World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 7.029

ARTIFICIAL INTELLIGENCE CRIME: EXPLORING MALICIOUS EXPLOITS AND RISKS

E. Sree Shanmitha[1]*, Kommineni Venkata Naga Hyma[1], Palle Venu Madhav[1] and Banothu Ramji[2]

[1]Student (S) from Department of Computer Science and Engineering (Data Science), CMR Technical Campus, Kandlakoya (V), Medchal Road, Hyderabad – 501401, Telangana, India.

[2]Assistant Professor from Department of Computer Science and Engineering (Data Science), CMR Technical Campus, Kandlakoya (V), Medchal Road, Hyderabad – 501401, Telangana, India.

Article Received on 16/02/2024 **Article Revised on 06/03/2024** **Article Accepted on 26/03/2024**

***Corresponding Author**
E. Sree Shanmitha
Student (S) from Department of Computer Science and Engineering (Data Science), CMR Technical Campus, Kandlakoya (V), Medchal Road, Hyderabad – 501401, Telangana, India.

ABSTRACT

The rapid evolution of Artificial Intelligence (AI) has impacted nearly every sector of society, with its integration into criminal activities posing new and expanded threats. This article reviews pertinent literature, reports, and incidents to establish a typology of malicious AI use and abuse. Our aim is to delineate various activities and associated risks, starting with an examination of AI model vulnerabilities and how they can be exploited by malicious actors. Subsequently, we explore both AI-enabled and AI-enhanced attacks, providing a comprehensive overview without striving for exhaustive classification. We propose four categories of malicious AI abuse: integrity attacks, unintended AI outcomes, algorithmic trading, and membership inference attacks. Additionally, we identify four types of malicious AI use: social engineering, dissemination of misinformation/fake news, hacking, and thedevelopment of autonomous weapon systems. By mapping these threats, we facilitate a deeper understanding of governance strategies, policies, and activities necessary to mitigate risks and prevent harm. Effective collaboration among governments, industries, and civil society actors is crucial to enhance preparedness and resilience against the malicious use and abuse of AI. This article contributes to the expanding

knowledge base on this issue and underscores the importance of proactive measures to address emerging threats in AI security.

KEYWORDS: Artificial intelligence, artificial intelligence typology, computer crime, malicious artificial intelligence, security, social implications of technology.

INTRODUCTION

The impact of systems using Artificial Intelligence (AI) is at the centre of numerous academic studies, political debates, and reports of civil society organizations. The development of AI has become the subject of praise due to unprecedented technological capabilities, such as enhanced possibilities for automated image recognition (e.g., detection of cancer in the field of medicine. However, it has also been criticized - even feared - due to aspects such as the uncertain consequences of automation for the labour market (e.g., concerns of mass unemployment. This duality of positive *vs* negative aspects of the technology can also be identified in the context of cybersecurity and cybercrime.

Governments use AI to enhance their capabilities, whereas the same technology can be used for attacks against them.

While the recent surge in AI development has been fuelled by the private sector and applications in customer-oriented applications, sectors such as defence might use similar capabilities in their operations. At the same time, it is increasingly difficult to distinguish between the actions of state and non-state actors. This has recently been demonstrated by a wave of ransomware attacks targeting public infrastructure in many countries, such as the Colonial Pipeline in the United States in May 2021. Additionally, programs and applications developed for non-malicious purposes can also be implemented or modified for malicious intent and potentially cause harm. The dual-use aspect of technology is not an entirely new problem when it comes to cybercrime1 or (cyber-) security. Nevertheless, how AI can be leveraged for malicious use and abuse constitutes novel vulnerabilities. Permanent assessment of the threat landscape is crucial to create and adapt governance mechanisms, develop proactive measures, and enhance (cyber-) resilience. To build on previous work and expand the understanding of how AI broadens the potential for malicious activities online, this article evaluates the main categories of use and abuse of AI in a criminal context. We provide several salient examples that allow us to illustrate the challenges at hand.

Based on these examples, we present a typology that catalogues the main harmful AI-based activities. Developing knowledge and understanding about the potential malicious use and abuse of AI enables cybersecurity organizations and governmental agencies to anticipate such incidents and increase their preparedness against attacks. Furthermore, a typology is greatly useful in structuring research efforts and identifying gaps in knowledge in areas where more research is warranted.

A. Malicious Abuse of AI: Vulnerabilities of AI Models

1) Integrity Attacks

Machine learning (ML) has become more prevalent in recent years. This has created incentives for attackers to manipulate models (e.g., the software itself) or the underlying data, making ML models prone to integrity attacks. In integrity attacks, hackers attempt to inject false information into a system to corrupt the data, undermining their trustworthiness. One of the risks associated with the vulnerability of AI models is the creation of 'adversarial examples'. According to "adversarial examples are malicious inputs designed to fool machine learning models" which causes misclassification of material scrutinized by the systems. In some cases, the perturbations are too subtle to be perceived by human observers, but they still cause AI systems to make mistakes.

One example of an adversarial ML is a 'poisoning attack'. The attacker influences the training dataof the system to alter the results of a predictive model by injecting a few corrupted points in the training process. In other words, poisonous samples can be injected into the training data to manipulate the classifier, leading to undesirable consequences. A concrete example is the attack on Tay, Microsoft's AI chatbot, which was released in 2016. The chatbot had the objective of creating tweets that could not be distinguished from a human actor. Within a few hours of release, users launched a coordinated attack in which they tweeted offensive words and phrases, exploringTay's "repeat after me" function.

This led the bot to reproduce similarly objectionable content. According to the Corporate Vice- President of Microsoft, "although we had prepared for many types of abuses of the system, we hadmade a critical oversight for this specific attack."

Consequently, after less than 16 hours, Microsoft had to suspend the account. This demonstrates that defending a chatbot against attacks is challenging, especially when the system is trained in online environments with unforeseeable live interactions.

Researchers at New York University (NYU) explored another risk associated with the context of outsourced training data. They demonstrated that an adversary might create a Bad Net (a maliciously trained network), which displays conventional behaviour until a potential attacker triggers an attack. To test this hypothesis, Bad Nets were implemented in a complex traffic sign detection system. They demonstrated that a stop sign could be correctly identified by a self-driving car until a stop sign with a pre-defined trigger (yellow 'Post-It' note) was presented. This study demonstrates that AI models might be susceptible to data poisoning and adversarial examples, resulting in misclassifications and errors with potentially grave consequences that are difficult to foresee for humans unfamiliar with the technology. This might be one of the reasons why the recently proposed EU AI Act entails specific requirements for training data of 'high-risk systems' in Article.

2) Unintended Outcomes of the Use of AI

Models used to train AI systems can present a different result from what was expected by the developer for various reasons. For instance, models based on neural networks may unintentionally memorize and disclose details. This can be problematic, especially when the data used to train the models are private or sensitive. explained the phenomenon: during the learning process, such models might memorize details unrelated to the primary task. To prevent harmful consequences from unintended memorization and disclosure of information by the algorithm, it is necessary to apply techniques that guarantee data privacy.

The team behind the development of Smart Compose, the real-time suggestion system used by Google's Gmail service, considered this carefully. To avoid unintended memorization, they conducted "extensive testing to make sure that only common phrases used by multiple users are memorized".

Their goal was to prevent the models from learning details (e.g., private information) that were not related to the primary task (e.g. general and commonly used phrases) while training the algorithm. For example, when a user enters a text prefix such as "my ID number is", the model should not suggest a text completion with the ID number of another user see). This challenge serves as one example in which the developer does not have the malicious intent of disclosing the user's personal information; the potential harm resides in the possibility that the model performs differently than previously expected (i.e., by memorizing private data).

3) Algorithmic Trading/Stock Market Manipulation

With the help of computers and AI-powered software programs, technology facilitates and accelerates the pace of financial analysis and decisions. The use of AI systems in market trading, which causes it to move "with lightning speed" has both positive and negative aspects. In terms of positive aspects, the current financial technology has, for instance, decreased transactional charges and costs of capital for businesses However, algorithmic trading with decisions that are difficult to follow for humans inserts instability into the market. As a result, a risk for high-speed crashes (i.e., flash crashes) emerges. David Weild IV, the former vice-chairperson of Nasdaq, bluntly argued that "we've created a stock market that moves too darn fast for human beings", which is the reason "we see shocking results".

The challenges of automated decision-making in the financial sector became apparent after the 2010 flash crash, which caused a loss of almost $ 1 trillion. Navinder Singh Sarao, a high-frequency trader, was sentenced in 2020 to a year of home incarceration for his involvement in this incident. Sarao was accused of using an automated program to create large sell orders to push down prices. Once the prices dropped, he canceled orders to buy at lower market prices to get thebenefits when the market recovered.

This first market crash in the era of algorithmic trading served as "a wake-up call" not only to traders but also to regulators, showing some of the challenges of high-speed automated trading andautomated-decision making more generally.

To prevent similar incidents in the future, some techniques used to manipulate high-frequency trading, such as spoofing and layering, were banned.

The discussions surrounding the development of regulatory frameworks usually focus on market harm caused by malicious actors . Even though this is a necessary evaluation, it is also important to consider what could be done in the case of a technological accident or insufficient testing. As trading on stock markets becomes increasingly driven by algorithms, investors could face similar flash crashes more often. In such an environment, things can change and "get out of hand in seconds". Among the potential policy responses to flash crashes is the creation of insurance systems. suggests that a financial market fund named the "National Protection Fund", which would compensate the investor eventually harmed by market disruptions caused by algorithms, could be a way of guaranteeing more stability and safety in trading. In addition, strengthening cybersecurity and an in-depth assessment of the

respective algorithms could help to prevent the harmful consequences of high-speed crashes.

4) Membership Inference Attacks

In membership inference attacks, the malicious actor aims to uncover and reconstruct the samplesused to train a ML model. These attacks can be effective on several systems, such as classification and sequence-to-sequence models . They can also be used against generative adversarial networks (GANs). GANs are a class of deep-learning model that creates seemingly realistic - but fake - examples of the data used in the training process. This technique is used in different applications.

In a recent study ,demonstrated that the faces produced by the "This person does not exist" algorithm are quite similar to the faces of the individuals that were part of the training data.

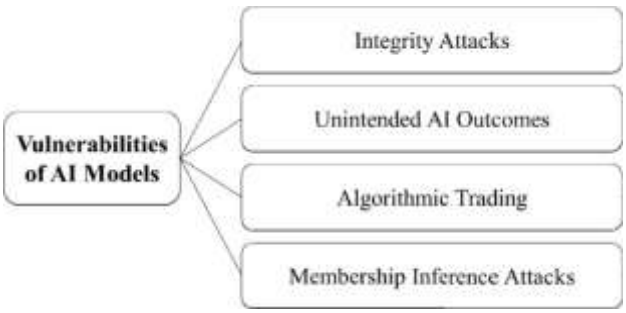

Figure 1: Vulnerabilities of AI.

Table 1: Summary of Vulnerabilities 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 incidenrs can be intentionalor accidental
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

B. Malicious Use of AI: AI-Enabled and AI-Enhanced Attacks

1) Social Engineering

Social engineering attacks use deception techniques to manipulate human subjects to share sensitive or personal information, which can be used for fraudulent purposes. Such attacks are performed in different ways using an array of AI techniques. Using these

techniques, cybercriminals can create elegant manipulation tactics, consequently increasing their chances of success and gains.

1.1) Deception and Phishing

Hackers can use AI techniques to develop a 'social bot', which can help them deceive and manipulate a person into complying with their request. These 'social bots' are algorithms designedto emulate human behaviour by producing content and interacting with users on the internet. Forinstance, the request of social bots can access a website that enables the criminal to take over the computer of the victim. One of the first known cyberattacks that used AI techniques was a datingchatbot known as 'Cyberlover'. It was released in 2007 to lure users of chat rooms into sharing personal information or click on fraudulent links. The bot used natural language processing (NLP) to deliver a customized dialog, which raised concerns about the capabilities being used incybercrime.

Similarly, attackers can masquerade themselves as trusted individuals or companies to induce thevictim to open an email or link to steal data. The technique, known as phishing, can also be enhanced by AI to maximize the reach and gain of criminals.

This was demonstrated by, who conducted an experiment using a model based on machine learning techniques to generate text to be posted on Twitter. The authors chose this social media platform because of the character limitation of each tweet, which makes posts with broken Englishand shortened links to be considered acceptable and normal. The results show that the dynamics of such platforms may facilitate the use of machine-generated text for phishing.

AI may enable growth in these types of attacks in social media because posts tend to be written in an informal tone, with occasional spelling and grammar mistakes, and with shortened links.

1.2) Big Nudging and Manipulation

In addition to the potential targeted action described in the previous section, large numbers of botsmight be created to support actions with malicious intent. Bots can potentially influence public opinion and the outcome of elections. For instance, by retweeting specific content or replicating hashtags, social bots can be used to create the impression that a candidate or political movement is more popular, deceiving users on social media platforms. A similar

strategy is astroturfing, a process that mimics a bottom-up activity to create the impression that a policy or individual has widespread grassroots support when little or no support exists . An example of this is when a given organization is responsible for publishing thousands of Twitter posts using different accounts to influence public opinion against or in favor of a candidate in an election. Astroturfing can be found in Twitter posts, blogs, news portals, and other online platforms, and they can be used as disinformation strategies.

Bots can also be used to create the perception of support for a cause in public consultations and interfere with polls. Concerns over this possibility spiked after the Federal Communications Commission's (FCC) consultation on net neutrality in the United States. As the FCC had plans toroll back net neutrality protections, the regulator opened a consultation to gather public opinion on the topic through a comment section. The data analytics company Gravwell identified that, out of the approximately 22 million comments received by FCC, more than 80% were submitted by bots. In this case, natural language generation was used to artificially inflate the support against net neutrality protection.

Another use of AI in this context is online profiling and targeting. The Cambridge Analytica scandal exemplifies this. According to reports and whistle-blowers, the app GSRApp was used to deceptively collect the personal data of their users, including personality traits, which were later used to train an algorithm.

This algorithm generated personality scores for app users and their Facebook friends, which werethen matched with the US elector records. Cambridge Analytica used the resulting data to develop voter profiling and targeted advertising services. With such information, politics could target specific groups of people by manipulating messages tailored to their psychological profile, in addition to disinformation and inflammatory material. Using these tools to change the behavior of individuals through manipulation can impact democratic processes and election outcomes.

2) Misinformation and Fake News

The development and diffusion of technology, blogging platforms, and social media have changedthe way individuals consume information, access news items, and form opinions. The fast pace of the Internet also enables anyone to create and rapidly share content, which can reach many people. This scenario has created an environment that allows the creation and spread of misinformation and fake news. Although the term "fake news" is contested by

some journalists and academics it is still relevant to promote debates on digital literacy and encourage scholarly work on the issue. Moreover, the justification behind the call for a ban has been demonstrated to be insufficient for abandoning the term.

Unsubstantiated rumours, speculation, and deliberately false information can lead to disastrous consequences, especially in times of uncertainty and social unrest, such as endemics and pandemics. During political events such as elections, it can also be harmful. AI systems can fuel the creation and spread of this type of content, which represents a risk to society and democratic processes, potentially even democracy as such.

Tools such as GPT-3 could boost the creation of written pieces aimed at misinformation. GPT-3 is an autoregressive language model that uses deep learning to complete tasks such as question- answering, text completion, and summarization. Due to format, choice of words, and consistency,texts created automatically with the tool might look like they were written by a human, misleadingthe reader due to apparent credibility.

Some examples of this can be seen on the website "NotRealNews.net", which uses AI to generate AI-written fake news pieces. The idea behind the project was to demonstrate how this tool can be used to support the work of journalists. Considering that the articles were mostly convincing, such a tool could easily be used to disseminate compelling fake news articles. This means that automatically generated texts, coupled with current targeting capabilities, could further increase the quantity, quality, and impact of fake news and disinformation campaigns. These might impact democratic processes to a greater (e.g., by convincing electors to change their vote) or to a lesser degree (e.g., by confirming or reinforcing electors' pre-existing views). In addition, as technology evolves, texts can be tailored to the audience's taste, increasing the proliferation of "filter bubbles"and polarization.

Some strategies could help to reduce the negative impact of the use of AI systems to create and disseminate fake news and misinformation. conducted a study that revealed that information literacy increases the likelihood of identifying fake news pieces.

According to 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". For this

reason, educating individuals about the adequate use of digital resources is of paramount importance. Following thislogic, the more citizens can navigate the online environment and critically evaluate the information, the less unfounded stories will impact them and their community.

3) Hacking

3.1 Forgery: Deepfakes

Prominent examples of forgery in the digital age are deepfake videos and images. Such hyper- realistic media may apply AI in its creation to portray a person saying or doing things that did nothappen. The use of AI for the forgery of videos and images enables more realistic material, making it difficult to distinguish between what is real and what is fake.

Although such manipulation is not new, especially after the popularization of programs such as Photoshop, AI makes forgery more elaborate and challenging to detect. For instance, Ali Aliev developed a method for creating deepfakes in real time. To test the tool, the programmer joined arandom Zoom meeting pretending to be Elon Musk. This example goes along with the current practice of mostly using the figures of well-known individuals, such as celebrities and politicians, in deepfake materials. The danger of these videos and images resides in the fact that they can be created for several malicious purposes: propaganda, disinformation, bullying, revenge porn, or blackmail to name just a few.

The malicious use of forged videos can have a direct impact on politics and international relations.The Democratic Party in the United States created a fake video of the chairman at a convention tohighlight their concern for the effect of deepfakes in democratic processes.

One of the alternatives to reduce the negative consequences of the use of forged videos is to raiseawareness of the population about such technology use.

Bruno Sartori, a deepfake creator, produces humorous videos depicting Brazilian national politics, especially involving politicians from the executive branch. Adding a level of absurdity in the videos, viewers understand that they are not real and the material produced constitutes an elaborate satire. More importantly, the material shared on social media platforms serves to demonstrate therisks of the technology to the public. Inoculation theory helps explain such interventions. According to this theory, prior exposure can help protect individuals against future threats.

In the context of deepfakes, by offering knowledge about the technology and convincing the population to interpret videos critically, such initiatives might help individuals to be "inoculated" against maliciously forged videos. In addition to raising awareness, it is important to further develop tools for deepfake detection. AI techniques can be particularly helpful, such as the use ofrecurrent neural networks.

Describe a phenomenon known as the "liar's dividend", which adds a layer of complexity to the problem. According to the authors, liar's dividend refers to the situation in which someone, a 'liar', takes advantage of the existence of deepfake videos to discredit a real video. This person would claim that the material was manipulated, creating doubt about its authenticity among the public. The more the public is aware of the use of AI to doctor videos, the more skeptical they will be, questioning videos and images that are, in fact, real. This is what the authors called the liar's dividend: "this dividend flows, perversely, in proportion to success in educating the public about the dangers of deep fakes . Therefore, there is a possibility that, during elections, a candidate thatwas caught on tape might lie about the video, saying it is a deepfake, convincing electors of their innocence. At this point, it remains to be seen whether and how regulations such as the EU AI Actwill be able to address deepfakes. In the current draft, no dedicated prohibition is visible. However, the People's Republic of China is introducing relevant legislation that will require platform operators to prevent the spread of deepfakes on their networks.

3.2 Repetitive Tasks

AI is also efficient in conducting repetitive tasks that can be used maliciously. One example is the incident involving the company Ticketmaster. AI tools were employed to bypass Captcha[3], whichenabled the purchase of thousands of tickets that would later be resold to generate profit. Pattern recognition is not a problem limited to Captcha-defeating purposes. Concerns about other hacking- based crimes, such as password-cracking, should also be considered.

One way to crack passwords is through brute force attacks, which can be time and resource consuming. However, it has been demonstrated that brute-force attacks using AI have a significantly higher success rate than non-AI based attacks. In other words, the advances in AI could lead to repetitive tasks being used for malicious purposes, such as password cracking.

3.3 Malware

Malware threats have been used for several decades. Creeper Worm, the first documented malicious software, appeared in the 1970s. Since then, these attacks have become a massive industry that is now a significant cybersecurity concern. The AV-TEST Institute registers more than 350,000 new malware and potentially unwanted applications (PUA) per day. This means that four new malware or PUAs are registered every second. As malware developers continue to innovate and create more elaborate malicious programs, it becomes challenging to establish proper and timely defense mechanisms. Currently, concerns revolve around the possibility of AI techniques being used to create more effective and difficult to detect malware. However, to the best of our knowledge, this technology is not yet well developed.

The current possibilities are mainly explored by academic research and as proof of concept by companies. For instance, IBM presented DeepLocker at the Black Hat USA 2018. This system enhances malware with AI and improves its evasion capabilities. DeepLocker explores the lack of explicability of AI systems, which is mainly considered a weakness of AI, to its advantage. It uses a deep neural network to select targets and conceal the intent until it reaches the desired destination. The main risk of this type of AI-enhanced malware is that it can infect many systems without being detected. In addition, the capabilities of developing systems such as DeepLocker are not constrained to states; civilians and private organizations can also work on the development of such high-risk malware.

Thus, even if AI-enabled or AI-enhanced malware are not well developed now, the potential risks associated with such a possibility need to be considered.

One way of addressing the challenges of AI-based or AI-enhanced malware is to improve capabilities in the field of cyber autonomy. The feasibility of cyber autonomy was demonstrated during the Cyber Grand Challenge, hosted by the Defence Advanced Research Projects Agency (DARPA) in 2016.

The finalist teams of the competition were asked to "develop automated cyber defence systems that can self-discover, prove, and correct software vulnerabilities at real-time". During the competition, the systems were able to auto-detect and correct. In addition, they were able to attack the software of other participants in their network. According to since this event, it was possible to identify a movement towards "security automation". This can be considered the first step toward cyber autonomy. Developing capabilities in autonomous

defensive cybersecurity is a way of leveraging AI systems against malicious actors. However, given the dual-use property of the technology, software created for defense can also be used for offensive purposes. To reduce this risk, there needs to be clear regulations around these systems' use and security safeguards.

4) Autonomous Weapons Systems (AWS)

Militaries have been exploring the possibility of autonomy in weapons for some time, practicallysince the inception of AI in the late 1950's. As machines can process data, analyze information, and make decisions in some situations in less time than humans, their use is particularly attractivein the context of defence. While Autonomous Weapons Systems (AWS) promise military and strategic advantages they also come with risks. AWS can be defined as AI systems designed to select (i.e., search for or detect) and engage (i.e., use force against) targets without the need for human control or human action after its activation . Autonomous functions can be applied to different platforms, such as ships or fighter jets.

One of the risks of this emerging technology is the possibility of the software embedded in militaryhardware (e.g., drones) being altered by malicious actors. If a drone is hacked and the GPS location of an attack changed, it would behave according to the new rules set in the software. This could result in unintended casualties due to the target being redirected. Similarly, if the data used to train the systems are poisoned, this could lead to disastrous consequences. In 2014, Reprieve publisheda report demonstrating that drone attacks aimed at killing 41 individuals resulted in the death of approximately 1,147 people, raising questions about the accuracy and precision of 'targeted killing'. Gibson, who led the report, argued that drone strikes are "only as precise as theintelligence that feeds them".

Such high risks associated with attacks on AI systems used in warfare are being discussed in academia civil society and at the government level. However, at present, there are no international regulations regarding the use of AWS.

The implications of the use of AI in warfare were first debated among state parties to the United Nations Convention on Certain Conventional Weapons (CCW). The main purpose of the CCW is"to ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately". Within the CCW, thetopic is mainly discussed through the lens of international humanitarian law. Ethical issues, for instance, play a secondary role.

From 2014 to 2016, annual Informal Meetings of Experts on AWS were held in Geneva. Later, the CCW created a Group of Governmental Experts (GGE) on AWS which is the main forum fordebating autonomous weapons systems at the international level.

Among the possibilities for regulation is the creation of an additional protocol to the existing convention. This would follow previously adopted additional protocols, such as those involving weapons with non-detectable fragments, landmines, incendiary weapons, blinding laser weapons,and explosive remnants of war. However, in the past, negotiations that started in the CCW, such as the one on cluster munitions, were moved outside the CCW due to a lack of consensus.

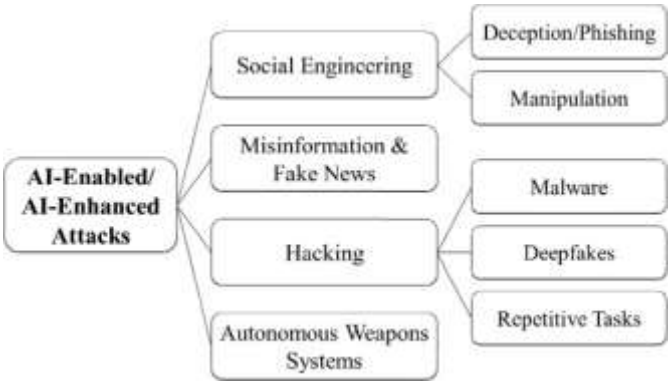

Figure 2: AI-Enabled Attacks.

Table 2: Summary of Malicious use of AI.

Deception and Phishing	To develop social bots, attackers can use AI techniques, suchas natural language processing. The bots are used to deceiveand manipulate people into complying with their requests. Example: Cyberlover.
Manipulation	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 spreadof unsubstantiated content aimed at content aimed at misinformation. Example: tools such as GPT-3
Deepfakes	With the advances in AI, algorithms support the creation ofhyper-realistic images and videos, known as deepfakes.
Repetitive Tasks	AI systems can perform repetitive tasks efficiently, whichmalicious actors can exploit. Example: Password cracking
Malware	Malware could be enhanced with AI techniques, improvingits capabilities. Example: Deep Locker

METHODOLOGY

This study refers to the categorization system of a ‘‘typology’’ rather than a taxonomy. The main difference between typologies and taxonomies involves the research methods used in

theirdevelopment: ‘‘typologies classify subjects by forcing deductive assignment into a priori predefined groups, while taxonomies determine membership into a posteriori categories that emerge from empirical analysis inductively’’.

Therefore, even though the terms taxonomy and typology have been used interchangeably in the literature at times this article refers to the classification scheme of malicious use and abuse of AI as a typology. The methodology is based on an analysis of the available literature on cybercrime and the potential malicious use and abuse of AI systems.

A literature review informs this study and findings using the following databases: IEEE Xplore, Science Direct, Wiley Online Library, and Google Scholar. We used keywords, titles, and screened abstracts. The search terms included are (Artificial Intelligence OR AI OR Machine Learning ORML) AND (malicious OR crime OR harmful OR cyberattack).

Additionally, we examined lists of references obtained from reviewed papers and reports, as well as news sources describing past AI incidents. We only reviewed papers/reports/web pages available in English and Portuguese. After analysing these sources, we were able to identify the different types of malicious use and abuse of AI systems.

With the typology presented in this paper, we hope to make the following contributions

a. Add to the emerging body of knowledge that maps types of malicious use and abuse of AI systems. To understand the main concepts, threat scenarios, and possibilities is necessary to develop much-needed preventive measures and proactive responses to such attacks.

b. Help in establishing a shared language among and across different disciplines, especially between STEM disciplines and legal practitioners, as well as policymakers. Interdisciplinary research on the topic can reduce confusion caused by excessively technical or monodisciplinary language and aid in bridging existing gaps.

c. Propose mitigation strategies, as well as demonstrating that a collective effort among government, academia, and industry is needed.

Implementation and Evaluation

Implementing effective strategies to address artificial intelligence (AI) crime and malicious use/abuse of AI requires a systematic approach, incorporating datasets and typologies to identify,analyze, and mitigate risks. Here's a step-by-step implementation process:

1. Dataset Acquisition and Preparation

- Identify and acquire diverse datasets related to AI crime, including incidents, attacks, vulnerabilities, and threat intelligence reports.
- Ensure datasets cover various domains and types of malicious activities, such as integrity attacks, misinformation, hacking, and autonomous weapon systems.
- Cleanse and preprocess datasets to remove duplicates, irrelevant information, and ensure data consistency and quality.

Dataset

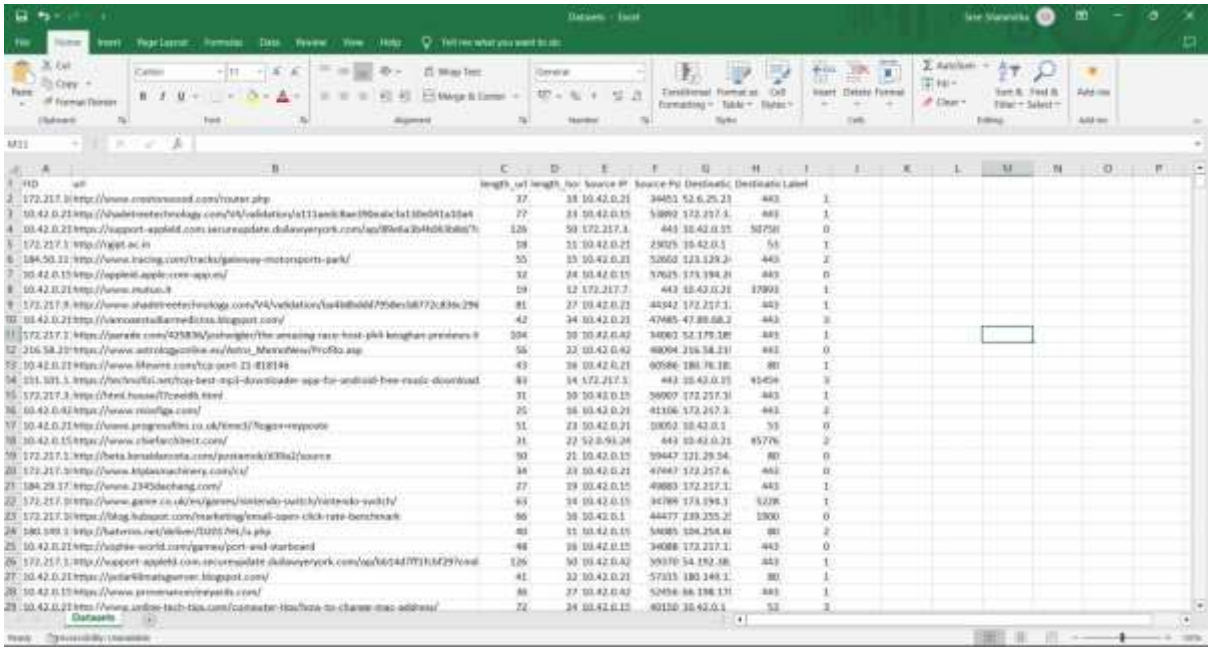

Figure 3: Dataset.

2. Typology Development

- Utilize existing literature, reports, and expert knowledge to develop a comprehensive typology of AI crime and malicious use/abuse of AI.
- Classify malicious activities into distinct categories, such as integrity attacks, social engineering, algorithmic trading manipulation, and autonomous weapon systems.
- Define subtypes within each category to capture nuances and variations in malicious behaviours and tactics.

3. Dataset Annotation and Labelling

- Engage domain experts and analysts to annotate datasets with relevant labels and attributescorresponding to the developed typology.
- Assign appropriate labels to dataset entries to categorize them according to the

identifiedtypology of AI crime.

- Ensure consistency and accuracy in labelling to facilitate subsequent analysis and modellingtasks.

4. Feature Engineering and Selection

- Extract meaningful features from annotated datasets to represent different aspects of AI crimeand malicious activities.
- Consider various types of features, including textual, numerical, temporal, and contextualattributes.
- Apply feature selection techniques to identify the most relevant and discriminative featuresfor subsequent analysis and modeling.

5. Model Development and Training

- Select appropriate machine learning or statistical modeling techniques to develop predictivemodels for detecting and mitigating AI crime and malicious use/abuse of AI.
- Divide annotated datasets into training, validation, and testing subsets for model developmentand evaluation.
- Train models using labelled data to learn patterns and relationships between features andmalicious activities defined in the typology.

6. Model Evaluation and Validation

- Evaluate the performance of trained models using appropriate metrics, such as accuracy, precision, recall, and F1-score.
- Validate models on independent datasets to assess their generalization capability and robustness across different scenarios and data distributions.
- Conduct sensitivity analysis to understand model behavior and identify potential vulnerabilities or biases.

7. Deployment and Integration

- Integrate validated models into operational systems and security infrastructure to enhancedetection and prevention capabilities against AI crime and malicious activities.
- Implement real-time monitoring and alerting mechanisms to detect suspicious behaviors andanomalies indicative of potential threats.
- Continuously update and refine deployed models based on feedback, new data, and emergingthreats to maintain effectiveness and relevance.

8. Continuous Improvement and Adaptation

- Establish mechanisms for ongoing monitoring, evaluation, and improvement of deployedmodels and detection systems.
- Solicit feedback from security analysts, incident responders, and end-users to identify areas forenhancement and refinement.
- Stay abreast of emerging trends, technologies, and threat landscapes to adapt and evolvedetection strategies and countermeasures accordingly.

RESULT

In the preceding sections, we have examined instances of AI misuse through the lens of our definition of malicious use and abuse of AI. This examination has led to the development of a typology that delineates between various types of malicious activities, as summarized in Tables 1and 2.

The category of 'malicious abuse of AI' (Figure 1) encompasses the exploitation of AI vulnerabilities, whether through integrity attacks targeting the learning models or the learning data. Additionally, we have included instances of unintended AI outcomes, exemplified by cases such as Google's Smart Compose, despite their divergence from our primary focus on intentional AI crime, due to their potential for intentional exploitation. Finally, algorithmic trading and membership inference attacks are incorporated within this category.

Within the realm of 'malicious use of AI' (Figure 2), we find AI-enabled and AI-enhanced attacks targeting both physical (e.g., human) and digital (e.g., data infrastructures and computer systems) entities. These attacks can be further subdivided into four categories: (1) social engineering, (2) hacking, (3) misinformation and fake news, and (4) Autonomous Weapons Systems (AWS), as summarized in Table 2.

While the resulting typology provides a comprehensive framework, it is not definitive or exhaustive. Some categories may overlap, while others may emerge as technology continues to evolve. Nevertheless, this structured overview of the current landscape and diverse attack vectorsoffers valuable insights into the burgeoning field of AI crime.

RESULT DATASET

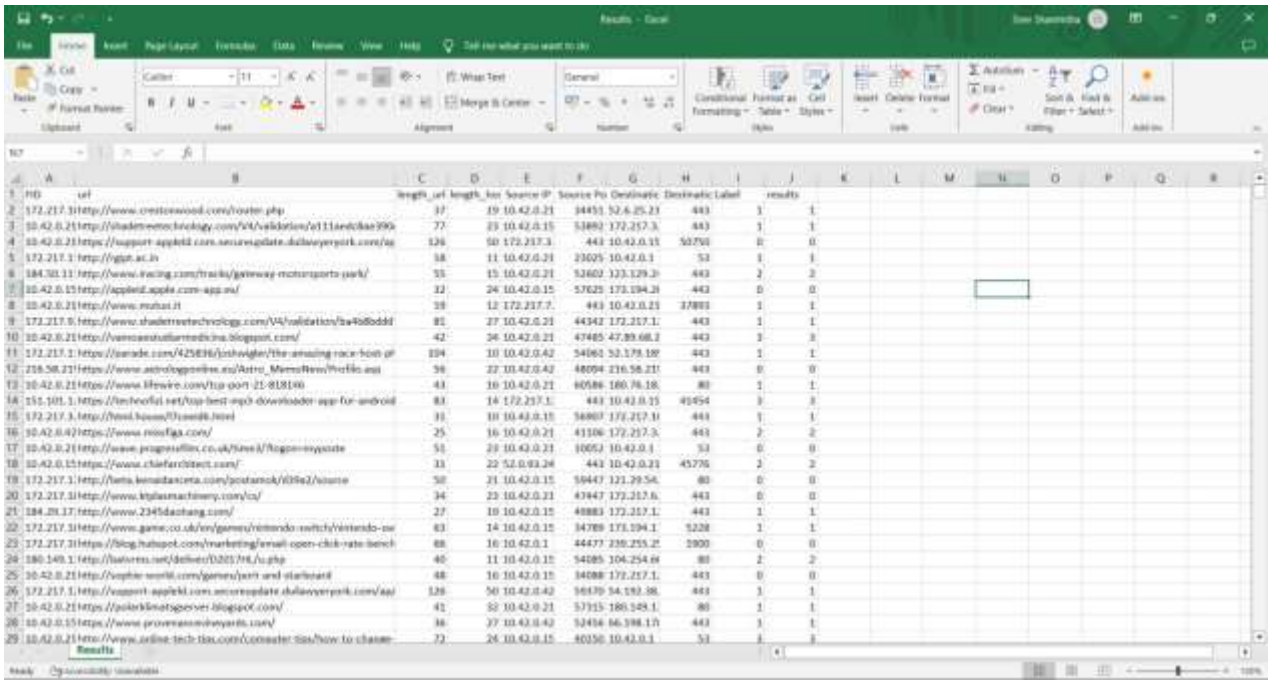

Figure 4: Result Dataset.

Trained and Test Accuracy Results Is Represented In

1) Bar Graph

2) Line Graph

3) Pie Chart.

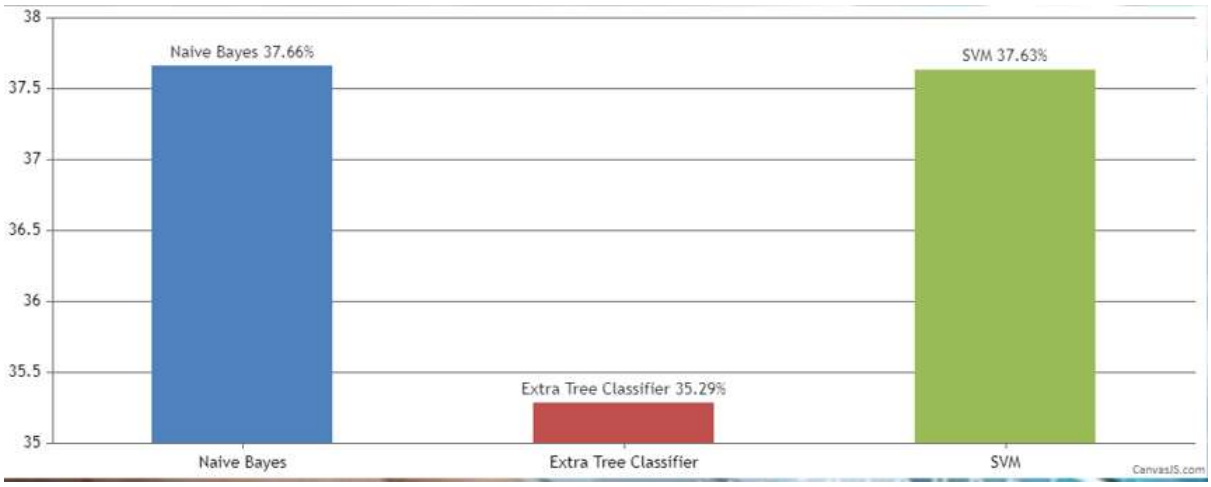

Figure 5: Bar Graph.

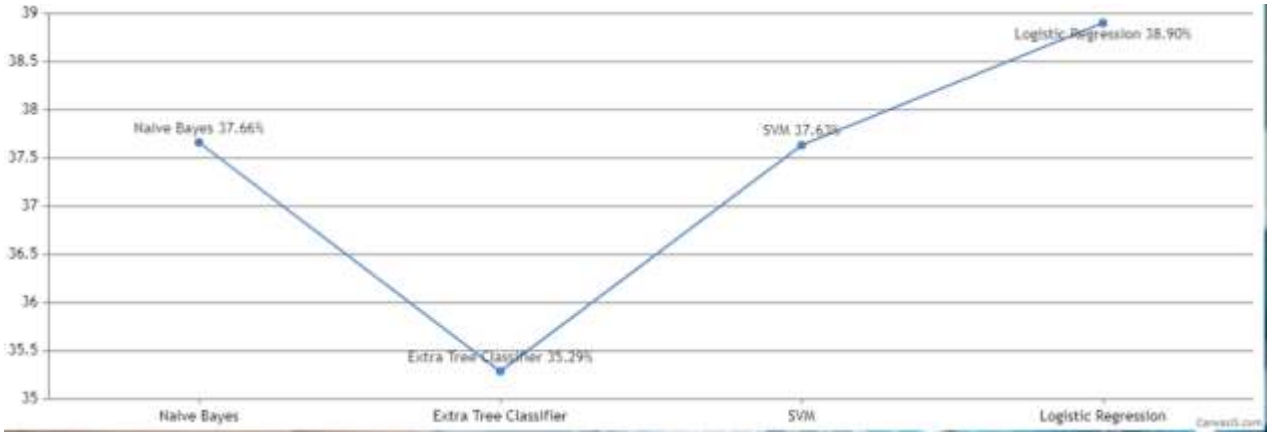

Figure 5: Line Graph.

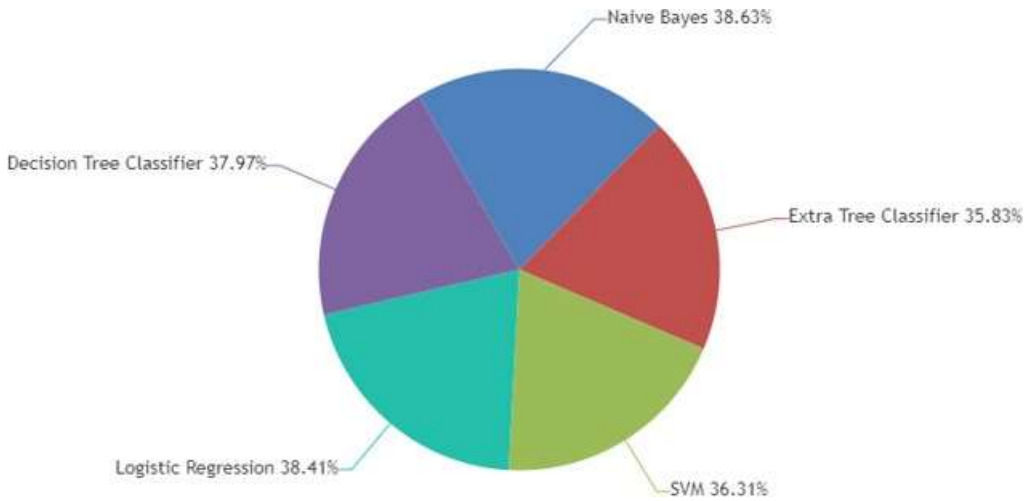

Figure 6: Pie Chart.

Crime Prediction Results

View Crime Prediction Type Details !!!

FID	url	length_url	length_hostname	Source_IP	Source_Port	Destination_IP
10.42.0.42-66.198.178.91-57417-443-6	https://www.missfiga.com/	25	16	10.42.0.211	41106	172.217.3.106
10.42.0.211-52.84.143.74-51243-80-6	http://sophie-world.com/games/port-and-starboard	48	16	10.42.0.151	34088	172.217.12.206
172.217.10.78-10.42.0.151-443-57693-6	http://www.ktplasmachinery.com/cs/	34	23	10.42.0.211	47447	172.217.6.202
172.217.10.74-10.42.0.211-443-33776-6	http://www.game.co.uk/en/games/nintendo-switch/nintendo-switch/	63	14	10.42.0.151	34789	173.194.175.188
172.217.3.97-10.42.0.42-443-40598-6	http://press-preview.weebly.com	31	24	172.217.10.42	443	10.42.0.42

Figure 7.1: Crime Prediction Type Details.

View Crime Type Prediction Type Ratio Details

Crime Type Prediction Type	Ratio
Social Engineering	57.14285714285714
Misinformation	14.285714285714285
Hacking	14.285714285714285
Autonomous weapon systems	14.285714285714285

Figure 7.2: Ratio Details.

Datasets Trained and Tested Results

Model Type	Accuracy
Naive Bayes	38.80139982502187
Extra Tree Classifier	35.30183727034121
SVM	37.79527559055118
Logistic Regression	39.28258967629046
Decision Tree Classifier	37.05161854768154

Figure 7.3: Datasets Results.

Crime Prediction Results In Graphs

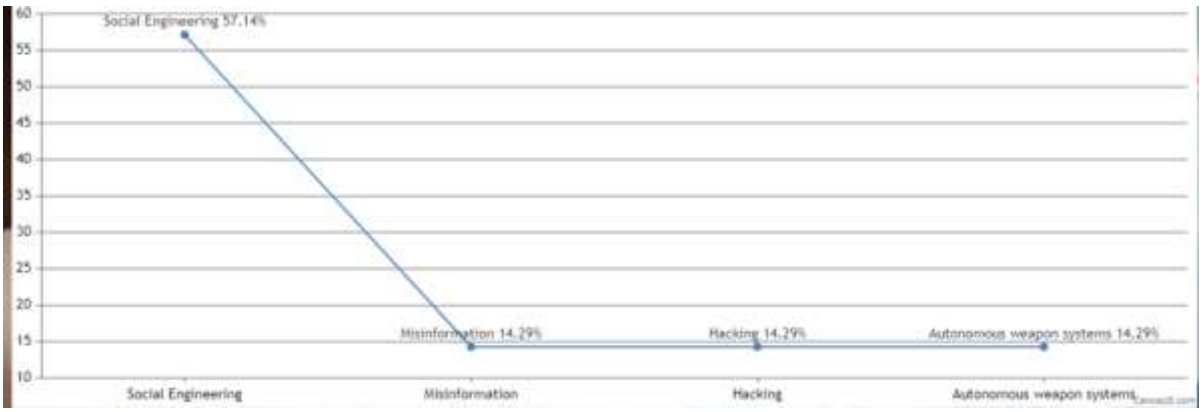

Figure 8: Crime Prediction Results in Line Graph.

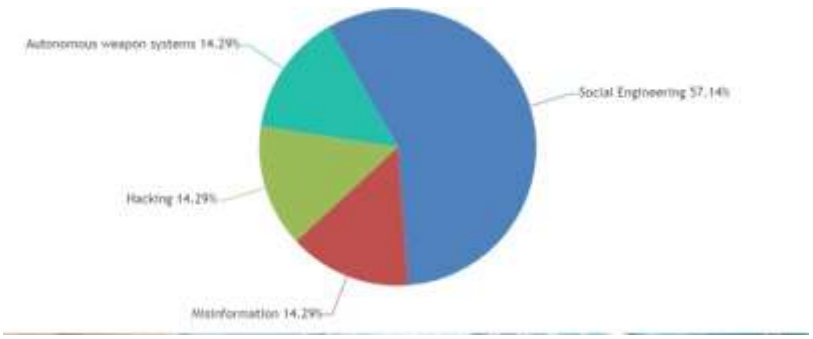

Figure 9: Crime Prediction Results in Pie Chart.

CONCLUSION

The threats posed by the use and abuse of AI systems must be well understood to create mechanisms that protect society and critical infrastructures from attacks. Based on the available literature, reports, and previous incidents, we focused on creating a classification of how AI systems can be used or abused by malicious actors. This includes, but is not limited to, physical, psychological, political, and economic harm. We explored the vulnerabilities of AI models, such as unintended outcomes, and AI-enabled and AI-enhanced attacks, such as forgery.

This article also describes past incidents, such as the 2010 _ash crash and the Cambridge Analytica scandal, manifesting the challenges at hand. We also outlined attacks that, to the best of our knowledge, have only been demonstrated through ``proof of concept", such as IBM's Deep Locker. In response to the risks presented in this paper, we have also exploredsome possible mitigation strategies. Industries, governments, civil society, and individuals should cooperate in developing knowledge and raising awareness while developing technical and operational systems and procedures to address the challenges.

Although this type of classification is a useful starting point, it does not come without drawbacks. Some AI-enabled or AI-enhanced attacks might not fit the categories established.

Further work could use empirical methods to assess whether the classification scheme presented is generalizable and representative. When sufficient data is available, methods such as statistical analysis could be helpful to reach a more complete overview of the threat scenario. Continuouslymapping the risks associated with malicious use and abuse of AI helps to enhance preparedness and increases the potential to prevent and adequately respond to attacks.

REFERENCES

1. K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Arti_cial Intelligence*. London, U.K.: Yale Univ. Press, 2021.
2. D. Garcia, ``Lethal arti_cial intelligence and change: The future of internationalpeace and security," *Int. Stud. Rev.*, Jun., 2018; 20(2): 334-341. doi: 10.1093/isr/viy029.
3. T. Yigitcanlar, K. Desouza, L. Butler, and F. Roozkhosh, ``Contributions and risks of arti_cial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature," *Energies*, Mar, 2020; 13(6): 1473. doi: 10.3390/en13061473.

4. I. van Engelshoven. (Oct. 18, 2019). *Speech by Minister Van Engelshoven on Arti_cial Intelligence at UNESCO, on October the 18th in Paris*. Government of The Netherlands. Accessed: Apr. 15, 2021. [Online]. Available: https://www.government.nl/documents/speeches/2019/10/18/speech-by-minister-van-engelshoven-on-arti_cial-intelligence-atunesco
5. O. Osoba andW.Welser IV, *The Risks of Arti_cial Intelligence to Security and the Future of Work*. Santa Monica, CA, USA: RAND Corporation, 2017. doi:10.7249/PE237.
6. D. Patel, Y. Shah, N. Thakkar, K. Shah, and M. Shah, ``Implementation of arti_cial intelligence techniques for cancer detection," *Augmented Hum. Res.*, Dec. 2020; 5(1). doi: 10.1007/s41133-019-0024-3.
7. A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S. H. Heywang-Köbrunner, I. Sechopoulos, and R. M. Mann, ``Detection of breast cancer with mammography: Effect of an artificial intelligence support system," *Radiology*, Feb. 2019; 290(2): 305-314. doi: 10.1148/radiol.2018181371.
8. J. Furman and R. Seamans, ``AI and the economy," Nat. Bur. Econ. Res., NBER, Cambridge, MA, USA, Work. Paper, 2018, doi:10.3386/w24689.
9. D. R. Coats, *Worldwide Threat Assessment of the U.S. Intelligence Com- munity*. New York, NY, USA, 2017; 32.
10. L. Floridi, ``Soft ethics: Its application to the general data protection regulation and its dual advantage," *Philosophy Technol.*, Jun., 2018; 31(2): 163-167. doi:10.1007/s13347-018-0315-5.
11. P. S. Chauhan and N. Kshetri, ``2021 state of the practice in data privacy and security," *Computer*, vol. 54, no. 8, pp. 125_132, Aug. 2021, doi: 10.1109/MC.2021.3083916.
12. S. Gordon and R. Ford, ``On the de_nition and classi_cation of cybercrime," *J. Comput. Virol.*, Aug. 2006; 2(1): 13-20. doi:10.1007/s11416-006-0015-z.
13. *Cybercrime*. United Nations: Of_ce Drugs. Accessed: May 19, 2021. http://www.unodc.org/unodc/en/cybercrime/index.html
14. M. Brundage, S. Avin, J. Clark, and H. Toner, ``The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018. *arXiv:1802.07228*.
15. T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi, ``Arti_cial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions," *Sci. Eng. Ethics*, Feb., 2020; 26(1): 89-120. doi:10.1007/s11948-018-00081-0.
16. V. Ciancaglini, ``Malicious uses and abuses of arti_cial intelligence," in *Trend Micro Research; United Nations Interregional Crime and Justice Research Institute (UNICRI);*

Europol's European Cybercrime Centre (EC3), Nov., 2020. [Online]. Available: https://www.europol.europa.eu/

17. publications-documents/malicious-uses-and-abuses-of-arti_cialintelligence K. D. Fiedler, V. Grover, and J. T. C. Teng, ``An empirically derived taxonomy of information technology structure and its relationship to organizationalstructure," *J Manage. Inf. Syst.*, Jun., 1996; 13: 9-34. doi: 10.1080/07421222.1996.11518110.
18. N. Bostrom, ``Information hazards: A typology of potential harms fromknowledge," *Rev. Contemp. Philosophy*, May, 2011; 10: 44-79.
19. W. B. Carper and W. E. Snizek, ``The nature and types of organizationaltaxonomies: An overview," *Acad. Manage. Rev.*, Jan., 1980; 5(1): 65-75.
20. (Apr. 21, 2021). *Proposal for a Regulation Laying Down HarmonisedRules on Artificial Intelligence Act*. European Commission. Accessed: May 19, 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulationlaying-down-harmonised-rules-arti_cial-intelligence-arti_cialintelligence

About the Authors

I am E. Sree Shanmitha, currently pursuing my undergraduate degree in Computer Science and Engineering with a specialization in Data Science at CMR Technical Campus in Hyderabad, India. My academic journey is driven by a deep interest in Data Science and Machine Learning. I've spent my time enhancing mytechnical skills in Data Science, Web Development, and Machine Learning.

Notable projects include a Machine Learning Based Heart Disease Prediction system and a Web Based project featuring a scientific calculator. Additionally, I'm developing a website dedicated to mythology, aiming to provide an easy-to-use platform for users to access and engage with various mythological stories. These experiences have given me a strong foundation in both theory and practical application in these fields, and I'm excited to keep exploring innovative solutions in technology.

I (K.V.N.Hyma) am currently pursuing Undergraduate(B. Tech) in CSE(Data Science) from CMR Technical Campus, Hyderabad, Telangana, India. I am veryenthusiastic in the fields of Data Science and Artificial Intelligence and actively exploring the current trends in these emerging technologies. I've dedicated my time to enhancing my technical skills in these domains and have undertaken significant projects. I always show my dedication to applying AI techniques to real-world challenges. I am enthusiastic about staying at the forefront of these dynamic fields and contributingto the ever-evolving landscape of artificial intelligence.

I'm Palle Venu Madhav, currently pursuing my undergraduate degree in Computer Science and Engineering with a specialization in Data Science at CMR Technical Campus in Hyderabad, Telangana, India. Passionate aboutdata science and its applications, I'm dedicated to honing my skills in this field. With a keen interest in technology and a drive to learn, I'm eager to contribute to innovative solutions in the realm of data science.

I am Mr. B. Ramji, currently serving as an Assistant Professor in the Department of Computer Science and Engineering (CSE) with a specialization in Data Science at CMR Technical Campus. Additionally, I am pursuing my Ph.D. in the field of Computer Science and Engineering at the esteemed National Institute of Technology Warangal. My academic journey includes the successful completion of my Master of Technology (M. Tech) degree in Computer Science and Engineering from JNTU Hyderabad. Prior to that, I earned my Bachelor of Technology (B.Tech)degree in Computer Science and Engineering from the same institution. My research interests encompass a wide spectrum of cutting-edge topics, with a primary focus on Internet of Things (IoT), Bioinformatics, and Deep Learning (DL). These fields offer exciting opportunities to contribute to the advancement of knowledge and technology, and I am dedicated to making meaningful contributions to these areas through my ongoing research and academic endeavours.