering, who showed a keen interest in my seminar and guided me until its completion by providing all the necessary information.

I am thankful and fortunate to have received constant encouragement, support, and guidance from all the faculty members of the Department of Computer Science and Engineering, which played a pivotal role in the successful completion of my seminar work. Additionally, I would like to extend my sincere appreciation to all the laboratory staff for their timely support.

Allwina Anna Soy Jose

CONTENTS **PAGE NO**

ABSTRACT

Artificial Intelligence (AI) has rapidly evolved and significantly impacted various sectors of society. While this powerful technology offers potential benefits, it has also been increasingly utilized for criminal and harmful purposes, exacerbating existing vulnerabilities and introducing new threats. In this article, an in-depth exploration of this phenomenon is conducted, drawing from relevant sources such as literature, reports, and real-world incidents to create a comprehensive classification of the misuse of AI-powered systems. The primary objective is to illuminate the various AI-related activities and the risks linked to them. The process begins with an examination of the inherent weaknesses in AI models, followed by an explanation of how malicious individuals can exploit these vulnerabilities. Also, delving into AI-driven attacks, which are made possible or more potent due to AI technology. Although the aim is to provide a broad overview, there is no intention to present an exhaustive categorization. Instead, the objective is to offer a comprehensive view of the risks stemming from the increased use of AI, contributing to the growing knowledge on this pressing issue. Specifically, four categories of malicious AI misuse are proposed, encompassing attacks on integrity, unintended AI consequences, algorithmic trading manipulation, and membership inference attacks. Moreover, four types of malicious AI usage are identified, covering social engineering, the dissemination of misinformation and fake news, hacking, and the development and deployment of autonomous weapon systems. This thorough assessment of threats forms a basis for more advanced discussions on governance strategies, policies, and proactive measures that can be developed or refined to mitigate these risks and prevent adverse outcomes. It underscores the urgent need for enhanced cooperation among governments, industries, and civil society stakeholders to strengthen preparedness and resilience against the malicious exploitation of AI.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ACRL	-	Association of College and Research Libraries
AI	-	Artificial Intelligence
CCW	-	Certain Conventional Weapons
DARPA	-	Defense Advanced Research Projects Agency
FCC	-	Federal Communications Commission
GANs	-	Generative Adversarial Network
GGE	-	Group of Governmental Experts
ML	-	Machine Learning
NLP	-	Natural Language Processing
PUAs	-	Potentially Unwanted Applications
STEM	-	Science, Technology, Engineering, and Mathematics

Chapter 1

INTRODUCTION

Artificial Intelligence (AI) systems are at the forefront of numerous academic studies, political discussions, and civil society organization reports[1]-[3]. The development of AI has been lauded for its unprecedented technological capabilities, such as improved automated image recognition (e.g., cancer detection in medicine[4]). However, it has also been criticized and even feared due to uncertain implications of automation on the job market (e.g., fears of widespread unemployment). This dichotomy of positive and negative aspects can also be seen in the context of cybersecurity and cybercrime. Governments are using AI to enhance their capabilities, but the same technology can be used for attacks against them. The recent acceleration in AI development, driven by the private sector and customer-oriented applications, suggests that sectors like defense might utilize similar capabilities. It is increasingly challenging to differentiate between the actions of state and non-state actors, as evidenced by recent ransomware attacks on public infrastructure in various countries, such as the Colonial Pipeline in the United States in May 2021. Moreover, programs and applications created for benign purposes can also be adapted or modified for malicious intent, potentially causing harm.

The dual-use nature of technology is not a new issue in the context of cybercrime or cybersecurity. However, the potential for AI to be exploited for malicious use and abuse presents new vulnerabilities. Constant assessment of the threat landscape is essential to develop and adjust governance mechanisms, create proactive measures, and enhance resilience.

Building on previous work, this article assesses the primary categories of AI use and abuse in a criminal context. We provide several notable examples to illustrate the challenges we face. Based on these examples, we propose a typology that catalogs the main harmful AI-based activities. Developing knowledge and understanding about the potential malicious use and abuse of AI allows cybersecurity organizations and government agencies to anticipate such incidents and increase their preparedness against attacks. Furthermore, a typology is extremely useful in organizing research efforts and identifying knowledge gaps where more research is needed.

Chapter 2

AIM OF THE STUDY AND CONTRIBUTIONS

To effectively counter AI-related threats, it is crucial to comprehend the diverse forms of malicious use and abuse of AI, along with their associated risks. Unfortunately, there is a noticeable absence of a comprehensive interdisciplinary assessment of AI-enabled and AI-dependent cyberattacks, hindering the development of robust countermeasures. This gap jeopardizes data security, personal safety, and political stability. This study seeks to address this by categorizing various forms of malicious AI use, expanding the understanding of the subject in a holistic manner.

The primary goal of this research is to propose a typology of malicious AI use based on empirical evidence and contemporary discourse. The analysis focuses on how AI systems compromise confidentiality, integrity, and data availability. Utilizing a classification technique established over 2000 years ago, this study aims to lay the groundwork for granular and in-depth analysis rather than developing a theory of malicious AI use[3].

The objectives are limited to identifying crucial elements of malicious AI use and collecting evidence of their practical application[1]. The compiled data will facilitate further exploration of potential ways in which AI systems can be exploited for criminal activities. It's important to note that this research does not aim to develop a comprehensive theory of malicious AI use.

Through the proposed typology, this study intends to make the following contributions:

1. Contribute to the growing body of knowledge mapping types of malicious AI use, aiding in the understanding of key concepts, threat scenarios, and possibilities. This understanding is essential for the development of preventive measures and proactive responses to AI-related attacks.

2. Facilitate the establishment of a shared language across disciplines, particularly between STEM fields and legal practitioners, as well as policymakers. Interdisciplinary collaboration can alleviate confusion arising from overly technical or monodisciplinary language and bridge existing gaps in understanding.

3. Propose mitigation strategies and emphasize the need for a collective effort involving government, academia, and industry. By doing so, this research aims to underscore the importance of collaborative measures in addressing the multifaceted challenges posed by malicious AI use.

Chapter 3

OVERVIEW OF MALICIOUS USE AND ABUSE OF AI

The advent of Artificial Intelligence (AI) heralds a transformative era across diverse sectors, promising groundbreaking advancements in healthcare, education, transportation, and beyond[1]. This technology holds the potential to revolutionize how we approach complex challenges, enhance decision-making processes, and improve overall efficiency. However, amidst its immense promise, the flip side of AI emerges—a potential for malicious use and abuse.

3.1 Malicious Abuse of AI: Vulnerabilities of AI Models

3.1.1 Integrity Attacks

The increasing prevalence of machine learning (ML) has sparked heightened interest among attackers seeking to manipulate either the models themselves or the underlying data. This susceptibility makes ML models vulnerable to integrity attacks, where malicious actors attempt to introduce false information into the system, compromising its trustworthiness[1]. An associated risk arising from the vulnerability of AI models is the emergence of 'adversarial examples.'These are inputs intentionally crafted to deceive machine learning models, leading to misclassifications of scrutinized material by the systems. At times, these manipulations are so subtle that they escape human detection, yet they induce errors in AI systems.

A specific instance of adversarial ML is a 'poisoning attack,' wherein an attacker influences the training data to manipulate the results of a predictive model by injecting corrupted points during the training process[6]. Essentially, poisonous samples introduced into the training data can alter the behavior of the classifier, resulting in undesirable outcomes. An illustrative example is the attack on Tay, Microsoft's AI chatbot, released in 2016. The chatbot aimed to generate tweets indistinguishable from those of a human. Shortly after its release, users orchestrated a coordinated attack, tweeting offensive content to exploit Tay's "repeat after me" function, causing the bot to reproduce objectionable material. According to Microsoft's Corporate Vice-President, "Although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack." Consequently, Microsoft had to suspend the account in less than 16 hours, highlighting the difficulty in defending a chatbot against attacks, especially when trained in unpredictable online environments with live interactions.

Researchers at New York University (NYU) identified another risk related to outsourced training data. They demonstrated the creation of a BadNet, a maliciously trained network exhibiting normal behavior until triggered by a potential attacker. In a practical test, BadNets were incorporated into a complex traffic sign detection system, revealing that a self-driving car could correctly identify a stop sign until one with a predefined trigger (a yellow 'Post-It' note) was presented. This study underscores the susceptibility of AI models to data poisoning and adversarial examples, leading to misclassifications and potentially severe consequences that may be unpredictable for individuals unfamiliar with the technology. This could be a motivating factor behind the specific requirements for training data in 'high-risk systems' outlined in Article 10 of the recently proposed EU AI Act.

3.1.2 Unintended Outcomes of the Use of AI

AI models, particularly those relying on neural networks, may produce unexpected results due to factors like inadvertent memorization of sensitive information during the learning process [5]. This becomes a significant concern when the training data involves private or confidential details. The phenomenon suggests that models might unintentionally memorize irrelevant information not directly tied to the primary task.

To tackle this issue, the team behind Smart Compose, the real-time suggestion system used in Google's Gmail service, took a careful approach. They conducted extensive testing to ensure the model memorizes only common phrases shared by multiple users. Their goal was to prevent the algorithm from learning and disclosing details unrelated to its primary task, such as private information. For instance, they aimed to avoid the model suggesting text completions containing another user's ID number when a user inputs a text prefix like "my ID number is". In this scenario, the challenge lies not in the developer's malicious intent but in the potential for the model to behave differently than anticipated, specifically by unintentionally memorizing private data.

3.1.3 Algorithmic Trading/Stock Market Manipulation

Through the utilization of computers and AI-driven software, technology plays a pivotal role in expediting financial analysis and decision-making. The incorporation of AI systems in market trading, known for their "lightning speed", brings about both positive and negative repercussions. On the positive side, contemporary financial technology has notably reduced transactional fees and capital costs for businesses. However, algorithmic trading, marked by decisions challenging for humans to comprehend, introduces market instability, giving rise to the risk of high-speed crashes, commonly referred to as flash crashes. David Weild IV, the former vice-chairperson of NASDAQ, straightforwardly asserts that the stock market has become too rapid for human comprehension, leading to unexpected and startling outcomes.

The challenges of automated decision-making in finance came to the forefront following the 2010 flash crash, resulting in a staggering loss of nearly $1 trillion. High-frequency trader Navinder Singh Sarao, sentenced in 2020 to a year of home confinement, was implicated in this incident[1]. Sarao was accused of utilizing an automated program to generate large sell orders, artificially lowering prices. Subsequently, he canceled these orders to buy at lower market prices, reaping benefits when the market rebounded. This inaugural market crash in the era of algorithmic trading acted as a wake-up call, prompting not only traders but also regulators to recognize the challenges of high-speed automated trading and decision-making. To avert similar incidents, techniques such as spoofing and layering, employed to manipulate high-frequency trading, were banned.

Discussions regarding regulatory frameworks often center on market harm caused by malicious actors. While this evaluation is crucial, it is equally important to consider potential actions in the event of a technological mishap or insufficient testing. As algorithm-driven trading gains prominence in stock markets, the likelihood of recurring flash crashes increases. In such a volatile environment, situations can rapidly escalate, leading to unforeseen consequences. One possible policy response to mitigate the impact of flash crashes involves the establishment of insurance systems. The creation of a financial market fund named the "National Protection Fund," designed to compensate investors adversely affected by market disruptions caused by algorithms, thereby ensuring greater stability and safety in trading. Additionally, reinforcing cybersecurity measures and conducting thorough assessments of algorithms could contribute to averting the detrimental consequences of high-speed crashes.

3.1.4 Membership Inference Attacks

Membership inference attacks involve a malicious actor attempting to unveil and reconstruct the samples utilized in training a machine learning (ML) model[1]. These attacks prove effective across various systems, encompassing classification, sequence-to-sequence models, and even generative adversarial networks (GANs)[2]. GANs, a type of deep-learning model, are adept at generating seemingly authentic but synthetic examples of the data employed in the training process. This capability finds application in various domains, exemplified by the website https://thispersondoesnotexist.com/. In a recent investigation, showcased that the faces generated by the "This person does not exist" algorithm closely resemble the faces of individuals included in the training data. The study's authors asserted that through membership inference attacks, it becomes feasible to pinpoint samples that may not be identical but share the same identity. This raises concerns as attackers could potentially unveil the true identity of individuals whose photos constituted the training datasets. Consequently, membership inference attacks carry privacy implications, particularly impacting individuals whose faces contributed to ML model training. For instance, in the context of a medical data model, a similar attack could enable attackers to link a disease to a specific individual. These attacks are not confined to models utilizing biometric data datasets (e.g., facial images, voice recordings,

gait detection) but extend to others built on susceptible information such as genetic data. Strategies to mitigate the risk of membership inference attacks include ensuring models are trained on diverse datasets, minimizing dataset bias, and conducting thorough pre-testing to fortify the system against such vulnerabilities. Figure 3.1 delineates a comprehensive overview of the thematic landscape surrounding the vulnerabilities inherent in AI models. The visual representation encapsulates a spectrum of critical topics, elucidating key facets that demand attention within the realm of AI security. This graphical depiction serves as a valuable reference, guiding stakeholders and researchers in understanding the multifaceted dimensions of challenges associated with AI model vulnerabilities

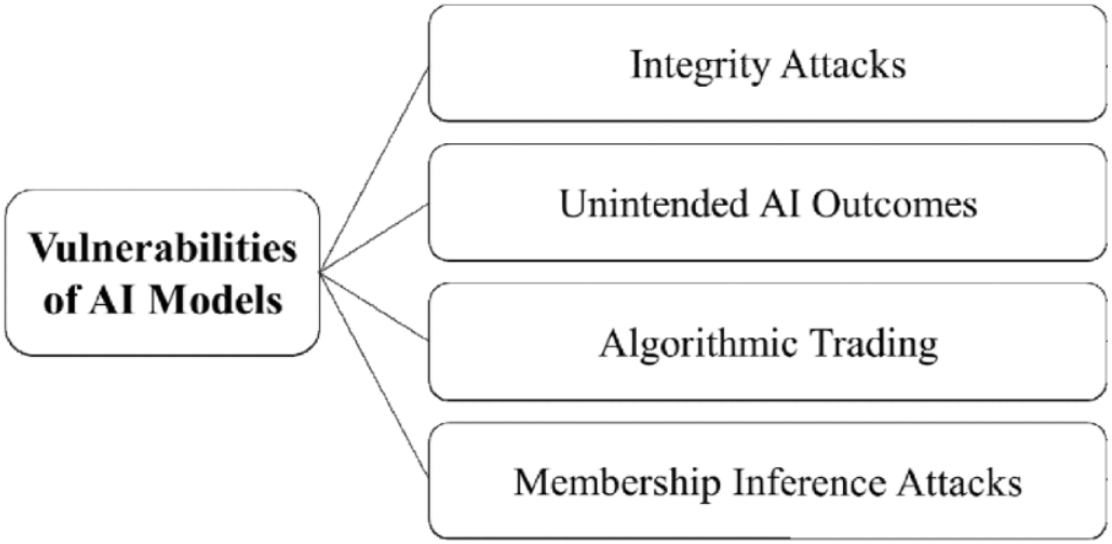


Figure 3.1: Malicious abuse of AI.

3.2 Malicious Use of AI: AI-Enabled and AI-Enhanced Attacks

3.2.1 Social Engineering

Social engineering attacks employ deceptive tactics to manipulate individuals into divulging sensitive or personal information, which can then be exploited for fraudulent purposes. These attacks are executed through various methods, utilizing a range of AI techniques. By leveraging these techniques, malicious actors can devise sophisticated manipulation strategies, thereby enhancing their likelihood of success and illicit gains.

Deception and Phishing

AI techniques empower hackers to create 'social bots' capable of deceiving and manipulating individuals into complying with their requests. These algorithms, designed to mimic human behavior, generate content and engage with users online. For example, these 'social bots' might initiate requests that lead the victim to a website, allowing the perpetrator to take control of the victim's computer. An early instance of a cyberattack utilizing AI techniques was the dating chatbot 'CyberLover,' introduced in 2007 to entice chatroom users into sharing personal information or clicking on fraudulent links[7]. This chatbot employed natural language processing (NLP) to deliver tailored dialogues, raising concerns about the application of advanced capabilities in cybercrime. Likewise, attackers may pose as trusted entities or companies to trick victims into opening emails or links, facilitating data theft. This tactic, known as phishing, can be further augmented by AI to enhance criminals' reach and success. An experiment utilized a machine learning-based model to generate text for posting on Twitter, demonstrating how AI could be employed to maximize phishing effectiveness. The chosen social media platform was Twitter due to its character limitations, making posts with broken English and shortened links appear acceptable. The findings suggest that the dynamics of such platforms may facilitate the utilization of machine-generated text for phishing.

Big Nudging and Manipulation

Beyond the potential targeted actions outlined in the preceding section, a substantial number of bots could be generated to support activities with malicious intent, including the manipulation of public opinion and election outcomes[1]. Social bots, for instance, can influence perceptions by retweeting specific content or replicating hashtags, creating a false sense of popularity for a candidate or political movement on social media platforms. A similar deceptive strategy is astroturfing, which simulates grassroots support for a policy or individual when little or no authentic backing exists. For example, an organization might publish numerous Twitter posts using various accounts to sway public opinion for or against a candidate in an election. Astroturfing can manifest in Twitter posts, blogs, news portals, and other online platforms, serving as a disinformation tactic.

Bots also have the capability to fabricate the perception of support for a cause during public consultations and manipulate poll results. This concern became evident after the Federal Communications Commission's (FCC) net neutrality consultation in the United States. As the FCC considered rolling back net neutrality protections, it sought public opinion through a comment section. Analysis by the data analytics company Gravwell revealed that over 80% of the approximately 22 million comments received by the FCC were submitted by bots. In this case, natural language generation was employed to artificially boost opposition to net neutrality protection.

AI's role in online profiling and targeting is another significant aspect. The Cambridge Analytica scandal provides a notable illustration, where the GSRApp collected users' personal data deceptively, including personality traits[1]. This data was then used to train an algorithm that generated personality scores for app users and their Facebook friends, matching them with U.S. voter records. Cambridge Analytica utilized this information to create voter profiles and deliver targeted advertising services. The manipulation of individuals through tailored messages based on their psychological profile, coupled with disinformation and provocative content, can impact democratic processes and influence election outcomes.

3.2.2 Misinformation and Fake News

The evolution and widespread use of technology, blogging platforms, and social media have transformed how individuals consume information, access news, and form opinions. The rapid nature of the internet enables anyone to create and swiftly share content, leading to an environment conducive to the creation and dissemination of misinformation and fake news. While the term "fake news" may be a subject of debate among some journalists and academics, it remains pertinent in fostering discussions on digital literacy and promoting scholarly research on the issue[1]. Furthermore, arguments against using the term have been shown to be insufficient.

Unverified rumors, speculation, and intentionally false information can have severe consequences, particularly in times of uncertainty and social unrest, such as during pandemics. Similarly, during political events like elections, the impact can be detrimental. AI systems pose a risk to society and democratic processes as they can fuel the creation and spread of such harmful content, potentially jeopardizing democracy itself.

Tools like GPT-3 have the potential to amplify the creation of misinformation. GPT-3, an autoregressive language model utilizing deep learning, excels in tasks like question-answering, text completion, and summarization. Due to its formatting, choice of words, and consistency, texts automatically generated by this tool may appear human-written, deceiving readers due to apparent credibility. "NotRealNews.net" exemplifies this, utilizing AI to generate fake news articles and demonstrating how such a tool could be used to support journalistic work. Given their convincing nature, these automatically generated texts could easily be disseminated as compelling fake news articles. This implies that when combined with current targeting capabilities, automatically generated texts could amplify the quantity, quality, and impact of fake news and disinformation campaigns. These campaigns might influence democratic processes to varying extents, from convincing voters to change their votes to reinforcing pre-existing views. Furthermore, as technology evolves, texts can be tailored to the audience's preferences, intensifying the prevalence of "filter bubbles" and polarization.

Several strategies can help mitigate the negative impact of AI systems in creating and disseminating fake news and misinformation. A study found that information literacy increases the likelihood of identifying fake news. As defined by the Association for College

and Research Libraries (ACRL), information literacy is the "set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating ethically in communities of learning". Therefore, educating individuals on the appropriate use of digital resources is crucial. The more citizens can navigate the online environment and critically evaluate information, the less impact unfounded stories will have on them and their communities.

Additionally, information systems and providers play pivotal roles. Given that many users access news and information based on algorithmic decisions, social media platforms (e.g., Facebook) and search engines (e.g., Google) face pressure to revise and improve their algorithms, structuring the presentation of content differently (e.g., in the context of the debate on a 'right to be forgotten'), and reducing the quantity of fake news on their feeds. Platforms are not mere intermediaries; their algorithms are designed to deliver specific types of content to users based on past activities and foster user engagement. In this model, a person who reads one or more pieces from a news outlet that disseminates false information is likely to receive additional content from the same source, as it creates engagement for the platform. Proactive enhancement of algorithms by tech companies can improve the detection of unreliable sources and contain their spread. Simultaneously, increased human monitoring and oversight are essential since only this type of control can understand information in context. Progress is being made in these areas, with the development of a somewhat independent oversight board by Facebook (now Meta) to "oversee" and check important decisions. However, much remains to be discovered and done.

Lastly, civil society organizations and activists can make positive contributions. An example is the initiative "Sleeping Giants Brasil," inspired by the Twitter account created in the United States after the 2016 election, called "Sleeping Giants". Activists, using a Twitter account, aimed to reduce the advertising revenue of certain news outlets known for spreading disinformation. Revenue is generated through the affiliation of these websites with Google's ad platform. Websites sharing fake news stories earn money based on the number of views and clicks on ads displayed on their sites. The Twitter account would publicly question brands about their support for such content after taking screenshots of

ads on these websites. Consequently, many companies would block their ads from appearing on these websites, reducing their stream of income. Similar accounts were created in Brazil to combat the spread of fake news.

3.2.3 Hacking

Forgery: Deepfakes

Prominent instances of forgery in the digital era involve deepfake videos and images, where highly realistic media leverages AI for creation, portraying individuals saying or doing things that never occurred. The use of AI in forging videos and images enhances the authenticity of the material, making it challenging to differentiate between reality and fabrication. While manipulation using programs like Photoshop is not new, AI introduces a higher level of sophistication and difficulty in detection. For instance, Ali Aliev developed a real-time deepfake creation method, showcasing its capabilities by joining a random Zoom meeting and impersonating Elon Musk[1]. Notably, deepfake materials predominantly feature well-known figures, such as celebrities and politicians. The threat posed by these videos and images stems from their potential use for malicious purposes, including propaganda, disinformation, bullying, revenge porn, or blackmail.

The malicious exploitation of forged videos can directly impact politics and international relations. An illustrative example is the creation of a fake video by the Democratic Party in the United States, featuring the chairman at a convention to emphasize concerns about the impact of deepfakes on democratic processes[7]. To mitigate the negative consequences of forged videos, one approach is to increase public awareness about such technology. Deepfake creator Bruno Sartori produces humorous videos depicting Brazilian national politics, incorporating absurdity to signal that they are not real and constitute elaborate satire. These videos, shared on social media, aim to demonstrate the risks of the technology to the public. The inoculation theory explains these interventions, suggesting that prior exposure and knowledge can help individuals interpret videos critically and be "inoculated" against maliciously forged content. In addition to awareness, there is a need to develop tools for deepfake detection, with AI techniques like recurrent neural networks being particularly useful.

The concept of the "liar's dividend," adds complexity to the issue. The liar's dividend arises when someone exploits the existence of deepfake videos to discredit a real video, claiming manipulation and casting doubt on its authenticity. The more the public becomes aware of AI's role in manipulating videos, the more skeptical they may become, questioning even authentic videos and images. This phenomenon, termed the liar's dividend, increases in proportion to the success of educating the public about the dangers of deepfakes. Consequently, there is a possibility that during elections, a candidate caught on tape might falsely assert that the video is a deepfake, convincing voters of their innocence. The effectiveness of regulations like the EU AI Act in addressing deepfakes remains uncertain in the current draft, as no dedicated prohibition is evident. However, the People's Republic of China is introducing relevant legislation that will mandate platform operators to prevent the spread of deepfakes on their networks.

Repetitive Tasks

AI demonstrates efficiency in executing repetitive tasks, which can potentially be exploited for malicious purposes. A notable example is the Ticketmaster incident, where AI tools were utilized to circumvent Captcha protections, allowing the purchase of thousands of tickets for later resale, and generating illicit profits. The application of AI in pattern recognition extends beyond defeating Captcha, raising concerns about other cybercrimes, including password cracking. Brute force attacks, a method for cracking passwords, can be laborious in terms of time and resources. However, it has been shown that AI-powered brute-force attacks achieve a significantly higher success rate compared to non-AI approaches. In essence, the progress in AI capabilities poses the risk of repetitive tasks being exploited for malicious activities, such as password cracking.

Malware

The malware threat has existed for several decades, dating back to the 1970s with the emergence of the Creeper Worm, the first documented malicious software. Since then, the landscape of cyber threats has evolved into a significant cybersecurity concern, with the AV-TEST Institute recording over 350,000 new malware and potentially unwanted applications (PUAs) daily. This translates to four new malware or PUAs being registered

every second. As malware developers continually innovate and create more sophisticated malicious programs, establishing effective and timely defense mechanisms becomes increasingly challenging. Presently, there are concerns about the potential use of AI techniques to craft more effective and harder-to-detect malware. However, as of our current knowledge, this technology is not yet extensively developed.

Current possibilities are primarily explored through academic research and as proof of concept by companies. IBM, for example, introduced DeepLocker at the Black Hat USA 2018. This system integrates AI with malware to enhance its evasion capabilities. DeepLocker exploits the lack of explainability in AI systems, typically viewed as a weakness, to its advantage. It employs a deep neural network to select targets and conceal its intent until it reach the desired destination. The primary risk associated with AI-enhanced malware like DeepLocker is its potential to infect numerous systems without detection. Moreover, the capabilities of such developing systems are not limited to states; civilians and private organizations can also contribute to the development of high-risk malware. Therefore, even if AI-enabled or AI-enhanced malware is not yet extensively developed, it is crucial to consider the potential risks associated with such a possibility.

Addressing the challenges posed by AI-based or AI-enhanced malware involves enhancing capabilities in the field of cyber autonomy. The feasibility of cyber autonomy was demonstrated during the Cyber Grand Challenge organized by the Defense Advanced Research Projects Agency (DARPA) in 2016[3]. Finalist teams in the competition were tasked with developing automated cyber defense systems capable of self-discovering, proving, and correcting software vulnerabilities in real-time. Throughout the competition, these systems demonstrated the ability to auto-detect and correct vulnerabilities while also launching attacks on the software of other participants in their network. Since then, a movement towards "security automation" has been identified, marking the initial steps toward cyber autonomy. Building capabilities in autonomous defensive cybersecurity is a strategy to leverage AI systems against malicious actors. However, given the dual-use nature of this technology, software developed for defense could also be repurposed for offensive activities. To mitigate this risk, clear regulations regarding the use of these systems and the implementation of security safeguards are essential.

3.2.4 Autonomous Weapons Systems (AWS)

The exploration of autonomy in weapons has been ongoing in military contexts since the advent of AI in the late 1950s. The appeal of Autonomous Weapons Systems (AWS) lies in the ability of machines to process data, analyze information, and make decisions faster than humans in certain situations, offering potential military and strategic advantages. However, along with these promises, AWS also introduces risks. AWS can be defined as AI systems designed to identify and engage targets without the need for human control or action after activation. The application of autonomous functions can be extended to various platforms, including ships or fighter jets.

One significant risk associated with this emerging technology is the potential for the software embedded in military hardware (e.g., drones) to be manipulated by malicious actors[1]. If a drone is hacked and the GPS location of an attack is altered, it would operate based on the new rules set in the software, potentially leading to unintended casualties. Similarly, if the data used to train these systems are compromised, it could have disastrous consequences. A 2014 report by Reprieve revealed that drone attacks, aimed at killing 41 individuals, resulted in the deaths of approximately 1,147 people, raising concerns about the accuracy and precision of 'targeted killing'[3]. Gibson, the lead researcher, emphasized that drone strikes are "only as precise as the intelligence that feeds them". The high risks associated with attacks on AI systems used in warfare are subjects of discussion in academia, civil society, and at the government level. However, as of now, there are no international regulations specifically addressing the use of AWS.

The implications of AI in warfare were initially discussed among state parties to the United Nations Convention on Certain Conventional Weapons (CCW), with a primary focus on international humanitarian law within the CCW. Ethical considerations play a secondary role in these discussions. From 2014 to 2016, annual Informal Meetings of Experts on AWS took place in Geneva. Subsequently, the CCW established a Group of Governmental Experts (GGE) on AWS, serving as the primary forum for international deliberations on autonomous weapons systems[7].

One potential avenue for regulation is the creation of an additional protocol to the existing convention, similar to previous protocols addressing weapons like those with non-detectable fragments, landmines, incendiary weapons, blinding laser weapons, and ex-

plosive remnants of war. However, past negotiations within the CCW, such as those on cluster munitions, have been moved outside the CCW due to a lack of consensus. In the case of cluster munitions, negotiations began outside the CCW in February 2007, leading to the adoption of the Cluster Munitions Convention in May 2008, which now has 110 state parties as of August 2021. The treaty was initially agreed upon by certain states and later adopted by others, suggesting a potential pathway for addressing AWS concerns. Figure 3.2 intricately maps the overarching themes associated with AI-enabled and AI-enhanced attacks. This visual representation offers a structured overview of the diverse topics that constitute the landscape of threats stemming from the intersection of artificial intelligence and cyber warfare. The figure serves as a comprehensive guide, delineating key considerations for stakeholders and researchers to navigate the nuanced challenges posed by malicious employment of AI in contemporary cybersecurity scenarios.

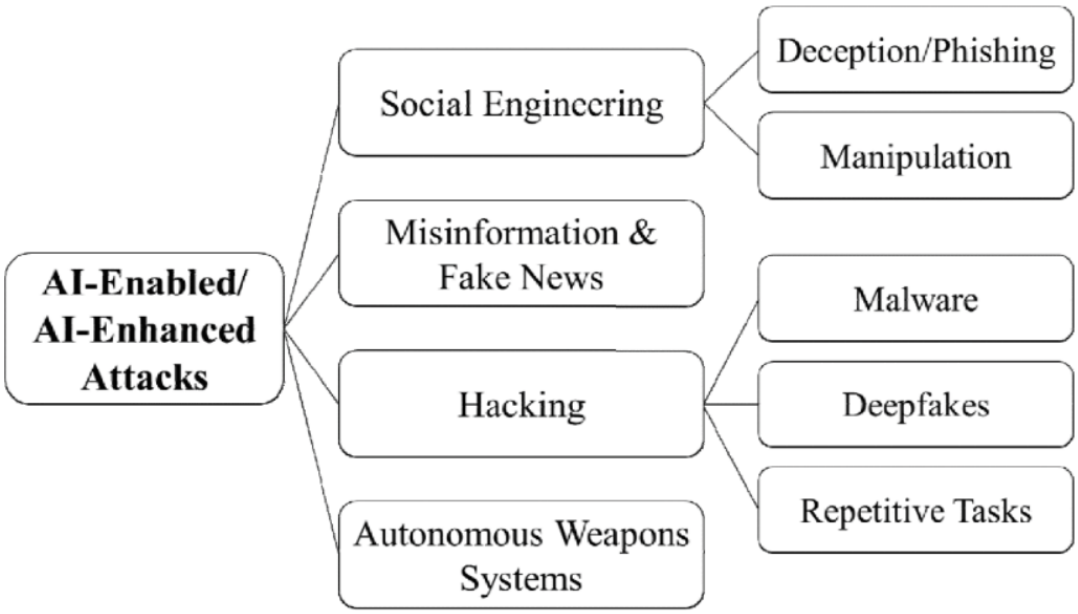


Figure 3.2: Malicious use of AI.

Chapter 4

RESULTS AND DISCUSSION

The paper titled "Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI" delves into the darker aspects of AI, focusing on how people can use it for less-than-positive purposes. It takes a deep dive into the weak spots of AI models and explains how individuals could exploit them. The analysis explores various types of attacks, both those powered by AI and those aimed at AI systems, based on a thorough review of relevant literature and real-life incidents.

It breaks down malicious AI abuse into four types: tampering with the integrity of AI, causing unintended outcomes, engaging in deceptive practices like algorithmic trading, and engaging in activities like figuring out who belongs to a certain group (membership inference attacks). On the flip side, it also identifies four ways people can misuse AI: manipulating others through social engineering, spreading fake news and misinformation, hacking into systems, and even using AI for autonomous weapon systems. Table 4.1 provides a condensed yet comprehensive overview of the malevolent exploitation of artificial intelligence. In Table 4.2, a synthesized summary is presented, shedding light on the deliberate and harmful applications of artificial intelligence. The paper organizes these threats in a structured way, helping us better understand how to address them by refining our rules, policies, and actions.

Table 4.1: Summary of Malicious Abuse of AI.

Integrity Attacks	Adversarial examples, a type of integrity attack, can be used to manipulate ML models, causing the algorithm to make mistakes. Example: Microsoft's Tay.
Unintended AI Outcomes	Algorithms can present an unexpected output due to, for instance, unintentional memorization by models based on neural networks. Example: Gmail's Smart Compose.
Algorithmic Trading	With the increase of algorithmic trading, the stock market is susceptible to high-speed crashes. The incidents can be intentional or accidental. Example: 2010 Flash Crash.
Membership Inference Attacks	Such attacks try to uncover and reconstruct data used to train Machine Learning models. The attacks can target datasets containing, for instance, biometric and genetic data. Example: thispersondoesnotexist.com.

Table 4.2: Summary of Malicious Use of AI.

Deception and Phishing (Social Engineering)	To develop social hots, attackers can use AI techniques, such as natural language processing. The bots arc used to deceive and manipulate people into complying with their requests (e.g., sharing personal information). Example: Cyberlover.
Manipulation (Social Engineering)	Malicious actors can use AI techniques to develop algorithms or social bots to manipulate public opinion. Example: Cambridge Analytica.
Misinformation and Fake News	AI systems can be used to accelerate the creation and spread of unsubstantiated content aimed at misinformation. Example: tools such as GPT-3.
Deepfakes (Hacking)	With the advances in AI, algorithms support the creation of hyper-realistic images and videos, known as deepfakes. Example: "fake" Elon Musk joining Zoom meeting.
Repetitive Tasks (Hacking)	AI systems can perform repetitive tasks efficiently, which malicious actors can exploit. Example: Captcha-defeating and password cracking
Malware (Hacking)	Malware could be enhanced with AI techniques, improving its capabilities. Currently, the possibilities are investigated by academic research and as proof of concept. Example: DeepLocker.
Autonomous Weapons Systems (AWS)	Malicious actors could alter the software embedded in weapons systems, resulting in unintended casualties. Example: hacking the GPS of drones.

Chapter 5

CONCLUSION

Understanding potential threats arising from the use and misuse of AI systems is crucial for devising effective mechanisms to safeguard society and critical infrastructures against malicious activities. Drawing insights from existing literature, reports, and historical incidents, a classification system has been developed that encompasses various dimensions of harm, including physical, psychological, political, and economic impacts. The examination delves into vulnerabilities inherent in AI models, such as unintended consequences, as well as the spectrum of AI-enabled and AI-enhanced attacks, including tactics like forgery. Shedding light on past incidents like the 2010 flash crash and the Cambridge Analytica scandal illustrates the complex challenges faced. Additionally, attacks demonstrated only in a "proof of concept" stage, exemplified by IBM's DeepLocker, are highlighted. To address identified risks, potential mitigation strategies are proposed. Emphasis is placed on collaboration among industries, governments, civil society, and individuals to develop knowledge, raise awareness, and establish technical and operational systems to effectively confront these challenges. While the classification provides a valuable foundation, it is not without limitations. Some AI-driven attacks may not neatly fit into established categories. Future research should consider empirical methods to assess the generalizability and representativeness of the classification scheme. As more data becomes available, statistical analysis could offer a more comprehensive understanding of the threat landscape. Continuously mapping risks associated with malicious use of AI enhances preparedness, offering better prospects for prevention and effective response to potential attacks.

REFERENCES

[1] T. F. Blauth, O. J. Gstrein and A. Zwitter, "Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI," in IEEE Access, vol. 10, pp. 77110-77122, 2022, doi: 10.1109/ACCESS.2022.3191790.

[2] R. Webster, J. Rabin, L. Simon and F. Jurie, "This person (Probably) Exists. Identity membership attacks against GAN generated faces", arXiv:2107.06018, 2021.

[3] T. Yigitcanlar, K. Desouza, L. Butler, and F. Roozkhosh, "Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature," Energies, vol. 13,no.6,p. 1473, Mar. 2020, doi: 10.3390/en13061473.

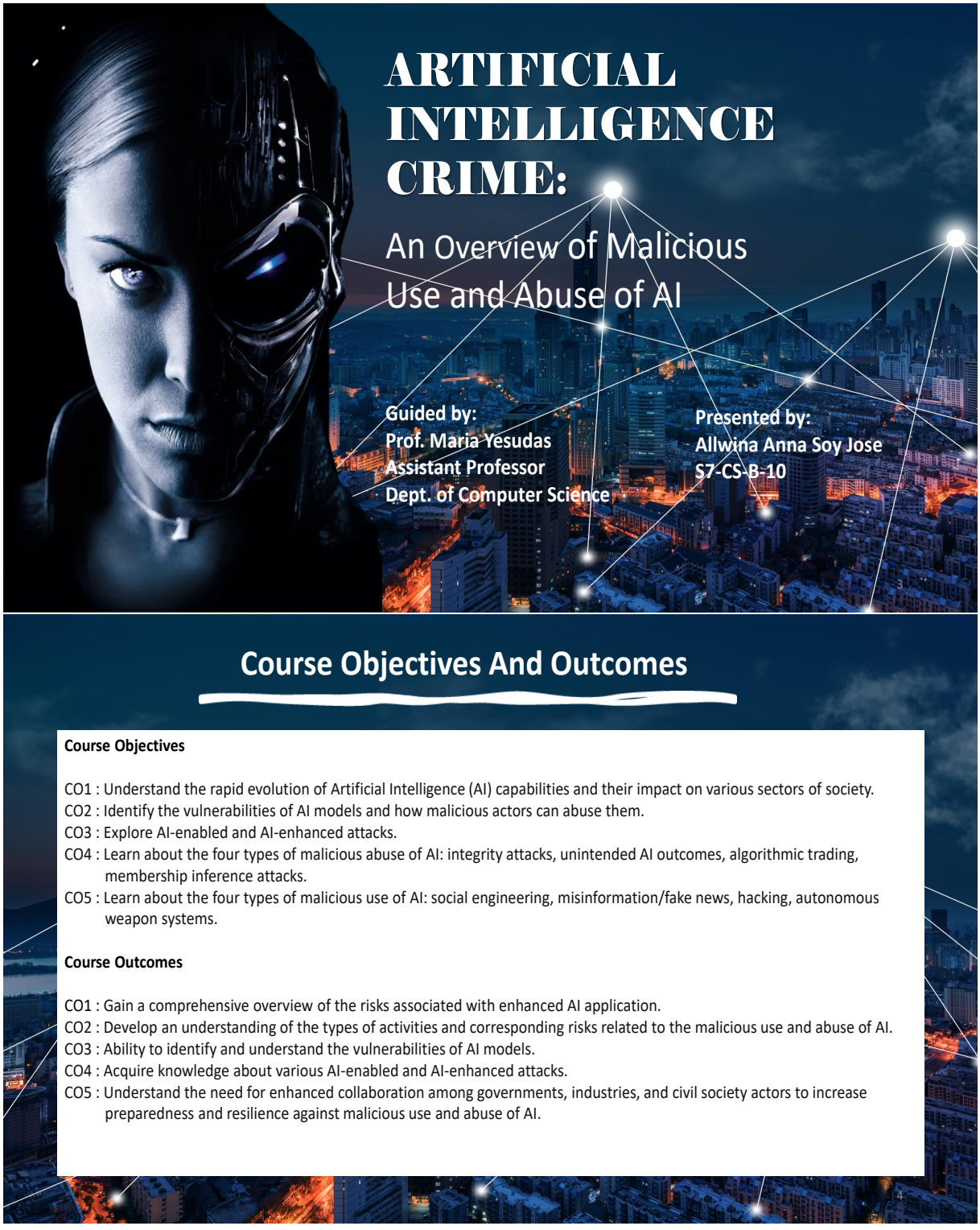
[4] D. Patel, Y. Shah, N. Thakkar, K. Shah and M. Shah, "Implementation of artificial intelligence techniques for cancer detection", Augmented Hum. Res., vol. 5, no. 1, Dec. 2020.

[5] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," 2018, arXiv:1802.08232.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world,"2016,arXiv:1607.02533.

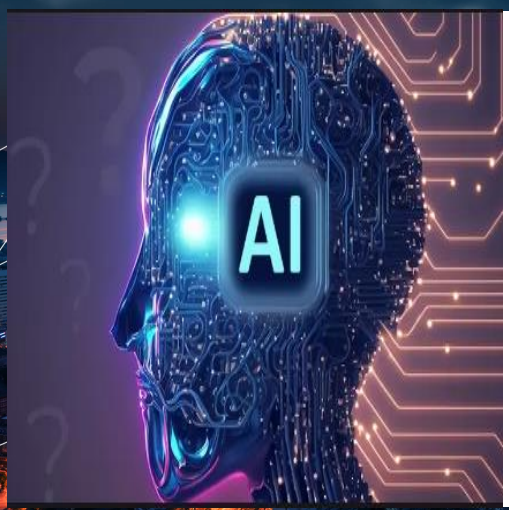
[7] J. Seymour and P. Tully. (2016). Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter.

ANNEXURE


ARTIFICIAL INTELLIGENCE CRIME:
An Overview of Malicious Use and Abuse of AI
Guided by:
Prof. Maria Yesudas
Assistant Professor
Dept. of Computer Science
Presented by:
Allwina Anna Soy Jose
S7-CS-B-10
Course Objectives And Outcomes
Course Objectives
CO1 : Understand the rapid evolution of Artificial Intelligence (AI) capabilities and their impact on various sectors of society.
CO2 : Identify the vulnerabilities of AI models and how malicious actors can abuse them.
CO3 : Explore AI-enabled and AI-enhanced attacks.
CO4 : Learn about the four types of malicious abuse of AI: integrity attacks, unintended AI outcomes, algorithmic trading, membership inference attacks.
CO5 : Learn about the four types of malicious use of AI: social engineering, misinformation/fake news, hacking, autonomous weapon systems.
Course Outcomes
CO1 : Gain a comprehensive overview of the risks associated with enhanced AI application.
CO2 : Develop an understanding of the types of activities and corresponding risks related to the malicious use and abuse of AI.
CO3 : Ability to identify and understand the vulnerabilities of AI models.
CO4 : Acquire knowledge about various AI-enabled and AI-enhanced attacks.
CO5 : Understand the need for enhanced collaboration among governments, industries, and civil society actors to increase preparedness and resilience against malicious use and abuse of AI.

Content:

What is Artificial Intelligence ?



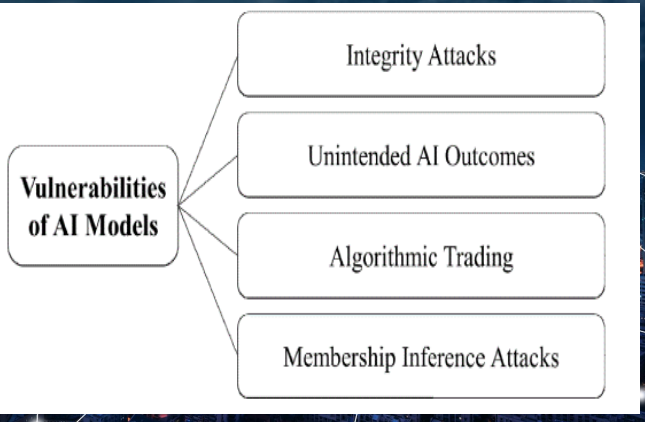
- Artificial Intelligence (AI) is the emulation of human intelligence in machines.
- It encompasses a wide range of technologies and techniques for tasks like problem-solving, learning, and decision-making.
- AI subfields include machine learning, natural language processing, computer vision, and robotics.
- AI is used in various industries for applications such as healthcare, finance, autonomous vehicles, and virtual personal assistants.

What is Artificial Intelligence Crime?

- "Artificial Intelligence crime" involves illegal activities using AI technology.
- It includes cybercrimes, hacking, automated fraud, and unethical AI practices.
- Discrimination concerns may also fall under this category.
- Law enforcement and regulators are taking action to combat AI-related criminal activities and promote ethical AI use.

7

Malicious Abuse of AI



This usually refers to scenarios where existing AI systems or models are manipulated or exploited in a way that causes harm.

8

Integrity Attacks

- Integrity attacks in AI crime involve malicious actions by hackers.
- The goal is to inject false information into a system to corrupt or manipulate data.
- These attacks can seriously undermine the trustworthiness and reliability of AI systems.
- Machine Learning (ML) models, in particular, are vulnerable to integrity attacks as they heavily rely on the quality and integrity of their training data.

Eg: Tay, Microsoft's Chatbot

9

Unintended AI Outcomes

- Unintended AI outcomes are situations where AI systems behave in ways not intended by their designers
- These outcomes can pose serious threats to privacy and security.
- It's crucial to understand that these unintended outcomes are not a result of the AI system's intent but rather issues related to design, implementation, or environmental interactions.

Eg: Google's Smart Compose

10

Algorithmic Trading

- It involves the misuse of AI to execute large orders of trades, manipulate market prices, or exploit information asymmetry.
- In a simulated market experiment, trading agents could learn and execute a “profitable” market manipulation campaign comprising a set of deceitful false-orders.
- AI can bring significant benefits to the financial sector, such as building and training bots that can trade better than humans do, and preventing and detecting everything from routine employee theft to insider trading.
  Eg: 2010 Flash Crash

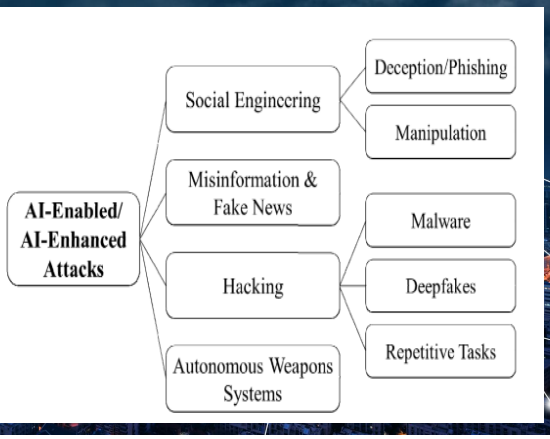
11

Membership Inference attacks

- Adversary queries a trained machine learning model to determine if a specific example was part of the model's training dataset.
- Adversary uses machine learning to train an inference model that detects differences in the target model's predictions between training and non-training inputs.
- This technique leverages adversarial use of machine learning to uncover information about the target model's training dataset.
  Eg:thispersondoesnotexist.com

12

Malicious Use of AI



This typically refers to scenarios where AI is intentionally used to carry out harmful activities

13

Social Engineering

- It uses deception techniques to manipulate human subjects to share sensitive or personal information , which can be used for fraudulent purposes.
- Deception and Phishing : The bots are used to deceive and manipulate people into complying with their requests. E.g. : Cyberlover
- Manipulation : Malicious actors can use AI techniques to develop algorithms or social bots to manipulate public opinion.
  E.g. : Cambridge Analytica

14

Misinformation and Fake News

- Misinformation and Fake News are false or misleading information presented as news.
- They spread rapidly through social media platforms, often becoming viral.
- They can undermine the electoral process, fuel hatred, and stoke outrage.
- Combating them involves fact-checking and raising awareness about their prevalence and impact.
  Eg: Tools such as GPT-3

15

Hacking

- Hacking is the act of identifying and then exploiting weaknesses in a computer system or network, usually to gain unauthorized access to personal or organizational data.
- Deepfakes : With the advances in AI, algorithms support the creation of hyper-realistic images and videos.
  Eg: “fake” Elon Musk joining Zoom meeting.
- Repetitive Tasks : AI systems can perform repetitive tasks efficiently, which malicious actors can exploit.
  Eg: Captcha-defeating and password cracking.
- Malware : It could be enhanced with AI techniques, improving its capibilities.
  Eg: DeepLocker

16

Autonomous Weapons Systems

- Autonomous Weapons Systems are weapons that select and apply force to targets without human intervention.
- They are triggered by sensors and software, which match what the sensors detect in the environment against a 'target profile'.
- The concern with this process is the loss of human judgement in the use of force. It makes it difficult to control the effects of these weapons.
- Eg: hacking the GPS of drones

17

Conclusion

- The capabilities of Artificial Intelligence (AI) evolve rapidly and affect almost all sectors of society[1].
- AI has been increasingly integrated into criminal and harmful activities, expanding existing vulnerabilities, and introducing new threats[1].
- The article provides an overview of the risks of enhanced AI application, contributing to the growing body of knowledge on the issue[1].
- It suggests four types of malicious abuse of AI (integrity attacks, unintended AI outcomes, algorithmic trading, membership inference attacks) and four types of malicious use of AI (social engineering, misinformation/fake news, hacking, autonomous weapon systems)[1].
- Enhanced collaboration among governments, industries, and civil society actors is vital to increase preparedness and resilience against malicious use and abuse of AI[1].

18

References

[1] K. Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. London, U.K.: Yale Univ. Press, 2021.

[2] D. Garcia, ‘‘Lethal artificial intelligence and change: The future of international peace and security,’’ Int. Stud. Rev., vol. 20, no. 2, pp. 334–341, Jun. 2018, doi: 10.1093/isr/viy029.

[3] T. Yigitcanlar, K. Desouza, L. Butler, and F. Roozkhosh, ‘‘Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature,’’ Energies, vol. 13, no. 6, p. 1473, Mar. 2020, doi: 10.3390/en13061473.

[4] Cybercrime. United Nations: Office Drugs. Accessed: May 19, 2021. http://www.unodc.org/unodc/en/cybercrime/index.html

[5] O. Osoba and W. Welser IV, The Risks of Artificial Intelligence to Security and the Future of Work. Santa Monica, CA, USA: RAND Corporation, 2017, doi: 10.7249/PE237.

19

THANK YOU

20

VISION & MISSION OF THE DEPARTMENT

VISION

To evolve as a school of computing with globally reputed Centre's of excellence and serve the changing needs of the industry and society.

MISSION

- The department is committed in bringing out career-oriented graduates who are industry ready through innovative practices of teaching and learning process
- To nurture professional approach, leadership qualities and moral values to the graduates by organizing various programs periodically
- To acquire self-sustainability and serve the society through research and consultancy