How Do Criminals Use Artificial Intelligence and Machine Learning?

Nguyen Huong Quynh Tran
Oulu, Finland
quynh.tran@student.oulu.fi

Federica Cesti
Oulu, Finland
fcesti19@student.oulu.fi

Abuelgasim Elfadul
Oulu, Finland
abuelgasim.elfadulgafar@student.oulu.fi

Tri Hong Nguyen
Oulu, Finland
tri.nguyen@oulu.fi

***Abstract*—Technology is primarily developed to serve human life. However, the dual-use nature of any powerful technology implies the inherent threat of reverse-engineering for malicious uses. Machine Learning (ML) and Artificial Intelligence (AI), as predicted to exceed human power, are the most existing features to enfeeble security systems under cybercrimes and physical attacks. Only criminals mindset can predict malicious intentions. In this report, we review the potential criminal uses of ML and AI, with an insight into data breach and phishing, for the further purpose of proactive prevention measures and security research orientation.**

***Index Terms*—cybersecurity, phishing, credential breach, criminal use.**

I. Introduction

In this digital and Internet era, our lives are not entirely under control. Human faces an ever-high risk of losing data. Traditional cybersecurity methods which we rely on and used to be processed manually such as signature and static rules becoming ineffective due to the deployment of big data, AI and ML[1].

Technology is neutral and it is human who decides the uses either for good or for bad, since scientists know how to create something, they also know how to reverse engineer it. The rapid development of AI and ML raise key questions regarding potential risks related to cybercriminal in particular and security in general.

For the purpose of further efficient control, protection and prevention, in this report we sketch the possible threats from the criminal perspective and how these technologies can play an increasingly role in crime acts in the future. Criminals can exploit AI and ML in several ways. Our study illustrates four most prominent methods: (1) confidential and personal data stealing, (2) impersonation, (3) bypass restrictions for getting access to an account or resources, and (4) automated attacks.

As illustration, we detail the use of AI and ML in data breach and phishing. Despite the large effort and the numerous solutions proposed by the security community, phishing attacks remain today one of the main threats on the Internet[1]. The frequency of stolen data is about 75 records every second which gives over 6 million every day[1]. Criminals use AI and ML to get a pattern for finding information from other tools, allowing criminal to develop methods to act like real. In developing phishing kit or data breach, criminals can deploy AI to automate decision making using highly advance algorithm that collects and interprets data to module behaviors and detect abnormalities [1].

In addition, we describe the frameworks and tools that serve criminals to succeed in phishing; (1) credential leaks which identify usernames and passwords; (2) phishing kit to cheat the users for submitting credential information to fake login pages; (3) Email flagging which is basically a message that tries to trick readers into giving away personal information such as, usernames, passwords, and other information [2].

The remainder of this report is organized as follows. In Section II, we provide an overview of the potential criminal uses of the AI and ML. The state-of-the-art of methodologies and criminal tools available are introduced in Section III. In the next Section IV, we continue to deliver an in-depth explanation of the phishing and credential leak's frameworks. Finally, we discuss some academic and industry implementation on the countermeasures against the future AI and ML criminals in Section V and conclude the report in the last Section VI.

II. How Holy grail can become a scourge

In the second quarter of 2015, the US tax authorities witness one of the hugest data beaches resulting in the illegal access to some 300.000 US taxpayers information. Those personal tax information serves to be filled fraudulent tax returns which values millions of dollars in the subsequent year. Although the attack is halted after the first thousands attempts, the nature of mechanism used provokes a security concern. Taking advantage of new innovations like big data, criminals in recent cyberattacks connect legitimate and stolen data from precedent attacks in new campaigns for the eventual purpose of monetizing[2].

Innovations are expected to open unlimited potential to humankind development, so are AI and ML. However, the dual nature of these innovations also raises concerns on the ever-existed threats. While the popular threats such as phishing, hijack, data breaches are well-documented, how far the potential cyberattacks using AI and ML can reach remains unknown. In this report, we intend to present potential trends and fictitious scenarios that might become a new reality.

Brundage et al. predict that the increase popularity of AI can lead to the cyberattack changes in 3 ways: the widening

[1]https://breachlevelindex.com/

[2]https://www.information-age.com/irs-reveals-how-cyber-criminals-use-big-data-too-tax-breach-200-worse-feared-123460016/

of existing threats, the occurrence of new threats, and the advances in threat characters. [3]

Like other cutting-edge technology innovation, AI and ML have the power to magnify traditional cyber-crimes. The erasing of scalability-efficiency tradeoff equipped with the ease of diffusion derived from the AI's open culture lead to easier, wider, and quicker malicious attacks than ever. It is predicted that in the future even low skill groups and individuals can perform attacks at a higher rate, targeting larger set of victims which are considered currently as unworthy from cost-benefit perspective. Spear phishing is traditionally a time-consuming work with low successful rate. Now criminals can feed big data to the AI system to mimic writing style at a high accuracy and persuasion level, automating the process on massive victims simultaneously. As a result, the world will witness more large scale labor-intensive attacks. [3]

According to Cummings, an AI system is an autonomous one which can "think" and select the best options based on given data inputs, possibly without human intervention. Besides attacking, AI systems are designed to erase trace or hide themselves, making the attributability more difficult [4]. Heinemeyer depicts that future AI-driven opportunistic malware is not only capable to quickly propagate, but also learn context from the target environment and choose attacking techniques accordingly. Autonomous malware does not require C2 channel to communicate, making the threat more secretive and dangerous.[5]

Alternative cyberattacks include the exploitation of nature flaws of the AI system like data poisoning. Criminals can try to feed modified data to confuse machines. They will also likely to automate the discovery of vulnerabilities, which means using past code flaws to speed up the subsequent flaws findings.[3]

New physical threats occur as the smart machines are capable to perform tasks which are currently seen as unrealistic. One of the most dangerous threat derives from the desire to weaponize AI. This threat becomes more critical when the government RD expenditures for AI defense and attack systems are underspent, compared to the technology giants for the commercial versions. While the ban of AI weapons remains a vigorous debate, it is also necessary to address whether those autonomous systems' safety and control are well designed and tested. [4]. Due to the dual use nature of the technology, any civilian automotive can be hijacked or engineered to endangered weapons. In fact, criminals can engineer a commercial drone or cleaning robots to be killer robots or explosives carriers, exploiting the face recognition and navigation system to precisely target victims.

Alternatively, instead in deploying their own machines, criminals can steal control of companies' delivery drones for swarm drone attack or altering the personal smart cars causing crashed. It is also possible that criminals set the digital attack first, using AI to analyze personal interests and behaviors to create personalized clickbait contents such as websites/emails/ads which contain malwares; from there they steal the organizational or personal control of the machines for further physical attacks.[3]

With the development of both digital and physical threats, the consequences are not limited in data breaches like identify information and bank account, but also life risk of any person. It is worth mentioned that under the AI regime, criminal minds are more prompt to malicious actions. While traditional terrorists and psycho individuals might suffer mental trauma, future attackers' ethics are even more blurred as they do not necessarily witness the suffering scenes since AI machines are capable to automate all the dirty jobs [3].

In 2017, Suwajanakorn et al. [6] successfully synthesizing Obama. Using neural network to map the mouth shapes, the authors create a highly realistic fake videos based on the original ones. This advancement makes fake news and impersonation as apparent new digital threats in the near future. Using the similar speech and image synthesis system, criminals can create fake videos and chatbots to deceive individual for traditional purposes as stealing personal information, or worse, to manipulate public opinions. Combined with the viral power of social networks, the fake information can easily cause social distress and distrust, in other word, another type of crisis.

Brundage et al. [3] lists this threat under the political attacks which implies an attack on politicians in presidential election via fake news and impersonation. Besides, this threat can be understood as a potential social and political manipulation of the government on the public. Precisely, the government deploys the AI and ML to create a mass surveillance infrastructure, secretively analyze social behaviors through mass data analysis, from there tightening censorship, creating propaganda to suppress debate or distract public opinions away from important content. As a consequence, citizens might not only be subject to privacy invasion but also social manipulation.

One of the first similar system is currently executed by China government, namely the Social Credit System. The Chinese government uses millions of cameras to track its citizens anywhere. Citizens' social data are combined together to create the fullest possible pictures of social and political behavior. Despite of the government advocates on the society management, this system is strongly criticized for potential Orwellian dystopia as it seizes citizen freedoms under strict surveillance. [7]

III. Survey of Methodologies

There are many tools and ways in which AI and ML may contribute in malicious attempts. In this section we analyze tools that involve AI and ML which can help cybercriminals.

A. Information gathering

The aim of this attack is to steal confidential and personal data from users and companies.

Hackers take advantage of Machine Learning's classifying algorithms to drive phishing attacks, targeted at specific individuals. After having collected data about user's preferences and typical behaviours, criminals send malware only to users who would click on the malicious link. These kinds of attempts

are called spear phishing, which trick users by using messages specifically selected for each individual in order to gather and share personal data or install malware. Moreover, Machine Learning models can easily and fast create large numbers of human-written messages[3] and help in detecting social media accounts by applying image recognition tools. For example, the software Social Mapper is a tool used to search for an individual in different social media[4].
While network scanners and sniffers are able to analyse traditional networks, AI is used to gather information about networks based on Software-Defined Networking (SDN). The Know Your Enemy attack aims to collect the configuration of a SDN network, like security tools or networking parameters. Information gathering can be automated, for example a tool like DirBuster, which scans directories and files, can become highly powerful if genetic algorithms, LSTMs or GANs are applied to produce more likely names[4].
An alternative automated tool which enhances phishing campaigns is SNAP_R. In practical, criminals deploy scam-like email sproofings carrying suspicious links to websites which forge legitimate ones. Victims is tricked to fill their private information in those fake websites. Besides suspicious websites, phishing emails can be used to distribute malware and spyware[5].

B. Impersonation

Once a cybercriminal gathers personal information about an individual, an impersonation attempt could easily be the following move. Social media phishing gives the opportunity to monitor and learn users' behaviour and use it against them: for example, Markov models can generate tweets based on user's old posts. With AI, fake voices and fake videos can be created. Special software like Google's WaveNet can make a bot that speaks exactly like a human, by applying generative adversarial networks (GANs). Tools like DeepFake can generate a video with a celebrity or a politician face making invented speeches, a software from Nvidia can design synthetic celebrity images[4].

C. Bypassing restrictions

Another common usage of AI in cybercrime regards bypassing restrictions, in order to get access to resources or accounts. Support Vector Machines can solve CAPTCHA images with accuracy of 82%, which may increase to 92% when deep learning is integrated[4].
CAPTCHA is used as a common tactic to stop bots from entering a website: the idea is to test the user to make him/her prove that he/she is a human being. Those tests usually involve text guessing form images or the "select all pictures containing a bus". Accuracy increases every year: scientists are worried that new techniques will be able to solve 100% CAPTCHA tests[4]. Markov models were the first method used to generate password guesses, before deep learning came. Neural networks and LSTMs generate text based on the trained texts and could be used to generate passwords: if the training set consists of the most used passwords, the model will be able to generate a lot of words which are comparable with the ones in the training set. In particular, attackers rely on PassGAN, which uses GANs[4]. They are Generative Adversary Networks, special neural networks consisting of two deep networks, one against the other: the generator creates new data instances and the discriminator evaluates their authenticity, returning probabilities that represent the validity of the instances[6].

D. Automated attacks

An additional application of Machine Learning involves real attacks, like malwares. A common technique is Fuzzing, which is a vulnerability discovery method: it implies putting random inputs in the application in order to crash it. A model can be trained to predict more relevant crashes: in this way the process will be faster and more efficient. Reinforcement learning is used to create malwares with Machine Learning. DeepLocker is one representative malware of this kind that has the self-hiding ability equipped with the postponed execution until a target is detected by its AI's face or voice recognition, or geography detection. Therefore, the new generation malware which behaves alike irregular guerrilla can become a greater source of danger than original ones[7].
The rise of AI is changing the way DDoS are used. The Distributed Denial-Of-Service (DDoS) attack aims to make an online application unavailable by overwhelming it with requests from multiple sources. It targets any kind of company, like banks, news websites and social media platforms. [8]Before AI, the limitation of DDoS was the need of human intervention, because attackers needed servers to distribute commands in order to steal data, spread malware, or change vulnerability types (bugs, weak passwords, virus, missing authorization, etc.). Nowadays, AI-driven DDoS no longer require human intervention as they are fully automated and can change their behavior by monitoring the respond from the defense side[9].
With the help of Machine Learning, cybercriminals might be able to automate the generation and the spread of false content like fake news, articles or social-media posts. For example, these techniques could be used during the elections to put a politician in a bad light. Fake news had already caused a lot of problems in these last few years and its automation will make its detection even harder[10].
Lastly, another way in which AI is becoming extremely efficient is its implementation in unmanned aerial vehicles (UAVs), the so-called drones, in commercial and military

[3]https://medium.com/latinxinai/machine-learning-security-still-all-hype-or-a-real-concern-e7e9c8e78a2e

[4]https://towardsdatascience.com/machine-learning-for-cybercriminals-a46798a8c268

[5]https://www.barracuda.com/glossary/phishing-campaign

[6]https://skymind.ai/wiki/generative-adversarial-network-gan

[7]https://www.dailydot.com/debug/ai-malware/

[8]https://www.digitalattackmap.com/understanding-ddos/

[9]https://www.acunetix.com/blog/articles/artificial-intelligence-ddos-attacks-part-2/

[10]https://www.technologyreview.com/s/612960/an-ai-tool-auto-generates-fake-news-bogus-tweets-and-plenty-of-gibberish/

contexts. This topic has been highly discussed, mainly for ethical reasons: is it correct that robots, which could fly, walk and swim, incorporate AI that would make them able of executing military mission on their own, especially with the possibility to put human lives are at risks? [4]
What these weapons can achieve is still partially unknown and the risk is that their power may fall into the wrong hands. Drone swarm technology may change forever the idea of conflicts. In fact, the capabilities of swarms of drones are endless and may become be really dangerous: they can search for adversary submarines in the oceans, identify and eliminate hostile missiles and other air defenses, spread chemical weapons, use biological, radiological, and nuclear detectors, apply facial recognition tools. Whereas on one hand drones offer defenses against a range of attacks, on the other hand they are extremely dangerous for civilians[11].

IV. Credential Leaks and Phishing Kits Framework

In this section, we would like to go more details about frameworks which can be built in order to serve criminal activities. In more detail, a credential leak and phishing kit frameworks are described in this section, respectively.

A. Credential Leaks

A strategy to identify usernames and passwords through data breaches is proposed by Thomas [2] in order to prove that credential leaks on private markets can be found on the Internet such as paste sites, public forums and also Google history. To demonstrate this idea, [2] collected data from private markets and public forms including pate sites. After that, a proposal framework as fig 1 is mentioned with three main tasks: crawling credential leak, classification and hash version.

1) Crawling Credential Leaks: In term of crawler, [2] collected data from 5 public blackhat forums and 115 paste sites. However, it should be noticed that there are several subsets of paste sites and public forums which can be candidate documents for later analyses. Therefore, to reduce the number of useless data from these sites, a filter is utilized to remove pastes which have the number of email addresses being lower than 100 email addresses ([2] kept the forum threads due to the less amount of data from forum threads). Additionally, Thomas et al. [2] collected Google's recent history as candidate documents which contain at least 10 over 1000 common passwords or some fixed-length strings as hash values going after email addresses (the common passwords are from previous credential leaks). However, this set of data cannot capture credential leaks which are protected. As a consequence, [2] captured 31446 candidate documents from public places and 258 huge credential leaks with 1.79 billion non-unique usernames and passwords from 11 private forums (from June 2016 to May 2017).

[11]https://mwi.usma.edu/era-drone-swarm-coming-need-ready/

2) Understanding and Analysis on Data: First of all, the set of data is separated as columns of data based on a "delimiter detection". After that, [2] recognized that confirmed credential leaks are produced based on highly structured format s such as key-value pairs, SQL or JSON blobs. Therefore, a set of multiple parsers is used to detect records containing at least two columns. Meanwhile, the remaining data is analyzed more details as an email column and a password column. To recognize the email column, Thomas et al. utilized a "regular expression" and other information (IP address, browser-agent and so on). Nevertheless, to obtain the password column, it becomes like a challenge due to the flexible characters and lengths. [2] separated two types of password column which can be happened (hashed and plain-text password). If the passwords are hashed, they will have the same length of character (e.g 32 in term of MD5). In the plain-text password, a binary classification to categorize "yes or no" password columns in candidate documents are utilized with the training set from plain-text credential leaks (private forums). Firstly, the training set is transformed into binary vectors of n-grams. In this case, there is a question that is about the size of n-grams which should be used. To answer this question, a grid search with 10-fold cross-validation on data from private forums (training data) is considered on the length of n-grams from 1 to 10 and binary vectors containing the top 1000 to 100000 most common n-grams in password value and non-password value classes. As a result, the classifier approach is built based on n-grams with a range between 2 and 5 and the number of binary vectors being 10000 of each n-gram in each class. For testing, [2] used 230 labeled candidates (157 stolen credentials and 73 not) to obtain 93.9% accuracy. Then, this model is applied to candidate documents from public sites in order to drop failed documents. As a consequence, 3527 candidate documents were classified to achieve 123055697 emails and passwords from MySpace, Badoo, Adobe, LinkedIn and so on. Last but not least, from plain-text passwords and current dictionaries, a set of 3416701663 keywords is utilized to inverse MD5 and SHA-1 passwords. From the report [2], 35.8% of hashed passwords were inverted. The result is quite low due to extra-secure approaches such as salted passwords.

B. Phishing Kits

A framework for phishing attacks is described in this subsection. The idea of this framework introduced by Kurt Thomas et al. [2] consists of four main components which are crawling phishing kits, template extraction, rule generation and email flagging. [2] used the phishing framework to understand the life cycle of phishing kits and their potential victims.

1) Data of Phishing Kits: At the beginning of the framework, a description of phishing kits which are collected from March 2016 to March 2017 by [2] through an undisclosed source is mentioned. The set of phishing kits is included 10037 phishing kits (PHP and HTML source code) which are utilized to collect 3779664 usernames and passwords from victims with the time-stamp.

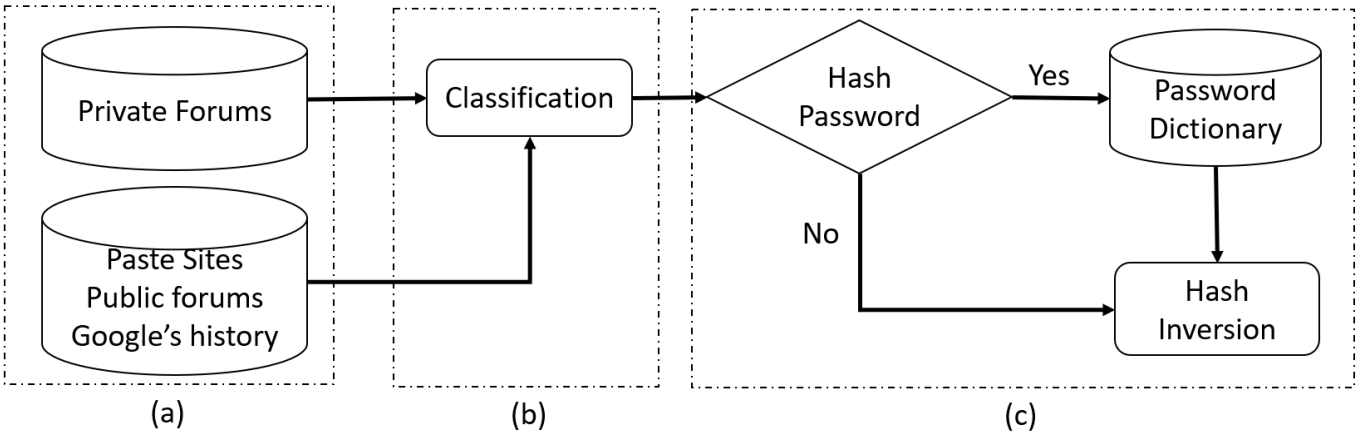

Fig. 1. Framework for identifying credential leaks [2]: (a) Crawling data, (b) Parsing and analyzing data, (c) Hash inversion.

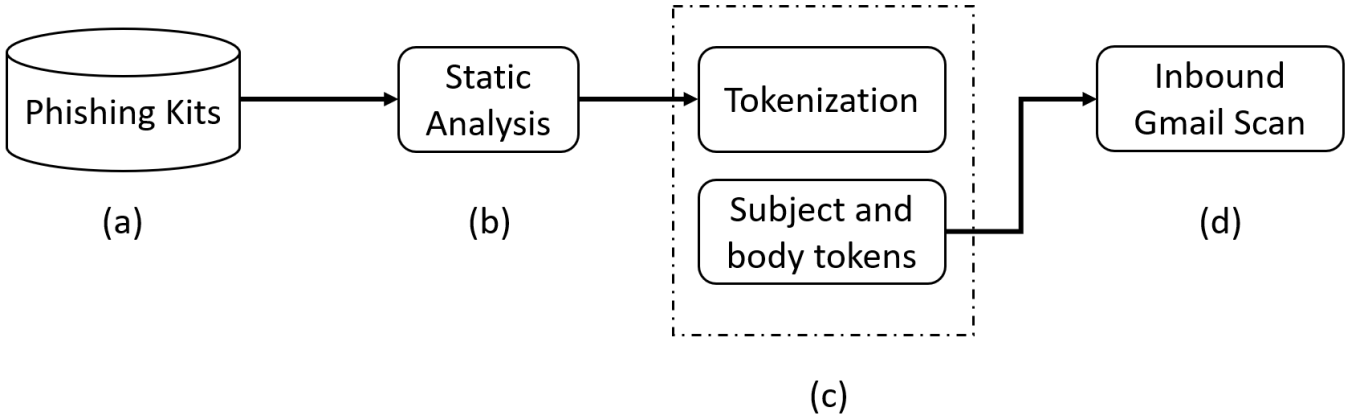

Fig. 2. Framework for phishing kits [2]: (a) Crawling data, (b) Template extraction, (c) Rule Generation, (d) Email flagging.

2) Template Extraction: After utilizing a static analysis, Thomas et al. [2] showed that the collected phishing kits use mail() command in PHP (as Listing 1) for notification of stolen credentials to victims for asking their sensitive information. As Listing 1, a phishing kit asks the victim (a hardcoded email) about their information including username, password, IP address, country and browser agent. In particular, with the use of static analysis, [2] mentioned that they can automatically obtain values from each variable after calling mail() command. Finally, this part of the framework returns a template which consists of the target email as exfiltration point, subject and body of the message. [2] used a set of 7780 unique exfiltration points or targets' email from different providers (Gmail 72.3%, Yahoo 6.8%, Yandex 5.1%, Hotmail 4.2%, Outlook 2.2%).

Listing 1. Phishing kit collects data (username password browser user-agent ip address) from victim [2]

```
$sub = 'Result from Gmail'
$msg .= '-- Gmail info --'
$msg .= 'Username: '.$gmail_user.'\n'
$msg .= 'Password: '.$gmail_pass.'\n'
$msg .= '-- Victim info --'
$msg .= 'Client IP: '.$ip_address.'\n'
$msg .= 'Browser: '.$browser_agent.'\n'
$msg .= 'Country: '.$country.'\n'
mail('victim@gmail.com', $sub, $msg)
```

3) Rule Generation: This part of the framework is about the tokenization of message templates to form a set of rules which combine the subject and body of inbound emails containing stolen credentials with a maximum expected length of the message. [2] mentioned that they preferred rules for identifying the exfiltration points due to the ability of reusable rules. As an example from Listing 1, a message matching the template if its subject is "Result from Gmail" and the body has "Username, Victim info" and so on. Finally, [2] extracted a set 7325 rules extracted through 10037 phishing kits.

C. Email Flagging

[2] configured Gmail's anti-abuse detection for testing on the number of exfiltration points which receives stolen credentials, the number of messages each account can obtain and the number of messages per each kit template. However, [2] set a restriction, which the exfiltration points are required to receive

at least 20 stolen credentials. As a result, there were 12449036 messages flagged by [2] and 19311 exfiltration points.

V. Discussion

The fictitious uses of AI and ML described in previous parts concentrate on dark sides of technology. For the technology advances are irresistible, staying proactive to foreseeable scenarios and interventions rather than reactive upon occurred damages is a common practice among academic and industry experts. In this section, we discuss the academic and industry advances in preventing the potential risks as well as a short guideline for individual users.

To prevent AI and ML cybercrimes, academic community intensively centers on the data poisoning vulnerability, forming a research field named Adversarial Machine Learning which mainly aims to corrupt classification models. For instance, [8] found that Multi-task Relationship Learning model is highly sensitive to input noises; from there they suggested to delve more into the strong correlation patterns of attacked tasks in future defense study.

Prior model protection methods include building resilient algorithms, deploying multiple classifiers hence corrupted points are treated as outliers, or hiding classifier model information. Taking a direct approach from the fact that corrupted points have identical features to genuine ones except for the flipped labels, [9] proposed Curie which serves as a filter for malicious input before retraining stage of Support Vector Machine (SVM) classifier. This method is efficient yet simple enough to incorporate in any existing ML systems. [9]

An additional non-technical solution for data poisoning includes responsible disclosure of vulnerabilities and models to accelerate the patches, such as the CleverHans library. Such attempts can discourage attackers as they cannot deploy new vulnerabilities on a massive scale of victims. [3]

In contrast to the proliferation of new vulnerability discoveries, a few studies focus on the defense methods' robustness [3]. [10] provided a theoretical upper bound determining the worst loss of defenders, which is efficiently scalable though applicable only in a design stage, not the deployment stage. Still, this field remains a point of concern of academic experts hence there poses a need for orientation and resource allocation.

Technology companies have more generous approaches to find system vulnerabilities. They can rely on outside stakeholders such as hackers by paying bounties for successful hacks and suppliers by buying commercial products/services that test source code flaws. From the inside, attack simulation via Red Teaming strategy is commonly used. This strategy, which comprises an attacking Red Team and a reacting Blue Team, is considered as efficient especially for AI cyberattacks and adversarial machine learning. Additionally, an automated vulnerability finding system using Fuzzing can be used internally. [3]

To prevent primitive cybercrimes like spear phishing and malware, Google uses a simple spam filter. This solution is highly efficient when equipped with big data of users. Other security protection measures involve the multi-factor authentication, frequent log and IP tracking, mail alerts of suspicious activities, and blocking suspicious IP addresses and accounts in severe malicious cases. Data encryption is another popular solution to ensure data transmissions [3]. Such security solutions are long put in practice by technology giants like Facebook. McAfee even examines behavioral analytics to distinguish legitimate logins for business and personal purposes from malicious ones. [11]

On physical threat prevention, the legal enforcement on the whole supply chain, from the hardware manufacturers to end users and payload, plays a vital role. Precise control measures for drones include the sale restriction of potentially destructive machines, registration requirements, and no-fly zone at sensitive areas such as military fields and airports [3]. Technical monitoring can be obtained nationwide using radar scan, ML-driven image recognition software [12], or surveillance cameras to spot suspicious individuals and flying objects [13].

Addressing misinformation threats, academic researchers are capable to detect fake news from authentic ones. [14] used the linear SVM classifier to distinct fake news based on their linguistic properties. [15] conducted a comprehensive study on fake news features, content and social context models, and performance evaluation ratios. [16] aimed at distinct trustworthy social media posts from untrustworthy ones by an ML-based Reputation Scoring system.

On detecting modified images, [17] developed PhotoProof – an image authenticator which leaves computational cryptographic marks on edited image, therefore smoothing the legitimate image verification and amendment. [18] studies live broadcasting videos' authenticity based on the consistency of video motion and camera movements; from there proving a mean to verify whether the video is broadcasted live or synthesized offline. However, synthesis multimedia generation seems to advance prevention measures [3] opening various prospects for defense studies.

In practicality, [3] insisted on the importance of social media platforms in terms of content governance against automated misinformation and public opinion manipulation. These platforms possess the administration power over the information accessibility and access rate, with exceptional large sets of users which strongly benefit AI algorithms. In fact, Facebook's AI algorithms can discard 99 percent of ISIS and Al-Qaeda terrorism contents[12], and prevent clickbait headlines by a classification system targeting fake news' common features including information withholding and exaggeration[13].

A final remark is that end users easily fall to simple tricks despite the available knowledge and security settings [3]. Taking psychology into consideration, they are more prompt to impulse reactions than rational consideration, given the fact that malicious links always seem very attractive to victims'

[12]https://uk.reuters.com/article/us-facebook-counterterrorism/facebook-reports-progress-in-removing-extremist-content-idUKKBN1DT003

[13]https://newsroom.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed/

interests. Notwithstanding the protection of service providers, users also bear the responsibility of their private data. Hence it is advisable to change the password every six months or one year, avoid easy-guessing passwords, diversify passwords for different accounts, follow companies' settings such as the multi-factor authentication, self-educate and always stay alert online.

To sum up, both industry and academic intensely focus on the defenses against the potential AI and ML crimes, however, higher attention and resources should be allocated to discover more confident defense methods and robustness studies. It is equally important for end-users to protect themselves in addition to the available configuration of the service providers.

VI. Conclusion

In conclusion, the AI and ML enhance modern life in various ways, but they also can magnify conventional security threats as well as create new crimes. Future life is predicted to witness more personalized cybercrimes which can be done automatically. There will also be new threats such as impersonation, robot killers, and drone swarm attacks.

It is possible to turn the Holy Grail's technology into a scourge without proactive preventions. With the power exceeding human capabilities, AI and ML are capable to create damages at a scale that human can never reach [3]. Our study examines some most prominent threats of AI and ML wrong uses. Further researches can delve more into potential alternative threats such as fraudulent advertisement and simulating fake traffic, as well as the efficient control and stringent prevention measures.

Acknowledgement

We would like to express our sincere gratitude to D.Sc.(Tech) Ekaterina Gilman, M.Sc.(Math) Lauri Lovén, and Hassan Mehmood through lectures and exercises in the course "Big Data Processing and Applications" at the University of Oulu.

References

[1] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[2] K. Thomas, F. Li, A. Zand, J. Barrett, J. Ranieri, L. Invernizzi, Y. Markov, O. Comanescu, V. Eranti, A. Moscicki, *et al.*, "Data breaches, phishing, or malware?: Understanding the risks of stolen credentials," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1421–1434, ACM, 2017.

[3] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[4] M. Cummings, *Artificial intelligence and the future of warfare*. Chatham House for the Royal Institute of International Affairs, 2017.

[5] H. M., "The next paradigm shift: Ai-driven cyber attacks," *DarkTrace Research White Paper*, 2018. [Online; accessed 19-April-2019].

[6] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.

[7] R. Creemers, "China's social credit system: An evolving practice of control," 2018.

[8] M. Zhao, B. An, Y. Yu, S. Liu, and S. J. Pan, "Data poisoning attacks on multi-task relationship learning," in *AAAI*, 2018.

[9] R. Laishram and V. V. Phoha, "Curie: A method for protecting svm classifier from poisoning attack," *arXiv preprint arXiv:1606.01584*, 2016.

[10] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in neural information processing systems*, pp. 3517–3529, 2017.

[11] L. McAfee, "Mcafee labs 2018 threats predictions," *Mission College Boulevard, Santa Clara, CA*, 2017.

[12] C. Aker and S. Kalkan, "Using deep networks for drone detection," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2017.

[13] P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, *et al.*, "Artificial intelligence and life in 2030. one hundred year study on artificial intelligence: Report of the 2015-2016 study panel," *Stanford University, Stanford, CA, http://ai100. stanford. edu/2016-report. Accessed: September*, vol. 6, p. 2016, 2016.

[14] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *arXiv preprint arXiv:1708.07104*, 2017.

[15] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[16] N. Khurana, S. Mittal, and A. Joshi, "Preventing poisoning attacks on ai based threat intelligence systems," *arXiv preprint arXiv:1807.07418*, 2018.

[17] A. Naveh and E. Tromer, "Photoproof: Cryptographic image authentication for any set of permissible transformations," in *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 255–271, IEEE, 2016.

[18] M. Rahman, M. Azimpourkivi, U. Topkara, and B. Carbunar, "Video liveness for citizen journalism: attacks and defenses," *IEEE Transactions on Mobile Computing*, vol. 16, no. 11, pp. 3250–3263, 2017.