IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 28, 2020, accepted October 3, 2020, date of publication October 7, 2020, date of current version October 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029280

Artificial Intelligence Security Threat, Crime, and Forensics: Taxonomy and Open Issues

DOOWON JEONG
College of Police and Criminal Justice, Dongguk University, Seoul 04620, South Korea
e-mail: doowon@dgu.ac.kr

ABSTRACT Advances in Artificial Intelligence (AI) have influenced almost every field including computer science, robotics, social engineering, psychology, criminology and so on. Although AI has solved various challenges, potential security threats of AI algorithms and training data have been stressed by AI researchers. As AI system inherits security threats of traditional computer system, the concern about novel cyberattack enhanced by AI is also growing. In addition, AI is deeply connected to physical space (e.g. autonomous vehicle, intelligent virtual assistant), so AI-related crime can harm people physically, beyond the cyberspace. In this context, we represent a literature review of security threats and AI-related crime. Based on the literature review, this article defines the term *AI crime* and classifies *AI crime* into 2 categories: *AI as tool crime* and *AI as target crime*, inspired by a taxonomy of cybercrime: *Computer as tool crime* and *Computer as tool crime*. Through the proposed taxonomy, foreseeable AI crimes are systematically studied and related forensic techniques are also addressed. We also analyze the characteristics of the AI crimes and present challenges that are difficult to be solved with the traditional forensic techniques. Finally, open issues are presented, with emphasis on the need to establish novel strategies for AI forensics.

INDEX TERMS Artificial intelligence, AI crime, AI forensics, security threats, malicious AI.

I. INTRODUCTION

Artificial Intelligence (AI) has become essential to almost all areas including computer science, security engineering, criminology, psychology, and robotics. Especially, Deep Learning [1], inspired by the structure and function of the brain, has been the major breakthrough in the AI field [2] and it has activated the AI study in various fields. Research on deep learning has been studied to process a huge amount of data (e.g. pictures, medical information, social media, crime information, etc.) to perform medical image analysis, speech recognition, and natural language understanding [3]–[7].

Although the fast development of AI has brought the benefits of innovation, it also has carried the significant risks [8]. This unprecedented growth reminds AI stakeholders of the early days of Information & Communication Technology (ICT). When ICT evolved at high speeds in the past, unexpected problems occurred (e.g. terrorism, security threat, cybercrime, privacy infringement, etc.) and it incurred the considerable social costs. Similarly, there are growing concerns about various problems that the AI can cause [9]. As Brundage *et al.* [10] stressed the importance of the changing threat environment, research on preventing and mitigating the dark side of AI also should be discussed and expanded seriously.

In this context, we explore AI security threats, foreseeable crimes, digital forensics for AI. A literature search for the subject covered various books, journals, and conference proceedings. Due to the heterogeneous nature of AI and digital forensics, we reviewed not only forensic researches but also researches of computer science, related law, criminology, etc. We used Google Scholar, IEEE Xplore, and ACM Digital Library to search for related paper using the keywords: 'AI', 'security threat', 'AI crime', 'forensic framework', etc. We also looked for studies that the retrieved works cited and were cited by the retrieved works. Among the papers, we tried to review articles published after 2015. However, we did not rule out papers published prior to 2015, as there would be works highly relevant to this survey. We note that this paper regards AI as a set of algorithms including training and inference processes to mimic human intelligence.

By reviewing previous studies, we define *AI crime*, and then propose a taxonomy for new types of crime: the *AI as tool crime* and *AI as target crime*, inspired by an existing

The associate editor coordinating the review of this manuscript and approving it for publication was Charith Abhayaratne.

taxonomy used in cybercrime: *computer as tool crime* and *computer as target crime* [11], [12]. The *AI as tool crime* is defined as the expansion of existing crimes, including traditional crime and cybercrime (e.g. advanced phishing, automated hacking, manipulation, fraud, etc.). The *AI as target crime* is a new area of potential criminal activity against AI system; adversarial attack [13] is a typical example. Based on proposed taxonomy of AI crime, this article discusses how to investigate the crime; we name this process *AI forensics*. Note, in the digital forensic field, a word in front of the 'forensic' implies the target to be analyzed (e.g. smartphone forensics, cloud forensics, memory forensics, IoT forensics). Therefore, AI forensics indicates not investigation using AI but investigating AI. We also perform a comparative analysis between traditional digital forensics and AI forensics.

In exploring the facets of AI crime, we make the following contributions:

- We discuss foreseeable AI-related crimes systematically and practically.
- We propose a taxonomy of AI crime based on comprehensive literature review.
- We introduce challenges that digital forensics can encounter when investigating AI crime with experiments.
- We highlight open issues in the field of AI forensics and propose corresponding suggestions. To the best of our knowledge, this paper is the first systematic study about AI forensics.

The rest of this paper is organized as follows. Section II introduces related work and background of AI security threat, AI crime, and digital forensics. In Section III, we define *AI as tool crime* and then describe foreseeable AI-related crimes. Section IV explains *AI as target crime*, which attacks training system and inference system. We discuss AI forensics aiming to investigate the AI crimes in Section V. Section VI highlights open issues of AI forensics by comparing the traditional forensics. Concluding remarks are drawn in Section VII.

II. RELATED WORK AND BACKGROUND

As already mentioned in the introduction, AI has been studied in various academics. This section describes researches about AI security threat and AI-related crime from various perspectives. In addition, we also explore cybercrime defined by cybersecurity and digital forensics community.

A. AI SECURITY THREATS AND CRIME

The term 'AI crime' was firstly provided by humanities field [9] as the term 'crime' is involved with law and ethics. Although the term AI crime has not been covered in computer science area, several studies have stressed security threats and malicious uses of AI that can cause various crimes.

A prime study of the malicious use of AI is about adopting online personas, called socialbot, that behaves like human [14]. Though the initial objective of socialbot was to advocate awareness and cooperation among people [15], it has often been used maliciously such as phishing, fraud, and political infiltration of a campaign on online social networks [16]. Seymour and Tully [17] presented that machine learning can be weaponized for social engineering; by using AI, mass-produced messages with phishing links could be posted on Twitter without any interruption. Because the malicious socialbot is based on a specific user's past behaviours and public profiles, detection of the socialbot has become the challenge of computer security [18], [19]. From the social science perspective, the technique may influence or inflame public opinion when malicious socialbots are designed to perform a political attack [20], [21].

Some researchers gave warning that hackers have already started to weaponize AI, in order to advance their cracking skills and develop new types of cyber attack [22]. The AI is utilized to sharpen techniques to commit traditional cybercrimes such as financial fraud, cyberterrorism, cyberextortion, etc. For example, when hackers try to voice phishing, the hackers can deceive victims by using the realistically imitated voices of the victims' family or friends [23].

Whereas the above studies focused on the problems that the specific techniques could cause, Brundage *et al.* [10] presented a comprehensive insight into the malicious use of AI. They addressed three changes in the landscape of threats: expansion of existing threats, the emergence of new threats, and change to the typical character of threats. By the scalable use of the AI system, the cost of tasks that require human labor may be lowered. Perpetrators then are able to attack more targets with the cost reduction techniques (e.g. mass spear phishing); this is the expansion of existing threats. The new threats also may be emerged to complete tasks that are infeasible for people (e.g. imitating individuals' voices, controlling multiple drones) [24]. When the highly effective attacks by AI become more common, the typical character of threats will be altered. Brundage *et al.* also classified security domains into digital security, physical security, and political security. The digital security domain includes cyberattacks that exploit vulnerabilities of human or AI systems. The physical security domain covers physical attacks such as causing autonomous vehicles to crash and controlling thousands of drones. The political security domain includes novel threats in profiling, repression, and targeted disinformation campaigns.

King *et al.* [9] provided a different view about AI security threats, by using the term 'AI crime'. They approached the problem from a broader perspective. In the article, AI crime is categorized based on criminal behavior: commerce, financial markets and insolvency (e.g. market manipulation, price fixing, collusion), harmful or dangerous drugs (e.g. trafficking, selling, buying, possessing banned drugs), offences against the person (e.g. harassment, torture), sexual offences (e.g. sexual assault, promotion of sexual offence), theft and fraud, and forgery and personation (e.g. spear phishing, credit card fraud). They insisted that the categorized crimes contain one or more threats. When classifying AI security threats, they focused on human's nature: emergence, liability, monitoring, and psychology. For example, the psychology threat means that AI can affect a human's mental state to the extent of facilitating or causing crime. This approach is quite different from that of the computer science field; this variety of perspectives is due to the inherently interdisciplinarity of AI.

Some studies focused on privacy issues of AI arising from processing of personal data. Li and Zhang [25] presented that AI applications in healthcare, finance, and education may occur privacy problems. As the number and quality of training data greatly affect the performance of AI, developers wish to collect as much data as possible. Li *et al.* insisted that the collection of comprehensive data has inherent privacy threats. Mitrou [26] approached the privacy problem with General Data Protection Regulation (GDPR). The author stressed that GDPR can be applicable to AI when AI handles personal data, though GDPR does not specifically address AI.

The previous studies give three implications to stakeholders in AI field. First, due to the dual-use nature of AI, researchers and engineers should perceive that AI technique may be used to commit criminal offences, even though the technique is designed for legitimate use. Since AI is a double-edged sword, stakeholders in AI field need strict professional ethics. Second, totally different types of security threats, that have not been considered so far, will emerge. As AI can complete tasks that have been regarded as impossible to be processed by people or traditional programs, the threats will be outside the primary scope of known threats; AI researchers should collaborate closely with professionals in diverse fields to prevent the security threats of AI and respond to AI crime. Finally, AI security area should learn from trials and errors of cybersecurity area. As described in previous studies, the foreseeable AI crimes are very closely involved in cybercrime. The cybercrime stemmed from the dual-use nature of ICT; the current situation of AI security resembles the initial phase of cybersecurity.

B. CYBERCRIME

Cybercrime is regarded as the dark side of cyber space [12]. It is categorized into two types, *computer as target crime* and *computer as tool crime* [11], [27]. As information has been digitized and connected by network, new types of crimes, such as cyberterrorism, cyberextortion, cyberwarfare, etc., have emerged; these crimes are called *computer as target crime*. The objective of *computer as target crime* is disrupting or destroying computer systems. Therefore, when perpetrators commit the *computer as target crime*, they use tools or techniques developed to intrude computer systems (e.g. viruses, worms, Trojan horses, and spyware). Meanwhile, all data in our daily life have been digitized from private area to business. This change makes offline crimes such as fraud, threats, child abuse, stalking, etc. enter the online environment; it is called *computer as tool crime* [12]. Cybercrime is intimately related to cybersecurity because most attack techniques in cybercrime are based on exploiting vulnerabilities of potential target [28], [29].

The taxonomy of cybercrime have helped to develop strategies against the crime in practice. When forensic investigators examine *computer as tool crime*, they focus on proving the perpetrator's past behavior to determine whether illegal behavior occurred or not. In *computer as tool crime*, the criminals generally use known tools and manipulate familiar infrastructures such as the mobile message, website, social media, etc. On the other hand, when investigators examine *computer as target crime*, they should focus on malicious programs, called malware. To quickly respond the crime and ascertain the extent of the damage, they must find the malware and then perform reverse-engineering to understand the purpose of the malware and identify the source of the attack [30]–[32].

1) DIGITAL FORENSICS

Digital forensics is defined as "the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence" [33]. In digital forensics area, many principles and guidelines have been suggested because each country and organization has various laws and policies. Nevertheless, they share an underlying foundation that forensic process is considered to be forensically sound only when it meets five principles: *Meaning*, *Errors*, *Transparency and trustworthiness*, *Reproducibility*, and *Experience* [34]–[39].

- *Meaning*: The original meaning of evidence should be unchanged; when change is inevitable, there should be minimal change.
- *Error*: Any unavoidable error in the forensic process should be documented.
- *Transparency and trustworthiness*: The reliability and accuracy of the forensic process should be tested and verified.
- *Reproducibility*: The result of the forensic process should show a consistent level of quality, no matter how many times it is repeated under the same conditions.
- *Experience*: The investigators should have sufficient experience or knowledge.

If the forensic process does not follow any of the five principles, the evidence would be hard to be accepted in court. Thus, investigators should collect and analyze the evidence while adhering to the principles.

In addition, forensic researchers proposed a proactive process that is used to manage incidents before they can occur [40]; the process is called Digital Forensic Readiness (DFR). DFR aims to collect digital evidence quickly and accurately while minimizing the cost of conducting forensic investigation during incident response [41]. In particular, DFR has been used to mitigate business risks of losing information assets due to security incident. As the incidents stem from vulnerability of information system, DFR also plays a role in preventing or detecting cybercrimes.

Similar to other fields, digital forensic researchers have also studied application of AI to investigation. Karbab and Debbabi [42] used natural language process and supervised machine learning to detect malware. They achieved over 94% f1-score in several datasets. Fidalgo *et al.* [43] also applied AI to digital forensics, to classify suspicious content posted on the Dark Web. By developing the monitoring system based on AI, it made the investigation efficient. In addition, several researchers have studied on forensic investigation methods using AI ([39], [44], [45]), but study on AI as the subject of forensic investigation has not been published yet.

III. AI AS TOOL CRIME

This section describes foreseeable *AI as tool crime* considering the dual-use nature of AI. Because AI system is also developed on digital infrastructure, the risk of cybercrime, including *computer as tool and target crime*, is embedded within AI security threats. In addition, AI can be used for physical crime by controlling autonomous devices like smart car, drone, Internet of Things (IoT) device, etc.

In this section, we explore how AI can be used to sharpen cyberattacks. Then, we focus on physical crime, regarded as a novel attack.

A. ENHANCED CYBERCRIME

As described in Section II, there are two types of cybercrime: *computer as tool and target crime*. They are traditional crimes in cyberspace, but the crimes are still serious threats. By using AI techniques, perpetrators can commit novel cybercrime that was considered an infeasible attack. This subsection discusses how AI techniques can be used for cybercrime.

1) COMPUTER AS TOOL CRIME

Previous researches notified that AI can be used for phishing and its effectiveness are already proved [9], [10]. One of the common phishing methods is scam email using profiling. The profiling using AI has been actively studied in the business field; targeted advertising is a typical example. However, the technique used in the targeted advertising, which is based on the customers' previous buying history or interest, may be instrumental for the attacker. The previous researches named the AI programme as a chatbot. Kietzmann *et al.* [46] and Paschen *et al.* [47] predicted that AI will enhance strategies to scam customers by using the malicious chatbot.

The chatbot is able to communicate with the customers without a break; it can collect mass data related to the customers' behavior and profile. The chatbot has been already developed and used in academia and industry. In the early days, the chatbot was mainly text-oriented [48]. However, the chatbot has been developed to verbally converse with people, as Natural Language Processing (NLP) technology has advanced [49], [50].

Whereas some studies suggested using the ability of AI's speech conversation for the common good, such as social therapy [51], education [52], medical diagnosis [53], and health [54], there are also concerns that AI-supported voice would raise theft and fraud. As the voice is one of the biometrics which is the irreplaceable measure in security mechanism, it can be a great weapon for attackers (e.g. voice phishing) [23].

Fake news is another example of the advanced crime. Although fake news has a long history in social engineering [55], it has begun to get noticed recently with the advent of social network services (e.g. Twitter, Facebook, YouTube) [56]. In particular, the fake news has a huge effect on political issues such as policy decision, propaganda, and election [57]. With the deepfake technique, the fake news gets more powerful. Citron and Chesney [58] presented that fake video mimicking prominent politicians can harm individuals by providing false information. News agencies have developed AI anchors to enhance efficiency and reduce costs [59]; It implies that it is possible to create fake news with virtual anchors that look like people.

2) COMPUTER AS TARGET CRIME

AI can complete tasks that have been previously unsolved, with even lower cost and labor. By making copies of the AI system, it can have a similar effect as hiring more human analysts. This characteristic gives attackers an opportunity to gain unauthorized access. For example, password authentication, the most fundamental technique of authenticating users, would be under the threat. Dictionary attack, regarded as one of the most effective ways to obtain the password, uses well-known words or phrases expected to have been used in the password [60], [61]. When creating the dictionary, social engineering technique that obtains victim's information from online (e.g. birthday, phone number, address, etc.) is used mostly [62]. The collecting information requires considerable time and effort, but the AI systems designed to automate social engineering can carry out the task effortlessly.

Automated detection techniques to find vulnerabilities would be a useful instrument for criminals. Russell *et al.* [63] provided the potential of using AI to detect vulnerabilities. They demonstrated that the usage of the convolutional neural network (CNN) and the tree ensemble has some advantages over traditional static analysis. Grieco *et al.* [64] presented a method to discover large-scale vulnerabilities. By using the proposed method, programs that have vulnerability could be identified without analyzing source code. Besides those studies, various methods to detect vulnerability have been actively researched. [65]–[67]. Though they were designed for the public good, perpetrators may use the techniques for finding vulnerable systems.

B. PHYSICAL CRIME

The AI security threat extends beyond cyberspace, particularly with the widespread use of IoT [68]. By manipulating AI system, a perpetrator can physically attack a target (e.g. human, pet, vehicle, house).

With respect to physical crime, the ethics of AI have been discussed in science ethics field. Lin *et al.* [69] represented robot ethics with an explanation that AI robots can kill people with or without intention. Scherer [70] also stressed that AI system can cause harm physically and there are arising challenges from difficulties in assigning moral and legal

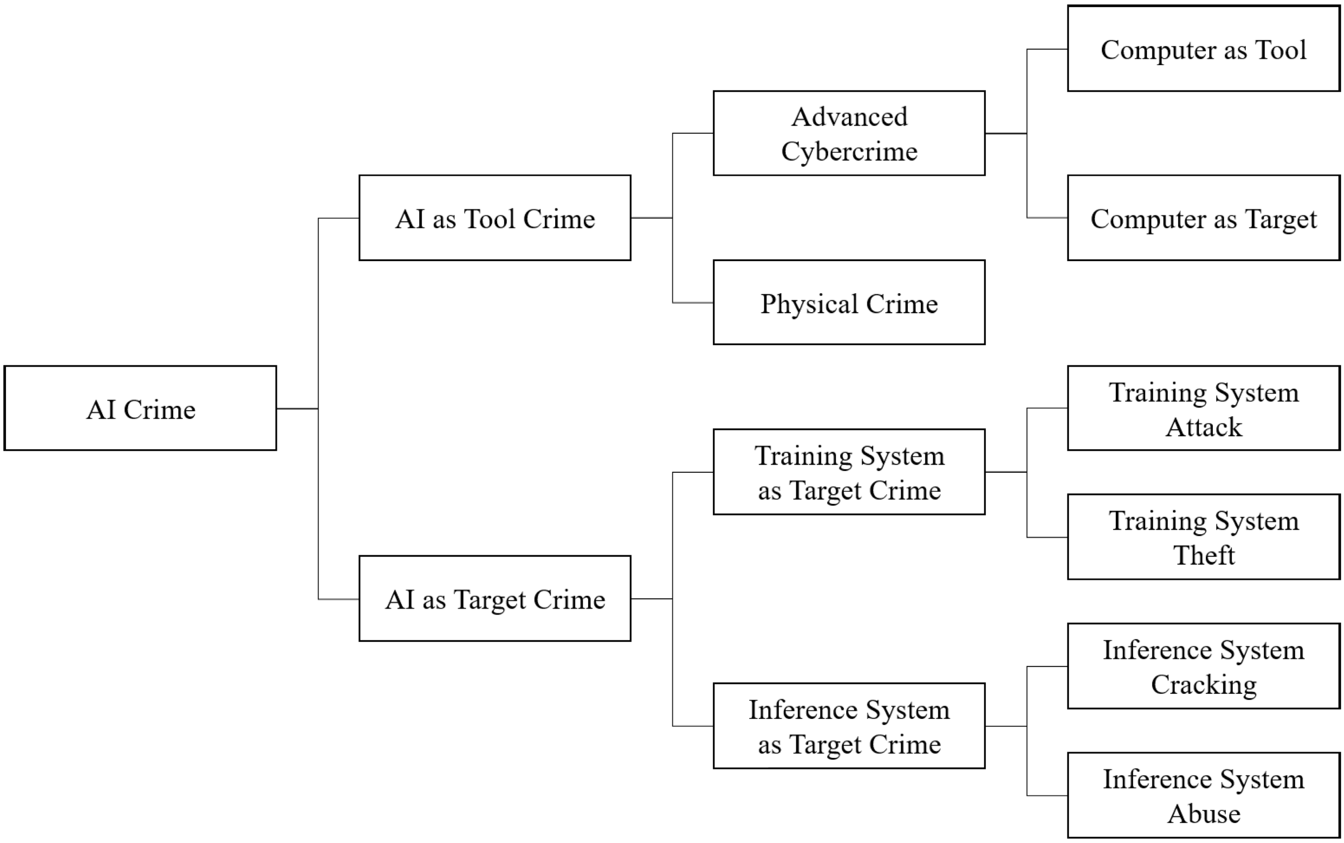

FIGURE 1. The proposed Taxonomy of the AI crime.

responsibility for the harm. The studies focused on harms that occurred by malfunction of autonomous devices.

On the other hand, AI inherently designed to attack physical targets has also been developed for military use; it is called military AI [71]. The military AI is developed for the public good, however, it can also be used as a technique to harm people on the outside of a military context [10]; the drone swarm is a typical example. To operate the drone swarm, the following requirements should be met, according to [72], [73].

- Autonomous (not under centralized control)
- Capable of sensing their local environment and other nearby swarm participants
- Able to communicate locally with others in the swarm
- Able to cooperate to perform a given task

Algorithms by traditional programming were difficult to meet the requirements, but the progress in AI enables swarming.

This swarming technology is applicable to robotic systems or vehicles; it may amplify the synergy with *computer as target crime*. Several studies already showed that it may be possible to remotely manipulate vehicles by exploiting vulnerabilities. Jafarnejad *et al.* [74] proposed possible attack scenarios on Renault Twizy 80, electronic car, by exploiting vulnerabilities of Sevcon Gen4 controller that is Electronic Control Units (ECU) installed in the Twizy. Though there was a limitation that the proposed method is only applicable when the car is turned on, they presented that the attacker can remotely control the vehicle system after hacking. Martinelli *et al.* [75] also presented the vulnerability of the Controller Area Network (CAN) protocol, regarded as the standard for the in-vehicle network. Based on the vulnerability, they were able to perform the message injection attack to cause malfunctions of ECU. The use of these cyberattack with the swarm technology is a serious threat as it can cause a lot of damage to physical space.

IV. AI AS TARGET CRIME

This article defines *AI as target crime* as an offence causing damage or impairment in processing data or operating AI system. This definition is inspired by a definition of *computer as target crime* from [76].

There are various AI systems, but the underlying concept of most AI systems can be expressed in Fig. 2. The AI system consists of training system and inference system. The training system generates a trained model based on training dataset. The trained model is used at inference system to classify the new data from endpoints. For instance, in Fig. 2, the training system creates an algorithm that distinguishes dogs from cats. The inference system loads the algorithm and then it

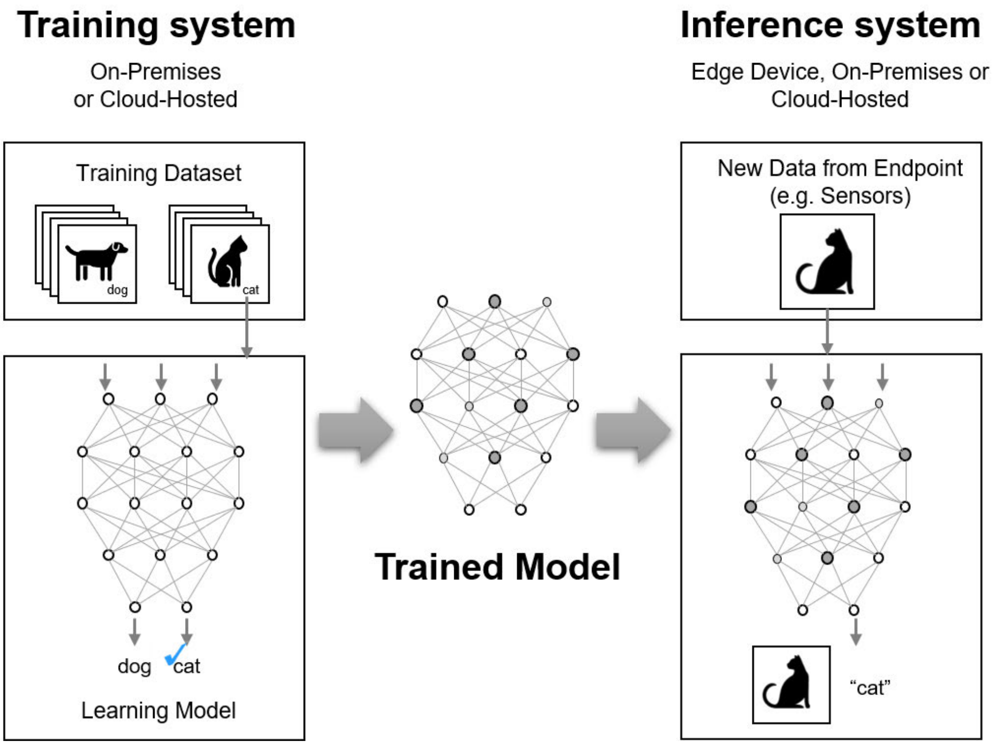

FIGURE 2. The structure of AI system.

determines whether the object image obtained from sensors is a dog or cat.

AI as target crime is committed primarily based on security threats of AI system. Several articles dealt with taxonomy of the security threats; white-box and black-box attacks are the typical threat model. The attack with knowledge of dataset, architecture, and parameters of targeted AI system is called the white-box attack. Whereas the black-box attack contains little or no information about the structure of the targeted system [77]–[79].

Some studies proposed threat models on adversarial example(AE)s. The AEs are input data with invisible noise, in order to misclassify the input and degrade the performance of AI [2]. They focused on impacts when malicious data is injected in training phase [80]–[82] or in inference phase [83]–[85]. The experiments demonstrated the performance reduction of AI system attacked by AEs such as malware detection, facial recognition, intrusion detection, etc.

Some researches categorized the security threats against the training phase and inference phase. Liu *et al.* [86] surveyed a variety of security threats and categorized them into the poisoning, evasion, impersonate, and inversion attacks. The poisoning attack is performed in the training phase, meaning an attacker injected AEs to the training data set [80], [87]. The evasion, impersonate, and inversion attacks behave in the inference phase. The evasion attack means that an attacker deteriorates the security of target systems by using AEs that can evade detection [88]. The impersonate attack means that imitated data samples that are able to wrongly classify the original samples are input to the inference system [89]. The inversion attack is applied to the output of the AI system to infer certain features of the input [90]. Papernot *et al.* [91] also proposed a comprehensive insight into the threat model. They presented the attack surface of AI systems, the trust model, adversarial capabilities, and adversarial goals. Adversarial settings on the training and inference system were also addressed. The model targets the integrity, privacy, and confidentiality of the training system. They also represent white-box and black-box adversaries of the inference system.

Referring to previous efforts, this article presents *AI as target crime* from the perspective of the victims. We focus on AI system including the training system and inference system, which would be targets of the attacks mentioned above.

A. TRAINING SYSTEM AS TARGET CRIME

Since the training system in practical AI system is protected with high confidentiality and not developed in common computer system [86], direct access to the training system seems hard to be achieved. However, it may be accomplished by insider spy, advanced persistent threat (APT), or malicious external storage (e.g. USB, external hard drive). If the security of the training system is compromised, there would be considerable damage to the AI system. Particularly, the training system includes a training dataset that significantly influences the performance of learning model; this is the very reason that crimes against the training system would be catastrophic. The following section describes the AI crimes by assuming that attackers have already intruded on the training system because investigation of the intrusion is the domain of traditional cyber forensics.

1) TRAINING SYSTEM ATTACK

The purpose of this crime is to reduce confidence of AI. By injecting AEs or modifying the existing dataset, AI may misclassify new data from the inference system. If a perpetrator can manipulate the learning algorithm, named logic corruption [91], there would be relatively more critical damage to the training system. This article classify the training system attack into three categories: data injection, data modification, and logic corruption.

Data injection crime disrupts the availability of AI system via injecting AEs. Goodfellow *et al.* [92] presented that AI erroneously recognizes a panda's picture as gibbon by adding a noise that people cannot perceive. The objective of AEs is to find the smallest perturbation deceiving AI.

$$\vec{x^*} = \vec{x} + \arg\min\{\vec{z} : \tilde{O}(\vec{x} + \vec{z}) \neq \tilde{O}(\vec{x})\} \qquad (1)$$

The x is original data and z is a perturbation that is the noise added to the original data to make it an AE x^*. The O is an oracle, which is a system that responds to every unique query, mainly used in the cryptography community [91] The method to generate and utilize AEs have been actively studied, especially for image recognition (See Fig. 3). Nguyen *et al.* [93] proposed a methodology to produce AEs totally unrecognizable to human eyes by using the multi-dimensional archive of phenotypic elites (MAP-Elites). In addition, Eykholt *et al.* [94] presented that AEs can be applied to physical space (e.g. self-driving car) by proposing a possible attack to misclassify signs on the road. Several studies also have represented that AEs can be generated and utilized for disturbing malware detection [95], [96] and intrusion detection [97], [98]; the forensic investigators should be aware of data injection crime.

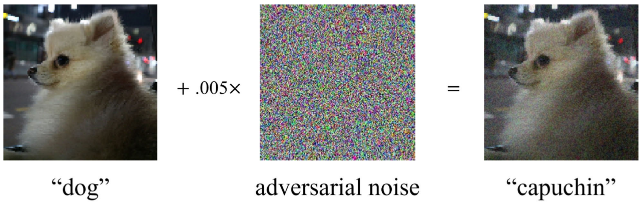

FIGURE 3. **Example of adversarial examples (AEs).**

If perpetrators have permission to modify or delete some training data, they can perform fatal attacks to AI system; it is the data modification crime. Zhao *et al.* [79] presented that label contamination attack (LCA) can significantly reduce the performance of AI by changing the labels of some training data. Hayes and Ohrimenko [99] showed that the accuracy of the classifier is obviously compromised by providing contaminated attacks to training system.

Logic corruption is the most serious crime in the training system. The criminals can manipulate the architecture and parameters of the trained model by tampering the learning algorithm [91]. For example, when the CNN system is attacked and then corrupted, the attacker can handle the input layer, classification layer, and training options.

2) TRAINING SYSTEM THEFT

Training system includes training dataset, learning model, and trained model. As they are directly related to the performance of AI, AI developers and manufacturers of AI-related products consider the training system as trade secrets.

The dataset is very important for AI stakeholders [100]. They create dataset by collecting data from various sources including open-source data (e.g. driving-related data [101], [102] and object data [103], [104]). Since the making dataset takes considerable time and labor, it has high economic value. For his reason, the dataset is favourite target for perpetrators. Indeed, serious privacy infringement may occur if perpetrators steal private data such as medical image, face image, and voice [105], [106].

The learning model and trained model are also important assets for AI developers because they are designed with know-how, insight, and expertise. Through this crime, the algorithm, distribution of training data, and parameters of fully trained model architecture could be leaked to the adversaries or public. In particular, the information may also be abused for white-box attack or black-box attack partially [2].

B. INFERENCE SYSTEM AS TARGET CRIME

Perpetrators may also attack inference system. Comparing to the attack to the training system, perpetrators can access the inference system relatively easily, because the inference system is usually implemented at end devices. The crime targeting inference system does not interfere with the learning model, however, it can cause the leakage of the trained model or malfunction of classification.

1) INFERENCE SYSTEM CRACKING

The parameters, which have been determined at the training phase, play an important role in the inference system. There are two types of operation methods depending on the location of the parameters: centralized and distributed model (See Fig. 4).

In the centralized model, a central server developed by an AI provider takes the inference operation. For example, when using a face recognition system developed for the centralized model, the role of end device (e.g. smartphone, IoT device, in-vehicle infotainment) is to send face images or extracted features to the central server, and then the server processes the image or feature. The end devices operate based on the results processed by the server.

The centralized model is theoretically appropriate for maintenance and security of AI service, but it may be less

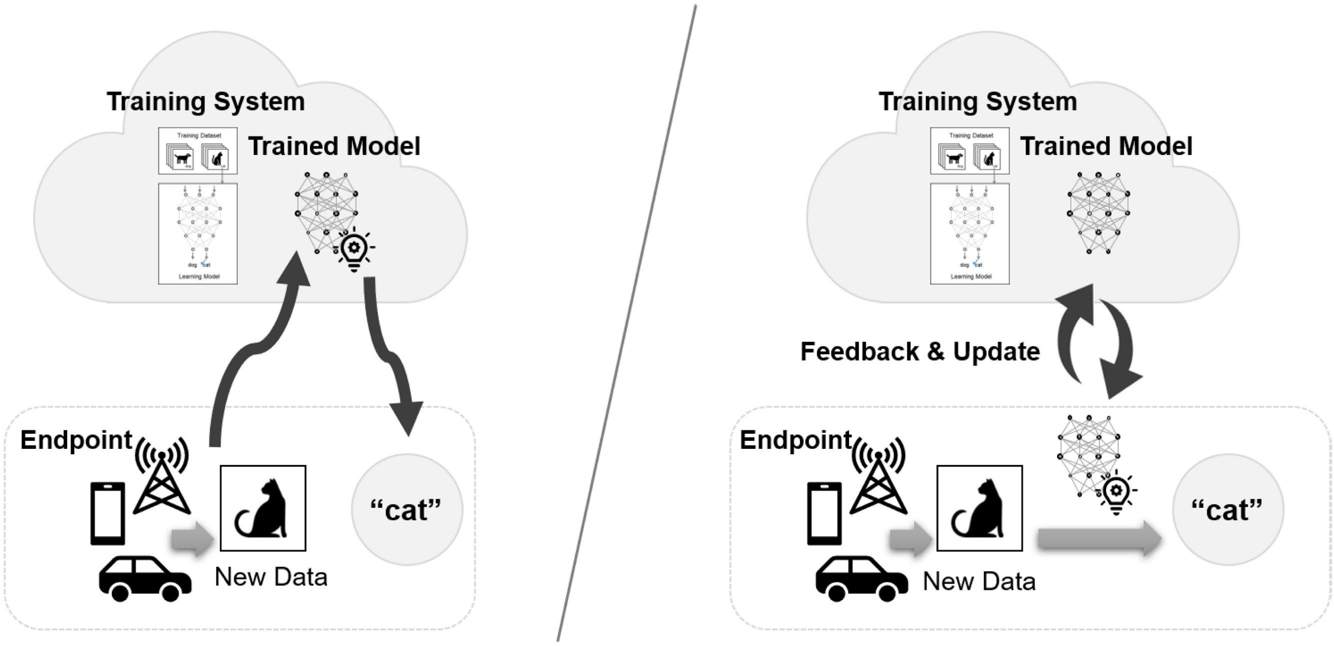

FIGURE 4. Comparison between the centralized model (left side) and distributed model (right side) in AI system.

useful in practice as it may cause a bottleneck. For this reason, the distributed model is being used more in practical fields. In the distributed model, the parameters determined in training system are managed in the inference system so that end devices take the operation of processing image, in the example of face recognition. With the emergence of Internet of Thing (IoT), the use of the distributed model for AI service is taken for granted [107], [108]; the relationship between central and distributed model is similar to that of cloud and fog computing.

The trend relying on the distributed model is beneficial for perpetrators who targets the parameters. By using traditional hacking techniques (e.g. reverse engineering, side channel attack), the trained model would be identified and then manipulated [109]. Indeed, this crime may enable perpetrators to perform white-box attack.

2) INFERENCE SYSTEM ABUSE

The abuse on the inference system is a crime that causes misclassification using AEs. The perpetrators may identify the learning model and its parameters through cracking the end device or noticing that the target system uses common libraries of open-source project [110]. According to the degree of understanding knowledge about the target system, the abuse is classified into white-box attack and black-box attack.

The perpetrators trying to white-box attack have knowledge of the AI model and its parameters. Based on their knowledge, they can simulate the targeted AI model by imitating the AI system and make a fake training dataset as the perpetrators already know the distribution of training data. The typical example of white-box attack is AEs, which is already described in Section IV-A1.

On the other hand, black-box attack is performed with restricted information or without the knowledge of the AI model. The black-box attack can be categorized into non-adaptive black-box attack, adaptive black-box attack, and strict black-box attack.

Having the knowledge of the distribution of training data, the non-adaptive black-box attackers can collect alternative dataset with the distribution although they can not figure out the architecture or structure of the target AI model. They can actually make AEs based on their local AI model trained by the alternative dataset. Generative adversarial networks (GAN) [111] is an example of the non-adaptive black-box attack.

The adaptive black-box attackers use input-output pairs by querying the targeted AI model. This attack is often likened to the oracle attack explained in Section IV-A1. Through collecting amounts of query data, the attackers can identify labels of queried data and then may reconstruct the model with the queried data corpus [112].

The strict black-box attack is also based on collecting input-output pairs, but this attack is more restricted than the adaptive black-box because the attackers can not issue queries to the inference system. Therefore, they should attack AI system without the oracle. Nevertheless, it may be powerful if the attackers obtain many input-output pairs and find a pattern or distribution of them [113].

V. AI FORENSICS

Forensic investigators should collect and analyze evidence to identify 5W1H (when and where the crime is committed, who is criminal, what is targeted, why the criminal commit, and how the crime occurs). As the method of collecting evidence and the type of digital data are heterogeneous

TABLE 1. The AI crimes and AI Forensics.

AI Crime	Techniques	Related Research	AI Forensics Challenges
Advanced Computer as Tool Crime	AI chatbot, Deep fake, etc.	[9], [10], [23], [49], [50], [56]–[58]	AI Exploration
Advanced Computer as Target Crime	Social engineering, Vulnerability scanner, etc.	[60]–[67]	
Physical Crime	Drone swarm, Hardware attack, etc.	[10], [71]–[75]	
Training System Attack	Data injection with adversarial examples, Data modification, Logic corruption	[2], [79], [91]–[99]	AI Exploration, Adversarial Attack Detection, Damage Assessment
Training System Theft	Vulnerability attacks used in computer as target Crime	[105]–[109]	AI Exploration, Similarity Analysis
Inference System Cracking			
Inference System Abuse	White-box attack, black-box attack (non-adaptive, adaptive, and strict)	[110]–[113]	AI Exploration, Adversarial Attack Detection, Damage Assessment

depending on the device or platform, forensic researchers have presented the challenges and solutions for forensic sub-fields such as smartphone forensic [114], [115], cloud forensics [116], [117], and IoT forensics [37], [38].

In this section, we suggest future research directions of AI forensics to investigate the AI crimes described in Section III and IV. Considering characteristic of AI system and techniques used for committing AI crimes, we describe 4 main parts of AI forensics that have been not covered in the forensic community: AI exploration, similarity analysis, adversarial attack detection, and damage assessment. AI forensics is currently in the beginning phase, so the research topics will inspire forensic researchers. Table 1 summarizes the AI forensics challenges against the AI crimes.

A. AI EXPLORATION

When investigating *AI as tool crime*, it is needed to identify how AI is used in the crime. The investigators should collect and analyze the dataset, learning model, trained model, inference model, and application of the AI system used to commit a crime. Based on the examination, investigators should also grasp the purpose of the AI. In this context, identifying a difference between the intention of developer and the result of AI is important for investigators.

Unlike traditional programming, AI program often result unintended consequences. Fig. 5 shows relationship between input, output, and program in traditional programming and AI. In traditional programming, data and program is processed on the computer to produce the output; otherwise, data and output are used to create a program in AI. In particular, parameters of AI are often determined with some randomness because many AI models use random weights in the learning phase. Therefore, even if same dataset and learning model are given, it may create programs that have different parameters and result different outputs. It means that it is hard to prove whether AI was actually used as a weapon, how AI was used, and how much damage AI caused, because investigators would fail to reproduce the case.

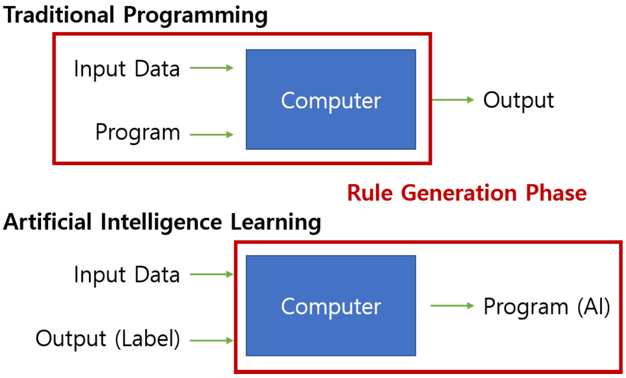

FIGURE 5. Traditional programming and artificial intelligence learning inspired by [118].

Another issue in practical AI forensics is complexity. Generally, an AI system uses various algorithms and libraries. Some AI systems rely on a mixture of the application program interfaces (APIs) provided by existing AI systems for their efficiency. This complexity refers that the investigators should have knowledge of AI and skill for reverse engineering the AI system.

The fact that only some elements of the AI system can be collected whirls forensic stakeholders. It is becoming nearly impossible to collect all elements such as training system, learning model, dataset, trained model, and inference system, due to technical and legal issues. Because perpetrators try to destroy traces of crime, a study is needed to identify the structure and activity history of AI using only limited information. The limited collection of evidence makes it difficult for investigators to reproduce the AI system that must be investigated. It is a important issue for the investigators, because, as described in Section II-B1, the reproducibility is one of the key principles of digital forensics. To show that it is difficult to reproduce the past state of AI system with limited collection data, we presents a simple experiment.

An experiment was conducted using a i7-8700 processor and a Nvidia GeForce 1070 Ti graphic card. We trained several models to categorize binary file into Malware or Benign. Assuming that the investigator had only collected only a portion of the dataset, we trained models according to the size of the dataset. The 1,000 PE files for each category were used as dataset. The 500 Malware files are collected from VirusShare,[1] which is publicly-available repository. The 500 Benign files are collected from Software Informer,[2] which is the most trustworthy sources-provider of benign files, and from system directories created when Windows 10 is newly installed. The model was based on Convolutional Neural Netowrk (CNN) and a voting-based ensemble technique was used to improve the performance for the model. The upper side of Fig. 6 shows accuracy of the models with variations of the size of the dataset. In order to identify performance changes in dataset selection, 60 percent of dataset were randomly selected 10 times, and then we trained models with the selected data. The under side of Fig. 6 shows accuracy of the models. It is shown that there is a deviation in accuracy depending on the data selected for training.

The experimental results show that it is impractical to reproduce the AI system with limited evidence. Indeed, because many AI systems adopt transfer learning where pre-trained models are used as the starting point, obtaining origin data will become more challenging. Therefore, technical and policy approach to overcome the challenge should be studied.

B. SIMILARITY ANALYSIS

Investigating violations, such as copyright infringement, leaking of the confidential document, and invasion of privacy, is a traditional field of digital forensics. In these cases, similarity analysis is one of the most important methods to identify the criminal activities (e.g. code plagiarism detection [119], document similarity [120], and digital image similarity [45]).

As described in Section III and IV, the training dataset, trained model, and learning model would be stolen by perpetrators. In this context, previous studies for similarity analysis can be used to compare between original dataset and suspicious dataset as the dataset consists of data type researched (e.g. image, text, audio, video, etc. [44], [45], [121]).

Identifying the similarity between two models is more complex when the models are based on a specific dataset. If the investigators could not collect the specific dataset, verifying or testing the models is more exhausting. For example, AI developed to distinguish or mimic a specific person, would not be able to be validated or tested, if the investigators can not obtain training data for some reasons (e.g. the person's rejection, death, or disappearance). To respond to the theft, similarity analysis for AI with or without a training dataset should be studied.

A study for a file format that stores trained model is also one of the important research area in AI forensics. For example, Python packages used for AI development, such as Keras and PyTorch, store and manage the trained model and parameters as a HDF5 file that is a binary data format unexplored in the forensics field [122]. In particular, the file is used to distribute the updated model to their edge devices in the distributed model described in Section IV-B1. Therefore, similarity comparison for models is essential to resolve infringement case, but the existing similarity algorithms are cannot be applied to the trained models.

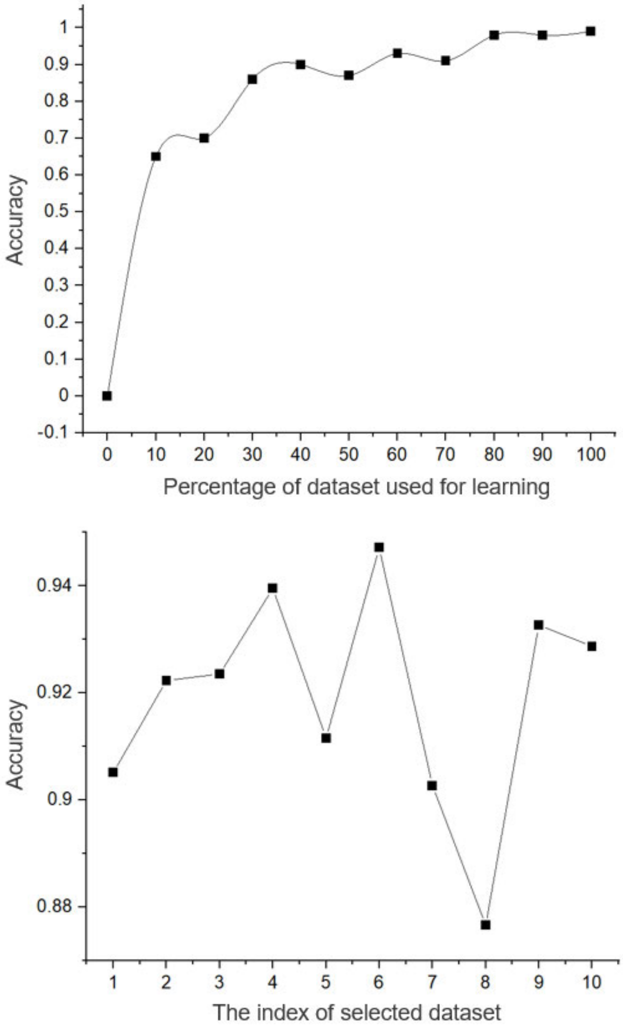

FIGURE 6. **Comparison of the accuracy of the trained models.**

To describe this challenge, we conduct similarity detection for the 10 trained models created in Section V-A. Because some of the training data was shared and the same learning model was applied, we say that the models are similar. We extract the trained models as HDF5 files and then calculate the probability of similarity between the files by using `ssdeep` [123] and `sdhash` [124] that are widely used in the digital forensics field. As seen in Table 2, results of `ssdeep` are all zero and results of `sdhash` shows smaller than three; the algorithms determine that the models are not similar because the threshold of `sdhash` is generally

[1] https://virusshare.com/ (last accessed 28 September 2020)

[2] https://software.informer.com/ (last accessed 28 September 2020)

TABLE 2. Results of `ssdeep` and `sdhash` matches. In each cell, the left figure represents `ssdeep` comparison result, and the right figure represents `sdhash` comparison result. For example, '0-1' means that `ssdeep` score is 0 and `sdhash` score is 1.

No.	1	2	3	4	5	6	7	8	9	10
1	100-100	0-1	0-1	0-2	0-1	0-1	0-1	0-2	0-1	0-1
2	-	100-100	0-1	0-1	0-2	0-2	0-2	0-2	0-2	0-1
3	-	-	100-100	0-1	0-1	0-1	0-1	0-1	0-2	0-1
4	-	-	-	100-100	0-1	0-2	0-1	0-1	0-1	0-2
5	-	-	-	-	100-100	0-1	0-2	0-2	0-1	0-2
6	-	-	-	-	-	100-100	0-1	0-1	0-2	0-2
7	-	-	-	-	-	-	100-100	0-1	0-1	0-2
8	-	-	-	-	-	-	-	100-100	0-1	0-2
9	-	-	-	-	-	-	-	-	100-100	0-1
10	-	-	-	-	-	-	-	-	-	100-100

21 [125]. This experiment shows the limitation of existing algorithms and identifies the need for new algorithm to calculate similarity between trained models.

C. ADVERSARIAL ATTACK DETECTION

With regard to DFR introduced in Section II-B1, it is important to prevent or detect the adversarial attack proactively. Many researchers proposed defence methods that attempt to classify AEs correctly, but the methods have been being defeated by newly developed attacks [126]–[128].

As it is difficult to defend against adversarial attacks, recent works have attempted to detect AEs instead [129]. Some works approached the problem statistically like two-sample hypothesis testing [130], principal component analysis (PCA) [131], and Bayesian uncertainty estimates [132]. Methods that use an additional neural network [133], [134] or an external model [135], [136] were also proposed.

The several techniques detect the adversarial attacks known at the time, but state-of-the-art AEs attacks that neutralize the detection techniques have been also developed. Because neural-network based classifiers have inherent vulnerability that leads to misclassification, it is fundamentally impossible to prevent current and foreseeable attacks. Therefore, making it difficult and time consuming to create AEs is considered as an alternative [137]. Nevertheless, a threat of adversarial attacks still exists during incremental training. The aim of incremental learning is to adapts to new data and to improve the model continuously, an attacker has an opportunity to insert AEs into AI system by deceiving AEs as Benign. In the current situation, as the technique of making AEs will become more sophisticated, enhancing detection technique remains on open issue for forensic researchers.

D. DAMAGE ASSESSMENT

The extent of the damage caused by AI crime should be ascertained by the forensic investigators. With respect to attack using AEs, the investigators need to identify which data are AEs, how many AEs were actually injected, and how much it affected the confidence.

Theoretically, the finding AEs is to identify data that raise the prediction error. We explain the process with the deep neural network (DNN) as an example. DNN uses a hierarchical composition of n parametric functions f_i. Each f_i for $i \in 1..n$ is modeled using a layer of neurons and each layer is parameterized by a weight vector θ_i. A DNN model F that is computed as follow:

$$F(\vec{x}) = f_n(\theta_n, f_{n-1}(\theta_{n-1}, \ldots f_2(\theta_2, f_1(\theta_1, \vec{x})))) \tag{2}$$

Assuming that AEs were already injected to the target dataset and investigators have knowledge of training F and dataset of input-output pairs $(\vec{x}, \vec{y})$, the AEs can be found using backward elimination method as follow:

$$k^* = \arg\min\{ \sum_{\neg(j=k)} \left| F_j(\vec{x_j}) - \vec{y_j} \right| \} \tag{3}$$

The k could be a single or collection of samples. The model F_j means trained DNN model without x_j, f_j, and θ_j. The example calculates the prediction error as the number of misclassification.

However, the approach is difficult to apply practical forensic investigation because it requires to calculate equation (2) and (3) n times. Because the AI algorithm including DNN has a large amount of training data, it is practically impossible to calculate the impact of each sample. The investigators also may not be able to obtain information about all samples of the dataset or AI model; this situation further complicates the problem. Therefore, it is a significant challenge to find an optimized method to identify AEs and calculate damage with limited knowledge of the AI model.

VI. DISCUSSION

Based on our observation from surveying the security threat of AI and exploring foreseeable AI crimes, this section highlights open issues in the context of AI forensics through comparison with traditional forensics. Table 3 shows related issues on the principles of traditional forensics.

A. LARGE-SCALE

Generating an AI model requires considerable resources and data, which would have been unimaginable before. This large-scale nature makes it even difficult to find data for investigators using forensic tools programmed in traditional computing. Even in traditional digital forensics, the large-scale issue has been dealt with, but a much larger number of data should be covered in AI forensics. Particularly, current AIs mainly focus on multimedia data like image

TABLE 3. Comparative Table of traditional Forensics and AI Forensics.

Principle	Traditional Forensics	AI Forensics
Meaning	The evidence and the meaning is unchanged once collected.	The meaning of AI system would be changed in learning phase.
Errors	Errors can be documented in forensic process	
Transparency and trustworthiness	Many methods have been tested and verified.	It is needed to verify AI forensic process.
Reproducibility	It shows a consistent level of quality.	It would be impossible even with identical dataset and learning model.
Experience	Various research and education exist.	It is needed to be studied.

or sound, which are even still challenges in digital forensics field. At the current level of forensic technology, the basic forensic process that collects the evidence at the scene of the crime and then analyzes the evidence in lab is needs to be adjusted to accommodate the AI system environment in practice.

B. IRREPRODUCIBILITY

The fact that the AI systems have inherent unpredictability would influence the forensic principles. Most AI algorithms use random values partially or completely; this nature often fails to satisfy the reproducibility of the forensic principles. If the reproducibility is strictly applied to the evidence of the AI crime such as copycat model and AEs, it can be rejected in the court as the evidence may not reproduce the situation at the time of the incident. Nevertheless, applying the reproducibility principles must be considered carefully, because it may trigger new issues like arresting wrong suspect. Therefore, a compromise between strict and tolerant appliance of the principles should be discussed among forensic examiners, policymakers, and AI professionals.

C. EXPERTISE

Finally, forensic stakeholders also need to develop their expertise in AI. To address challenges of AI forensics, they must have a clear understanding of AI crime and AI forensic techniques. As forensic investigators should understand traditional programming (e.g. memory structure, compiler, assembly language) when analyzing malware [138], they need to have the background knowledge about AI system, AI structure, and AI environment to suggest probable solutions for the AI forensic challenges. To achieve this, forensic researchers should be interested in AI and collaborate with AI stakeholders.

VII. CONCLUSION

AI is becoming widely used in various systems and applications. Due to the dual-use nature, there are also growing concerns that AI can be harmful to people. To perform illegal activities, perpetrators may use AI maliciously or attack AI system by exploiting the inherent vulnerabilities of the victim AI system.

This paper have studied foreseeable AI related crimes. Based on the literature review of security threats of AI and AI-related crime, we have identified that the previous studies focused on the malicious use of AI to sharpen existing criminal techniques or the vulnerabilities of AI algorithms and training dataset. We have also presented that existing crimes would be more powerful with AI and new types of crimes may be appeared, which have not been identified before. To cope with the AI crime, this paper have provided a systematic taxonomy for AI crime: *AI as tool crime* and *AI as target crime*.

Furthermore, we have represented the novel strategies against the AI crime, named AI forensics. By providing comparative analysis of AI forensics and traditional forensics, we have found that some principles of digital forensics are not suitable for AI forensics.

Future works and open issues of AI forensics that inspire forensic researchers to better understand challenges to face have been suggested. We hope that this article can serve as a valuable reference for researchers in digital forensics, security engineering, computer science, and criminology.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[6] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[7] S. V. Albrecht and P. Stone, "Autonomous agents modelling other agents: A comprehensive survey and open problems," *Artif. Intell.*, vol. 258, pp. 66–95, May 2018.

[8] V. C. Müller and N. Bostrom, "Future progress in artificial intelligence: A survey of expert opinion," in *Fundamental Issues of Artificial Intelligence*. Cham, Switzerland: Springer, 2016, pp. 555–572.

[9] T. King, N. Aggarwal, M. Taddeo, and L. Floridi, "Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions," *SSRN Electron. J.*, vol. 26, no. 1, pp. 1–32, 2019.

[10] M. Brundage *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," 2018, *arXiv:1802.07228*. [Online]. Available: http://arxiv.org/abs/1802.07228

[11] S. Gordon and R. Ford, "Cyberterrorism?" *Comput. Secur.*, vol. 21, no. 7, pp. 636–647, 2002.

[12] M. Yar and K. F. Steinmetz, *Cybercrime and Society*. Newbury Park, CA, USA: Sage, 2019.

[13] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1765–1773.

[14] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "Key challenges in defending against malicious socialbots," presented at the 5th USENIX Workshop Large-Scale Exploits Emergent Threats, 2012.

[15] R. W. Gehl and M. Bakardjieva, "Socialbots and their friends," in *Socialbots and Their Friends*. Evanston, IL, USA: Routledge, 2016, pp. 17–32.

[16] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, and J. H. Park, "Social network security: Issues, challenges, threats, and solutions," *Inf. Sci.*, vol. 421, pp. 43–69, Dec. 2017.

[17] J. Seymour and P. Tully, "Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter," *Black Hat USA*, vol. 37, pp. 1–39, Aug. 2016.

[18] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "BotOrNot: A system to evaluate social bots," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 273–274.

[19] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016.

[20] N. Abokhodair, D. Yoo, and D. W. McDonald, "Dissecting a social botnet: Growth, content and influence in Twitter," in *Proc. 18th ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, 2015, pp. 839–851.

[21] M.-A. Rizoiu, T. Graham, R. Zhang, Y. Zhang, R. Ackland, and L. Xie, "#DebateNight: The role and influence of socialbots on Twitter during the 1st 2016 us presidential debate," in *Proc. 12th Int. AAAI Conf. Web Social Media*, 2018.

[22] G. Dvorsky. *Hackers Have Already Started to Weaponize Artificial Intelligence*. Gizmodo.com. [Online]. Available: https://gizmodo.com/hackers-have-already-started-to-weaponize-artificial-in-1797688425

[23] O. Bendel, "The synthetization of human voices," *AI & Soc.*, vol. 34, no. 1, pp. 83–89, 2019.

[24] G. Allen and T. Chan, *Artificial Intelligence and National Security*. *Belfer Center for Science and International Affairs*. Cambridge, MA, USA: Belfer Center for Science and International Affairs, 2017. [Online]. Available: https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf

[25] X. Li and T. Zhang, "An exploration on artificial intelligence application: From security, privacy and ethic perspective," in *Proc. IEEE 2nd Int. Conf. Cloud Comput. Big Data Anal. (ICCCBDA)*, Apr. 2017, pp. 416–420.

[26] L. Mitrou, *Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) 'Artificial Intelligence-Proof'?*. SSRN, 2018. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3386914

[27] K. Dashora, "Cyber crime in the society: Problems and preventions," *J. Alternative Perspect. Social Sci.*, vol. 3, no. 1, pp. 240–259, 2011.

[28] S. Schjolberg and S. Ghernaouti-Helie, "A global treaty on cybersecurity and cybercrime," *Cybercrime Law*, vol. 97, 2011. [Online]. Available: http://pircenter.org/kosdata/page_doc/p2732_1.pdf

[29] T. Tropina and C. Callanan, *Self- and Co-regulation in Cybercrime, Cybersecurity and National Security*. Cham, Switzerland: Springer, 2015.

[30] M. Brand, C. Valli, and A. Woodward, "Malware forensics: Discovery of the intent of deception," *J. Digit. Forensics, Secur. Law*, vol. 5, no. 4, p. 2, 2010.

[31] C. H. Malin, E. Casey, and J. M. Aquilina, *Malware Forensics Field Guide for Windows Systems: Digital Forensics Field Guides*. Amsterdam, The Netherlands: Elsevier, 2012.

[32] B. Ruttenberg, C. Miles, L. Kellogg, V. Notani, M. Howard, C. LeDoux, A. Lakhotia, and A. Pfeffer, "Identifying shared software components to support malware forensics," in *Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment*. Cham, Switzerland: Springer, 2014, pp. 21–40.

[33] B. Carrier and E. H. Spafford, "An event-based digital forensic investigation framework," in *Proc. Digit. Forensic Research Workshop*, 2004, pp. 11–13.

[34] L. Pan and L. Batten, "Reproducibility of digital evidence in forensic investigations," in *Proc. 5th Annu. Digit. Forensic Res. Workshop (DFRWS)*, 2005, pp. 1–8.

[35] R. McKemmish, "When is digital evidence forensically sound?" in *Proc. IFIP Int. Conf. Digit. Forensics*. Boston, MA, USA: Springer, 2008, pp. 3–15.

[36] M. Damshenas, A. Dehghantanha, and R. Mahmoud, "A survey on digital forensics trends," *Int. J. Cyber-Secur. Digit. Forensics*, vol. 3, no. 4, pp. 209–235, 2014.

[37] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, "Internet of Things security and forensics: Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 78, pp. 544–546, Jan. 2018.

[38] J. Hou, Y. Li, J. Yu, and W. Shi, "A survey on digital forensics in Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 1–15, Jan. 2020.

[39] F. Amato, A. Castiglione, G. Cozzolino, and F. Narducci, "A semantic-based methodology for digital forensics analysis," *J. Parallel Distrib. Comput.*, vol. 138, pp. 172–177, Apr. 2020.

[40] V. R. Kebande and H. S. Venter, "Novel digital forensic readiness technique in the cloud environment," *Austral. J. Forensic Sci.*, vol. 50, no. 5, pp. 552–591, Sep. 2018.

[41] R. Rowlingson, "A ten step process for forensic readiness," *Int. J. Digit. Evidence*, vol. 2, no. 3, pp. 1–28, 2004.

[42] E. B. Karbab and M. Debbabi, "MalDy: Portable, data-driven malware detection using natural language processing and machine learning techniques on behavioral analysis reports," *Digit. Invest.*, vol. 28, pp. S77–S87, Apr. 2019.

[43] E. Fidalgo, E. Alegre, L. Fernández-Robles, and V. González-Castro, "Classifying suspicious content in Tor darknet through semantic attention keypoint filtering," *Digit. Invest.*, vol. 30, pp. 12–22, Sep. 2019.

[44] W. Anwar, I. S. Bajwa, M. A. Choudhary, and S. Ramzan, "An empirical study on forensic analysis of urdu text using LDA-based authorship attribution," *IEEE Access*, vol. 7, pp. 3224–3234, 2019.

[45] O. Mayer and M. C. Stamm, "Forensic similarity for digital images," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1331–1346, 2020.

[46] J. Kietzmann, J. Paschen, and E. Treen, "Artificial intelligence in advertising: How marketers can leverage artificial intelligence along the consumer journey," *J. Advertising Res.*, vol. 58, no. 3, pp. 263–267, 2018.

[47] J. Paschen, M. Wilson, and J. J. Ferreira, "Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel," *Bus. Horizons*, vol. 63, no. 3, pp. 403–414, May 2020.

[48] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "SuperAgent: A customer service chatbot for E-commerce websites," in *Proc. ACL, Syst. Demonstrations*, 2017, pp. 97–102.

[49] S. A. Abdul-Kader and D. John, "Survey on chatbot design techniques in speech conversation systems," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 7, 2015.

[50] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction," *Future Gener. Comput. Syst.*, vol. 92, pp. 539–548, Mar. 2019.

[51] S. D'Alfonso, O. Santesteban-Echarri, S. Rice, G. Wadley, R. Lederman, C. Miles, J. Gleeson, and M. Alvarez-Jimenez, "Artificial intelligence-assisted online social therapy for youth mental health," *Frontiers Psychol.*, vol. 8, p. 796, Jun. 2017.

[52] F. Clarizia, F. Colace, M. Lombardi, F. Pascale, and D. Santaniello, "Chatbot: An education support system for student," in *Proc. Int. Symp. Cyberspace Saf. Secur.* Cham, Switzerland: Springer, 2018, pp. 291–302.

[53] S. Divya, V. Indumathi, S. Ishwarya, M. Priyasankari, and S. K. Devi, "A self-diagnosis medical chatbot using artificial intelligence," *J. Web Develop. Web Designing*, vol. 3, no. 1, pp. 1–7, 2018.

[54] J. L. Z. Montenegro, C. A. da Costa, and R. da Rosa Righi, "Survey of conversational agents in health," *Expert Syst. Appl.*, vol. 129, pp. 56–67, Sep. 2019.

[55] J. M. Burkhardt, "History of fake news," *Library Technol. Rep.*, vol. 53, no. 8, pp. 5–9, 2017.

[56] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[57] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, 2017.

[58] D. K. Citron and R. Chesney, *Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy?* Washington, DC, USA: Lawfare, 2018.

[59] X. Bo, "Xinhua presents AI anchors at news agencies world congress," Xinhua, Beijing, China, Tech. Rep., Jun. 2019. [Online]. Available: http://www.xinhuanet.com/english/2019-06/15/c_138146148.htm

[60] D. Wang and P. Wang, "Offline dictionary attack on password authentication schemes using smart cards," in *Information Security*. Cham, Switzerland: Springer, 2015, pp. 221–237.

[61] A. K. Kyaw, F. Sioquim, and J. Joseph, "Dictionary attack on wordpress: Security and forensic analysis," in *Proc. 2nd Int. Conf. Inf. Secur. Cyber Forensics (InfoSec)*, Nov. 2015, pp. 158–164.

[62] S. Gupta, A. Singhal, and A. Kapoor, "A literature survey on social engineering attacks: Phishing attack," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, Apr. 2016, pp. 537–540.

[63] R. Russell, L. Kim, L. Hamilton, T. Lazovich, J. Harer, O. Ozdemir, P. Ellingwood, and M. McConley, "Automated vulnerability detection in source code using deep representation learning," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 757–762.

[64] G. Grieco, G. L. Grinblat, L. Uzal, S. Rawat, J. Feist, and L. Mounier, "Toward large-scale vulnerability discovery using machine learning," in *Proc. 6th ACM Conf. Data Appl. Secur. Privacy*, 2016, pp. 85–96.

[65] S. M. Ghaffarian and H. R. Shahriari, "Software vulnerability analysis and discovery using machine-learning and data-mining techniques: A survey," *ACM Comput. Surv.*, vol. 50, no. 4, pp. 1–36, 2017.

[66] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, Z. Deng, and Y. Zhong, "VulDeePecker: A deep learning-based system for vulnerability detection," 2018, *arXiv:1801.01681*. [Online]. Available: http://arxiv.org/abs/1801.01681

[67] H. Xue, S. Sun, G. Venkataramani, and T. Lan, "Machine learning-based analysis of program binaries: A comprehensive study," *IEEE Access*, vol. 7, pp. 65889–65912, 2019.

[68] H. Ashrafian, "AIonAI: A humanitarian law of artificial intelligence and robotics," *Sci. Eng. Ethics*, vol. 21, no. 1, pp. 29–40, Feb. 2015.

[69] P. Lin, K. Abney, and R. Jenkins, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. London, U.K.: Oxford Univ. Press, 2017.

[70] M. U. Scherer, "Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies," *Harv. JL Tech.*, vol. 29, no. 2, p. 353, 2015. [Online]. Available: https://heinonline.org/HOL/LandingPage?handle=hein.journals/hjlt29&div=15&id=&page=

[71] M. Cummings, "Artificial intelligence and the future of warfare," Chatham House Roy. Inst. Int. Affairs London, London, U.K., Tech. Rep., 2017. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/59339185/2017-01-26-artificial-intelligence-future-warfare-cummings-final20190521-119589-196oqd3.pdf?1558436451=&response-content-disposition=inline%3B+filename%3DArtificial_intelligence_future_warfare_c.pdf&Expires=1602336950&Signature=HKEsP2BLOP-ejS4xEtWLmUDyK73alNdyoR6VKfwztl9UtE6JWYXedrK5m70LyENBBBiz31Gh5us1fPYBUzisLn9oawuXphtS8n17G9m0pI4kNyL01baxyqdYSKrB7G-gYcmHY2D49fqW7L6Y7voswcghWv1gctqVq5vS8KprSY8zTb6XJCZXdBNjWxbNqpw2nLplzXG7gghignIFYj9ncYGrXcVH6IV7KKYjpX9DfvE01dROHibDLS1Ixhj~ry5faZ6BZvPEJS5UWOVk-T0H91fUIURVW3Y2DRG5tqWWO2V8vJJt8DRFnQCbof70RE9MlgWIgWNKcF04GVE3ul0ZoA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[72] A. Ilachinski, *AI, Robots, and Swarms: Issues, Questions, and Recommended Studies*. Arlington County, VA, USA: CNA Corporation, 2017.

[73] D. S. Hoadley and N. J. Lucas, *Artificial Intelligence and National Security*. Washington, DC, USA: Congressional Research Service, 2018

[74] S. Jafarnejad, L. Codeca, W. Bronzi, R. Frank, and T. Engel, "A car hacking experiment: When connectivity meets vulnerability," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.

[75] F. Martinelli, F. Mercaldo, V. Nardone, and A. Santone, "Car hacking identification through fuzzy logic algorithms," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2017, pp. 1–7.

[76] J. Clough, *Principles of Cybercrime*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[77] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1322–1333.

[78] B. Biggio, "Machine learning under attack: Vulnerability exploitation and security measures," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, 2016, pp. 1–2.

[79] M. Zhao, B. An, W. Gao, and T. Zhang, "Efficient label contamination attacks against black-box learning models," in *Proc. IJCAI*, 2017, pp. 3945–3951.

[80] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, "Data poisoning attacks on factorization-based collaborative filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1885–1893.

[81] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2018, pp. 19–35.

[82] J. Zhang and C. Li, "Adversarial examples: Opportunities and challenges," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2578–2593, Jul. 2019.

[83] K. R. Mopuri, U. Garg, and R. V. Babu, "Fast feature fool: A data independent approach to universal adversarial perturbations," 2017, *arXiv:1707.05572*. [Online]. Available: http://arxiv.org/abs/1707.05572

[84] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The space of transferable adversarial examples," 2017, *arXiv:1704.03453*. [Online]. Available: http://arxiv.org/abs/1704.03453

[85] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1528–1540.

[86] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE Access*, vol. 6, pp. 12103–12117, 2018.

[87] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016.

[88] L. Tong, B. Li, C. Hajaj, C. Xiao, N. Zhang, and Y. Vorobeychik, "Improving robustness of {*ML*} classifiers against realizable evasion attacks using conserved features," in *Proc. 28th USENIX Secur. Symp. (USENIX Secur.)*, 2019, pp. 285–302.

[89] H. Dang, Y. Huang, and E.-C. Chang, "Evading classifiers by morphing in the dark," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 119–133.

[90] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.

[91] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and privacy in machine learning," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Apr. 2018, pp. 399–414.

[92] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*. [Online]. Available: http://arxiv.org/abs/1412.6572

[93] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 427–436.

[94] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1625–1634.

[95] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial examples for malware detection," in *Proc. Eur. Symp. Res. Comput. Secur.* Cham, Switzerland: Springer, 2017, pp. 62–79.

[96] O. Suciu, S. E. Coull, and J. Johns, "Exploring adversarial examples in malware detection," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2019, pp. 8–14.

[97] I. Corona, G. Giacinto, and F. Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues," *Inf. Sci.*, vol. 239, pp. 201–225, Aug. 2013.

[98] N. Martins, J. M. Cruz, T. Cruz, and P. H. Abreu, "Adversarial machine learning applied to intrusion and malware scenarios: A systematic review," *IEEE Access*, vol. 8, pp. 35403–35419, 2020.

[99] J. Hayes and O. Ohrimenko, "Contamination attacks and mitigation in multi-party machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6604–6615.

[100] C. Catal and B. Diri, "Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem," *Inf. Sci.*, vol. 179, no. 8, pp. 1040–1058, Mar. 2009.

[101] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road scene segmentation from a single image," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 376–389.

[102] R. Zhang, S. A. Candra, K. Vetter, and A. Zakhor, "Sensor fusion for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1850–1857.

[103] S. D. Jain and K. Grauman, "Supervoxel-consistent foreground propagation in video," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 656–671.

[104] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas, "A scalable active framework for region annotation in 3D shape collections," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Nov. 2016.

[105] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 19–38.

[106] Y. Mao, S. Yi, Q. Li, J. Feng, F. Xu, and S. Zhong, "A privacy-preserving deep learning approach for face recognition with edge computing," in *Proc. USENIX Workshop Hot Topics Edge Comput. (HotEdge)*, 2018, pp. 1–6.

[107] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, 4th Quart., 2018.

[108] X. Fei, N. Shah, N. Verba, K.-M. Chao, V. Sanchez-Anguix, J. Lewandowski, A. James, and Z. Usman, "CPS data streams analytics based on machine learning for cloud and fog computing: A survey," *Future Gener. Comput. Syst.*, vol. 90, pp. 435–450, Jan. 2019.

[109] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[110] Q. Xiao, K. Li, D. Zhang, and W. Xu, "Security risks in deep learning implementations," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 123–128.

[111] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2014, pp. 2672–2680.

[112] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," 2019, *arXiv:1904.01067*. [Online]. Available: http://arxiv.org/abs/1904.01067

[113] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.

[114] R. Ahmed and R. V. Dharaskar, "Study of mobile botnets: An analysis from the perspective of efficient generalized forensics framework for mobile devices," in *Proc. Int. J. Control Automat. Proc. Nat. Conf. Innov. Paradigms Eng. Technol. (NCIPET)*, 2012, pp. 5–8.

[115] S. Mohtasebi and A. Dehghantanha, "Towards a unified forensic investigation framework of smartphones," *Proc. Int. J. Comput. Theory Eng.*, vol. 5, no. 2, p. 351, 2013.

[116] S. Simou, C. Kalloniatis, S. Gritzalis, and H. Mouratidis, "A survey on cloud forensics challenges and solutions," *Secur. Commun. Netw.*, vol. 9, no. 18, pp. 6285–6314, Dec. 2016.

[117] S. Simou, C. Kalloniatis, S. Gritzalis, and V. Katos, "A framework for designing cloud forensic-enabled services (CFeS)," *Requirements Eng.*, vol. 24, no. 3, pp. 403–430, Sep. 2019.

[118] J. Brownlee. (2020). *Basic Concepts in Machine Learning*. [Online]. Available: https://machinelearningmastery.com/basic-concepts-in-machine-learning/

[119] L. Luo, J. Ming, D. Wu, P. Liu, and S. Zhu, "Semantics-based obfuscation-resilient binary code similarity comparison with applications to software and algorithm plagiarism detection," *IEEE Trans. Softw. Eng.*, vol. 43, no. 12, pp. 1157–1177, Dec. 2017.

[120] F. Benedetti, D. Beneventano, S. Bergamaschi, and G. Simonini, "Computing inter-document similarity with context semantic analysis," *Inf. Syst.*, vol. 80, pp. 136–147, Feb. 2019.

[121] O. Mayer and M. C. Stamm, "Learned forensic source similarity for unknown camera models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2012–2016.

[122] M. Ferguson, S. Jeong, K. H. Law, S. Levitan, A. Narayanan, R. Burkhardt, T. Jena, and Y.-T. T. Lee, "A standardized representation of convolutional neural networks for reliable deployment of machine learning models in the manufacturing industry," in *Proc. Int. Design Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, vol. 59179, 2019, Art. no. V001T02A005.

[123] J. Kornblum, "Identifying almost identical files using context triggered piecewise hashing," *Digit. Invest.*, vol. 3, pp. 91–97, Sep. 2006.

[124] V. Roussev, "Data fingerprinting with similarity digests," in *Proc. IFIP Int. Conf. Digit. Forensics*. Berlin, Germany: Springer, 2010, pp. 207–226.

[125] V. Roussev, "An evaluation of forensic similarity hashes," *Digit. Invest.*, vol. 8, pp. S34–S41, Aug. 2011.

[126] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on Speech-to-Text," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 1–7.

[127] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," 2018, *arXiv:1802.00420*. [Online]. Available: http://arxiv.org/abs/1802.00420

[128] T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 2253–2260.

[129] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, 2017, pp. 3–14.

[130] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (Statistical) detection of adversarial examples," 2017, *arXiv:1702.06280*. [Online]. Available: http://arxiv.org/abs/1702.06280

[131] X. Li and F. Li, "Adversarial examples detection in deep networks with convolutional filter statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5764–5772.

[132] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," 2017, *arXiv:1703.00410*. [Online]. Available: http://arxiv.org/abs/1703.00410

[133] J. Hendrik Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," 2017, *arXiv:1702.04267*. [Online]. Available: http://arxiv.org/abs/1702.04267

[134] H. Hosseini, Y. Chen, S. Kannan, B. Zhang, and R. Poovendran, "Blocking transferability of adversarial examples in black-box learning systems," 2017, *arXiv:1703.04318*. [Online]. Available: http://arxiv.org/abs/1703.04318

[135] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 135–147.

[136] S. Tian, G. Yang, and Y. Cai, "Detecting adversarial examples through image transformation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[137] N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 658–659.

[138] E. Eilam, *Reversing: Secrets of Reverse Engineering*. Hoboken, NJ, USA: Wiley, 2011.

DOOWON JEONG received the B.S. degree from the Division of Industrial Management Engineering, Korea University, in 2011, and the Ph.D. degree from the Graduate School of Information Security, Korea University, in 2019. He is currently an Assistant Professor with the College of Police and Criminal Justice, Dongguk University. His research interests include digital forensics, information security, artificial intelligence, and digital profiling.