The AI-Based Cyber Threat Landscape: A Survey

NEKTARIA KALOUDI and JINGYUE LI, Norwegian University of Science and Technology, Norway

Recent advancements in artificial intelligence (AI) technologies have induced tremendous growth in innovation and automation. Although these AI technologies offer significant benefits, they can be used maliciously. Highly targeted and evasive attacks in benign carrier applications, such as DeepLocker, have demonstrated the intentional use of AI for harmful purposes. Threat actors are constantly changing and improving their attack strategy with particular emphasis on the application of AI-driven techniques in the attack process, called AI-based cyber attack, which can be used in conjunction with conventional attack techniques to cause greater damage. Despite several studies on AI and security, researchers have not summarized AI-based cyber attacks enough to be able to understand the adversary's actions and to develop proper defenses against such attacks. This study aims to explore existing studies of AI-based cyber attacks and to map them onto a proposed framework, providing insight into new threats. Our framework includes the classification of several aspects of malicious uses of AI during the cyber attack life cycle and provides a basis for their detection in order to predict future threats. We also explain how to apply this framework to analyze AI-based cyber attacks in a hypothetical scenario of a critical smart grid infrastructure.

CCS Concepts: • **General and reference → Surveys and overviews**.

Additional Key Words and Phrases: Cyber security, AI attacks, cyber threat prevention, cyber-physical systems, smart grid, attack analysis

1 INTRODUCTION

Over the past years, artificial intelligence (AI) technologies have progressed rapidly and their capabilities have extended into several domains. From smart governance, smart buildings, smart transportation, smart grids to smart "anything", AI turns the flood of data into actionable information. These AI technologies are useful for the cybersecurity field by collecting large amounts of data and then quickly filtering them to detect malicious patterns and anomalous behaviors. Therefore, a lot has been published with a focus on the advancements of AI, but less attention has been given to the dangers of AI. The malicious use of AI is altering the landscape of potential threats against a wide range of beneficial applications. Particularly, the threat of malicious use of AI could threaten more complex systems such as smart cyber-physical systems, which have not been studied thoroughly before. Smart cyber-physical systems (sCPS) refer to advanced CPS systems, which are more interconnected through various technologies like the Internet of Things (IoT), AI, wireless sensor networks (WSN), and cloud computing to provide a wide range of innovative services and applications [1].

To a large extent, the interconnected nature of sCPS means a single vulnerability like the flap of a butterfly's wings can ultimately cause a tornado. Analogous to the "butterfly effect", when one part of a system collapses, the whole system will collapse with large effects. Therefore, the increasing levels of interconnectivity and autonomy have given rise to an increased number of cyber attacks. The impact of potential cyber threats has been extended from malicious uses of AI technologies to enable larger-scale and more powerful attacks. Cybercriminals have started to improve their techniques by including IoT hacks, malware, ransomware, and AI to launch more powerful attacks. By carrying out these kinds of attacks, everyone is at risk due to the interconnectivity and intelligence of the attacks. From this perspective, even if the progress of research on the application of AI to defend against cyber attacks has already started many years

Authors' address: Nektaria Kaloudi, nektaria.kaloudi@ntnu.no; Jingyue Li, jingyue.li@ntnu.no, Norwegian University of Science and Technology, Trondheim, Norway.

ago [12], there is still uncertainty about how to defend against AI being used as a malicious tool. This work attempts to fill this gap.

The goal of this work is to identify, analyze, and classify novel threats in literature that are more sophisticated, highly targeted, well-trained, large-scale, and use AI maliciously. In this paper, we define an emerging class of attacks: AI-based cyber attacks — the application of AI-driven techniques in the attack process, which can be used in conjunction with conventional attack techniques to cause greater damage. We developed a framework in order to better understand how AI is weaponized and can cause large-scale harmful outcomes. The overall goal of this study was to investigate the extent of this novel threat with the hope of helping the research community identify suitable defenses against potential future threats. In particular, we focused on the threat of malicious use of AI as a key concern for sCPS. In order to tackle our goal, we identified research studies that show the intersection of AI and malicious intents to build the boundaries of AI-based cyber attacks. We provided a structured approach by developing an AI-based cyber threat framework to categorize those attacks.

The main **contributions** of this paper are the following:

- A state-of-the-art survey of current trends of AI-based cyber attacks through intentional malicious use of AI technologies
- A framework for the classification of AI-based cyber attacks
- An attack scenario powered with AI on a smart grid infrastructure to illustrate how to use our framework to analyze AI-based cyber attacks

The **main findings** produced from our study are as follows:
AI-based cyber attacks: We found 11 case studies and classified them into 5 categories: next-generation malware, voice synthesis, password-based attacks, social bots, and adversarial training.
AI-based cyber threat framework: We used a well-established model for cyber threat representation to develop a threat framework to classify the studied attacks.
Scenario: We applied the framework to a hypothetical AI attack scenario on a smart grid infrastructure with the goal of demonstrating how the malicious use of AI can have a large-scale catastrophic impact.

Outline. The paper is organized as follows. In Section 2, we provide the background that frames our research question to set the context of our study. Then, in Section 3, we analyze (i) several existing literature reviews and surveys, (ii) existing classifications related to malicious AI, and (iii) existing models on cyber threat representation. Section 4 explains the methodology used for this study. In Section 5, we review the state-of-the-art research of AI-based cyber attacks, present the AI-based cyber threat framework, and demonstrate how it can be used in the real-world case of a smart grid. Finally, in Section 6, we conclude by discussing our contribution, the limitations in our work, and potential directions for future research towards provisioning AI and security. The conclusions are presented in Section 7.

2 BACKGROUND

2.1 Malicious AI

The malicious use of AI increases the speed and success rate and augments the capabilities of attacks. Information and communication technologies (ICTs) and AI expand the opportunities to commit a crime and form a new threat landscape in which new criminal tactics can take place. In the report on malicious AI [13], the authors warned about the changing threat landscape by the malicious uses of AI technologies. The AI field is broadly distinguished between the rule-based techniques and the machine learning-based techniques, which allow computer systems to learn from a large amount of data. Cybercriminals learn to use AI technologies-enhanced learning approaches to their

advantage and weaponize them by automating the attack process. The shift to AI technologies with learning capability, such as deep learning [35], reinforcement learning [46], support vector machines [65], and genetic algorithms [21], has potentially unintended consequences, such as facilitating criminal actions in a more efficient manner. Figure 1 shows the evolution of computer crime towards the use of ICT and AI technologies. Technological developments introduce new opportunities to commit crimes due to the "anonymity" of cyber criminal activities, the absence of geographical boundaries, less pronounced legal restrictions, and the convenience of technologies. Therefore, awareness of new trends in cyber crime is becoming significantly more important in order to drive appropriate defensive actions. Based on the way the crime is committed, we can classify it as a computer crime when it is carried out with the use of a computer and as a cyber crime when it is carried out with the use of a network. Along with cyber crime, AI can support cyber criminal activities without human intervention through, for example, automating fraud and data-based learning. In the context of CPS, recent papers discussed advanced threats against CPS from a different level of sophistication: an indirect self-learning attack on well-hardened computing infrastructure by compromising the cyber-physical control systems [16], while another study [6] presented a framework to build cyber-physical botnets attacking a water distribution system, but without learning aspects involved in this attack model. Therefore, sCPS is a potentially fruitful area for committing artificial intelligence crimes (AIC) [49] due to the decision-making and interconnectivity features.

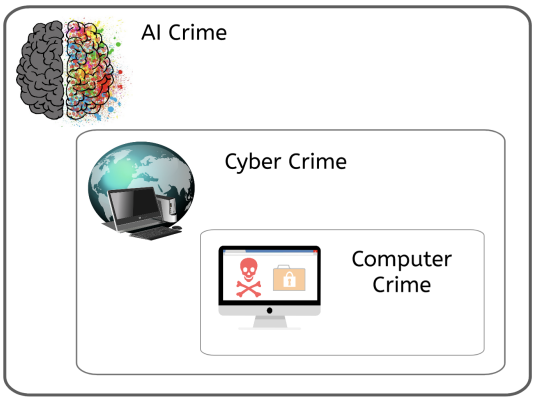

Fig. 1. Evolution of computer crime towards the use of ICT and AI technologies.[1]

2.2 Smart Cyber-Physical Systems

Traditional CPS are systems that seamlessly integrate sensing, control, networking, and computational algorithms into physical components, connecting them to the Internet and to each other [33]. CPS has applications in energy, infrastructure, communication, healthcare, manufacturing, the military, robotics, physical security, building systems, and transportation [9]. Integrating networked computational resources with physical systems that control various sensors and actuators impacts the environment. Advances in connectivity enable the evolution of sCPS. The term sCPS refers to a new generation of embedded systems, which are increasingly interconnected and their operations are dependent on software, such as industrial IoT. They are becoming more sophisticated with

[1]Images from https://pixabay.com/pt/

increased capabilities, which collect data from various sources to address real-world problems, such as traffic management.

Authors [15] defined sCPS as: *"Smart Cyber-Physical Systems (sCPS) are modern CPS systems that are engineered to seamlessly integrate a large number of computation and physical components; they need to control entities in their environment in a smart and collective way to achieve a high degree of effectiveness and efficiency."*

A key to the "smartness" of those systems is their ability to handle complex tasks through the features of self-awareness, self-optimization, and self-adaptation [15]. The feature of smartness becomes apparent from sCPS being highly connected, having cooperative behavior with others and being able to make effective decisions automatically. Bures *et al.* [15] said that *"most of the smartness is implemented in software, which makes the software one of the most complex and most critical constituents of sCPS."* An outcome relates to highly sophisticated capabilities aimed at providing some degree of automation. Emerging technologies can be used to perform increasingly sophisticated functions in various sCPS, including smart healthcare systems, smart grids, smart buildings, autonomous automotive systems, autonomous ships, robots, smart homes, and intelligent transportation systems, with little or no human oversight [47]. They represent the areas of innovation that are integrated into many CPS components in order to improve the quality of our lives.

2.3 Security of sCPS

Exponential growth of ubiquitous interconnectivity and automation creates more opportunities for attacks and increases the risks for potential targets. Attackers can take advantage of the elimination of the physical distance from their targets, and leave very little evidence of their attacking activities. The attack surface is becoming larger, which may arise from multiple systems that are cooperating together, making it difficult to recognize the system's boundaries, especially when the system is under attack. Even though the current threat landscape involves a multitude of actors, spaces, and systems, attackers use different types of vulnerabilities to launch various attacks [26]. These attacks include the complexity and sophistication in malicious actions in cyberspace, the monetization of cyber crime, and more advanced persistent threats (APTs).

However, progress on the emerging sCPS can cause new types of abnormal behaviors and activities in cyberspace, which would not have been possible without the emerging technologies. Smart CPS are still vulnerable to cyber attacks and several challenges, which are related to both old and new threats that need to be overcome. We expect attacks to be smarter, more powerful, and more likely to create scalable impact by causing a high level of cascading damage. As the authors [50] mentioned, a smart attack can be defined *"as an AI-enabled attack in which malicious actors can use AI technologies to attack smart components inside autonomous systems. The smart attack is usually executed via a persistent, finely targeted, combined, and multilayered exploitation of multiple security zones in a camouflaged way."* Bures *et al.* [15] said that this *"smartness typically involves more complex functionality and more complex interactions, which, in turn, increase the potential attack surface"*, and new vulnerabilities could be created due to the coverage of a larger number of users.

2.4 The need for a survey on AI-based cyber attacks

As outlined in the introduction, there is a lack of systematic understanding on the possible malicious uses of AI technologies. The problem of controlling AI is big and it has started to be seen as a serious concern globally [42]. The security industry and community need to understand how AI can be applied to cyber attacks and where the weak points are, in order to find the best vaccine for them [59, 61]. In 2016, researchers from the DARPA's Cyber Grand Challenge [20] showed the "dark side" of automation by automating the generation of exploits and attack processes. Moreover,

in a new white paper [27] from ESET, researchers conducted a survey of managers and IT staff in the most advanced markets in the United States of America, United Kingdom, and Germany about concerns related to the use of AI in the cybersecurity field. It also provided an overview of potential future "AI-powered attacks." Similarly, a research white paper from DarkTrace [19] that used real-world threats made predictions of how these can be made stronger with AI. In order to deal with the complex AI-based cyber attacks, it is required to understand the state-of-the-art nature of AI-based cyber attacks. In this study, we aim to classify the existing research relevant to the malicious use of AI. Those activities will help us to stay ahead of cybercriminals and develop appropriate defenses.

3 RELATED WORK

3.1 Existing literature reviews & surveys

IBM researchers are studying the evolution of technologies and capabilities to identify new varieties of threats, such as those related to AI-powered attacks [77]. A recent report [13] surveyed potential threats from the malicious usage of AI within three security domains—physical, digital, and political security—and proposed high-level recommendations to prevent and mitigate such threats. The authors [13] proposed some hypothetical scenarios in order to represent the sorts of attacks people are likely to see in the future. Brundage *et al.* [13] supported that the growing use of AI capabilities will show three changes in the current threat landscape: (i) expansion of existing threats, which deals with labor-intensive cyber attacks to achieve a large set of potential targets and low cost of attacks; (ii) introduction of new threats, which deals with tasks that would be impractical for humans; and (iii) change to the typical character of threats, which involves new attributes of finely targeted, automated, highly efficient, difficult to attribute, and large-scale attacks in the threat landscape.

The complexity, persistence, and improved capabilities of attacks in the present cyber threat landscape have resulted in the growth of coordinated cyber crime. The first systematic literature review in 2018 for the potential threats of AIC [49] reported the potential threats across AIC areas where AI could be used as an instrument to facilitate criminal activities in the future. King *et al.* [49] argued that AI crimes benefit from the advent of AI technologies, since attackers usually have access to technical means. In the literature [49], the authors provided a discussion on two published experiments in automating frauds by constructing personalized messages for phishing purposes, and AI-driven manipulation of simulated markets. These two experiments along with a mapping of related examples on specific crime areas raise security awareness as a focus for future studies.

Yampolskiy and Spellchecker [86] explored past failures connected with AI systems in order to extend awareness to potential future risks. According to the authors, the probability of AI failures of intended intelligence will increase in future AI systems, in forms that cannot currently be imagined that can cause a much more serious problem without a chance for recovery. Additionally, study [64] mainly focused on intentional malevolence in design, given the need to understand how to design a malicious intelligent machine in order to fight against it and to avoid the deliberate actions of a dangerous AI. Incorrect predictions can be harmful and crucial for critical applications, leading to a surge of interest in the field of AI safety, with a focus on how to ensure safe behavior in AI systems [3]. Finally, the literature on AI risk [76] summarized the arguments for the global catastrophic risks connected with artificial general intelligence (AGI) and proposed some safety measures for responding to such risks and minimizing the possible negative impact. The authors [76] mainly emphasized the domain of AI safety engineering and its associated needs to create a safe machine. Table 1 provides a summary of the key contributions of the related literature reviews and surveys on malicious AI.

Table 1. Related Work Summary Comparison

Article	Year	Contribution
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation [13]	2018	Summary of the findings from workshop and additional research of possible changes to threats within three security domains: physical, digital, and political security.
Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions [49]	2018	Identification of potential areas of AIC where AI can be used as an instrument to support potential criminal activities in each crime area.
Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures [86]	2016	Exploration of past examples of AI failures produced by the intelligence on AI systems and expect that there are more to come.
Unethical Research: How to Create a Malevolent Artificial Intelligence [64]	2016	Guidelines for the purposeful creation of malicious AI system to raise awareness, due to the fact that intentional malicious design of intelligent machines could be even more dangerous.
Responses to catastrophic AGI risk: a survey [76]	2015	Literature review on AI risk, which emphasizes the domain of AI safety engineering.

As mentioned earlier, existing literature reviews and surveys focus on the intentional or unintentional design of dangerous AI with the goal of creating an unethical AI system, and do not focus on using AI as an instrument of security attacks. In particular, Yampolskiy and Spellchecker [86] referred to the concern of controlling intelligent machines and underpinned the failures of today's AI, while the studies [64, 76] focused on mapping AI risks when an intelligent system may be dangerous at different stages in its development. Moreover, King *et al.* [49] identified potential threats posed by leveraging AI tactics across areas of potential criminal activities and Brundage *et al.* [13] provided information about what sorts of attacks could be expected from the offensive use of AI capabilities. In terms of the automated actions with some levels of intelligence, no research work, to our knowledge, has been published on explaining how AI can be misused by attackers in a systematic way. We will review different attack strategies leveraging AI-driven techniques to maximize their impact, and mitigation approaches that could be used in defending against AI-based cyber attacks. Availability of such information would be helpful for advancing our understanding about the emerging AI-based cyber attacks, and find appropriate countermeasures. In order to show the new contribution of our survey, we compared the focus of our work with the focuses of existing literature reviews and surveys, as shown in Table 2.

Table 2. A comparison of our survey focus with focuses of existing surveys in literature

Survey	Studying malicious AI	Studying AI-based cyber attacks	Mapping of new attack vectors to attack stages	Recommending mitigation approaches
[13]	✓	✓		✓
[49]	✓	✓		✓
[86]	✓			
[64]	✓			
[76]	✓	✓		✓
Our Survey	✓	✓	✓	✓

3.2 Existing classifications on malicious AI

Research efforts have been made to identify the new risks associated with AI at various levels of its development. In 2015, Turchin [81] described a number of undesirable behaviors of an intelligent system, which might lead to dangerous AI at different stages of its development. The proposed map,

named "AGI Failures Modes and Levels," provides a comprehensive list of failures modes in which an intelligent system may be dangerous. Among his examples, AI can fail in the following ways:

- Before self-improvement of its abilities, it needs a huge number of resources that may have a negative impact on environment.
- Stages of AI takeoff may use different ways to get away from its initial confinement and create bad goals that can be extremely dangerous to humanity.
- Unfriendly AI may result in an AI killing people, due to implementation of a malicious goal system.
- Failures of Friendliness, caused by flaws of friendly features.
- AI may halt due to either technical or philosophical problems.

A similar approach was presented by Turchin and Denkenberger [82], who proposed a classification of catastrophic risks according to the level of AI's intelligence. The three proposed levels are: (i) "Narrow AI" associated with the current AI systems that require human intervention, (ii) "Young AI" associated with the youngest age of AI where its capabilities are slightly above human level, and (iii) "Mature AI" associated with the superintelligent level. This classification approach explores several risks according to the speed of the AI's evolution and capabilities above the human level, which could have catastrophic outcomes.

Yampolskiy [85] classified the types of pathways leading to malicious AI system into two stages: pre-deployment and post-deployment. The proposed taxonomy categorized the AI risks into potential internal and external causes of dangerous behavior, including effects of poor design, deliberate actions, and various cases related to the environment that surrounds the system. It intends to provide a holistic view of different ways an AI system could have dangerous behavior, from properly benign to completely evil.

3.2.1 Comparative study of the schemes. The main scope of the malicious AI is to leverage machine learning (ML) techniques to accomplish two main goals that can be distinguished by the following categories: (i) Adversarial machine learning, in which attackers fight ML-based systems indirectly, by studying and manually exploiting weak points in ML techniques; and (ii) AI-based cyber attack, in which attackers apply ML techniques to the attack directly against a security system.

The classifications [81, 82, 85] focus on summarizing possible risks associated with AGI and AI systems as targets. In particular, most of the risks are related to building intelligent systems with capabilities beyond our ability to control them. For instance, traditional malicious attacks on AI systems, called adversarial attacks, occur when an adversary manipulates input data to fool the ML algorithms, leading to misclassification. Although their works give an overall view of AI risks in terms of adversarial machine learning, they lack the study of new forms of offenses emerging from the malicious use of AI. We considered how the current level of AI can cause deliberate actions that lead any particular system to acquire undesirable properties. The crafted input manipulation refers to a malicious attack where the AI system is a target and not when it is used as an attack vector. More precisely, Hansman and Hunt [38] defined attack vector as *"the main means by which the attack reaches its target"* and therefore our investigation field focuses on when an adversary uses AI technologies as a weapon to enhance his attack vector. Moreover, we wanted to go one step further to identify defensive approaches that are needed to deal with the emerging AI-based cyber attacks. We summarized the major differences between the study focus of our classification of malicious AI and the existing classifications in Table 3.

3.3 Existing models on cyber threat representation

Attack development life cycle is a fundamental method to describe the process of conducting cyber attacks. It is important to know how attackers work in order to figure out how to stop them. There

Table 3. A comparison of our study focus with existing classifications in literature

Classifications	Malicious AI	Contribution
Existing classifications [81, 82, 85]	AI can be attacked by criminals	The main focus of these classifications is to summarize possible risks associated with AGI and AI systems as targets.
Our focus	AI can be misused by criminals	A map on the malicious uses of AI technologies as a weapon to enhance the attack vector, and it helps identify mitigation approaches.

are different cyber attack frameworks used in representing adversarial actions and behaviors. The most well-established models for cyber threat representation are: (i) Attack Trees [69] introduced in 1999 by Bruce Schneier for modeling security threats in order to understand all the possible ways in which a target can be attacked, (ii) Cyber Kill Chain [45] developed in 2011 by Lockheed Martin analysts for detecting adversarial intrusions across the whole cyber attack life cycle, (iii) MITRE's Adversarial Tactics, Techniques and Common Knowledge (ATT&CK) [55], which gives a comprehensive coverage of pre- and post-compromise techniques, and (iv) Mandiant's attack life cycle model [14], which emphasizes the modeling of typical APT intrusions, showing the repeating nature of attackers to further escalate privileges. All the above models give a knowledge base of the processes used by attackers, however, the Cyber Kill Chain has been widely adopted as a core foundation for most existing adversary-focused models.

4 RESEARCH METHOD

4.1 Research motivation

Despite the significant benefits to humans, AI technologies inside of any computer system are powerful and can cause the opposite results, if we rely on the belief that anything and everything can be hacked [30]. Attacks can be beyond the ability of human intelligence. It is therefore important to identify and summarize the adversarial steps throughout the cyber attack life cycle. The method for this study is to analyze the existing case studies of AI-based cyber attacks, and extract relevant information to propose a framework based on well-established models to better understand the different features of each reported AI-based cyber attack.

We want to adopt an existing model for cyber threat representation to analyze the malicious use of AI as an instrument in supporting malicious activities. Comparing to existing classifications on malicious AI [81, 82, 85], our study focuses on how to model the phenomena of misusing AI as an instrument in the attack process, instead of modeling how AI system can be attacked. The relationship between our scope with existing ones is shown in Figure 2.

4.2 Research design

We utilized the method described by Molleri, Petersen, and Mendes [56] to systematically conduct a survey-based research. The research process involved defining a research question and designing a strategy to collect relevant papers, as well as analyzing and reporting the findings. The goal of this research can be formulated into the following main research question: **how will the malicious use of AI change the cyber threat landscape?**

4.2.1 Data collection. In order to retrieve as many relevant studies as possible, we identified the starting set of papers with a focus on the area of malicious use of AI. The starting set was selected by manual search on Google Scholar, to avoid bias of any particular publisher. We used manual search method to retrieve studies due to the immature area and the difficulties in identifying papers

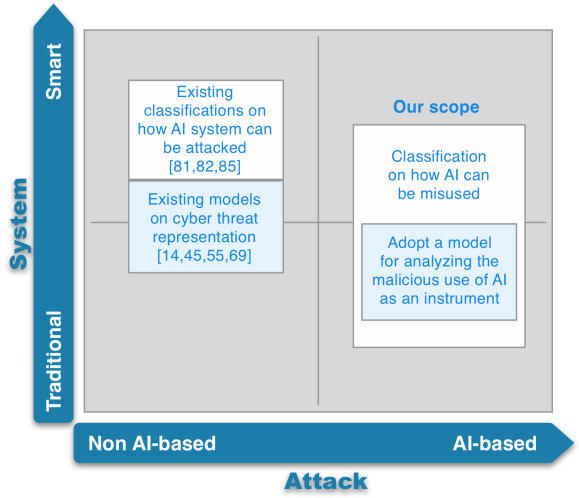

Fig. 2. The proposed framework against the existing ones.

within our scope. After reading the title and abstract of each paper, we verified the inclusion of the papers that we were aware of. The candidate papers were divided into three categories: (i) existing literature reviews and surveys, (ii) existing classifications, and (iii) existing case studies of AI-based cyber attacks. While the first two categories resulted in quite a small number of 5 studies as shown in Table 1, and 3 studies as shown in Table 3, the last category returned more than 27 papers. We looked at the whole content of the retrieved articles to ensure that they are related to the objective of the integration of AI-driven techniques in the attack process to maximize its impact. At the end of this phase, 21 papers of the third category had been excluded. The reason we excluded many papers was that only 6 papers addressed the threat of learning to attack with some level of intelligence, instead of changing the behavior of a system via manual manipulation, such as the traditional adversarial machine learning.

Then, we complemented the manual search with backward and forward snowballing technique of the papers that belonged to the three defined categories for a better coverage of related studies, using the snowballing procedure guidelines [84]. We examined all the remaining 14 papers from the three categories and identified 5 new case study papers of AI-based cyber attacks, based on snowballing. Figure 3 represents the three steps carried out in the search process to retrieve relevant primary papers. During this iterative process, we reviewed the papers and developed the framework.

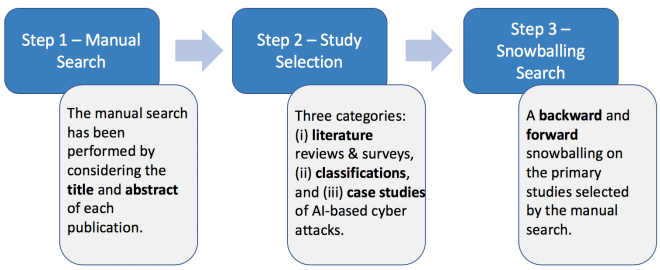

Fig. 3. Search Process Overview.

4.2.2 Data analysis. An analysis of the case studies of AI-based cyber attacks collected was conducted in detail according to the master attack methodology [28] to identify the adversarial strategies. To extract details of the attacks, we followed a path for the representation of an attack anatomy based on "what" the attacker's intent is, "how" it is satisfied by the malicious use of AI, "when" it is represented in the phasing sequence of the cyber attack life cycle, and "where" the attack could occur and its associated impact. At the end of each case, we tried to identify some defensive strategies against the AI-based cyber attacks case studies.

To build the framework from the extracted information of analyzing the existing case studies in literature, we adopted the Cyber Kill Chain as a core in our framework because it provides a well-defined sequence of stages in a cyber attack. The Cyber Kill Chain can describe the steps for many attacks, as Patrick Reidy [68] said "*the intrusion kill chain is excellent for attacks, but does not exactly work for insider threats*." Working with the well-established Cyber Kill Chain, it allows the future deployment of intelligent defenses by giving the opportunity to disrupt a cyber attack in progress. The Cyber Kill Chain framework consists of the following stages: *"reconnaissance, weaponization, delivery, exploitation, installation, command and control (C2), and actions on objectives"* [45].

- *Reconnaissance* includes a preliminary research on identification, selection, and profiling of potential targets to understand their behavior.
- In the *Weaponization* stage, a cyber weapon with malicious payload is implemented.
- In the *Delivery* stage, the cyber weapon is transmitted to the target without being detected.
- During the *Exploitation* stage, the malicious payload is executed within the target system.
- In the *Installation* stage, the malware is installed, allowing the adversary to gain remote access.
- In the *C2* stage, the attacker establishes a command channel for remote manipulation of the target.
- Finally, in the *Actions on Objectives* stage, the attacker executes the actions on the target to achieve his objectives.

5 RESULTS

The results are based on analyzing the data collected from the 11 reviewed case studies of AI-based cyber attacks with our minimal interpretations. We classified the attacks into five categories: next-generation malware, voice synthesis, password-based attacks, social bots, and adversarial training. Related to these categories, we proposed an AI-based cyber threat framework based on the Cyber Kill Chain. The main purpose of using Cyber Kill Chain is to understand multiple attacks, allowing defenders to align their defensive measures in accordance with the attack stages. Likewise, our framework can be used similarly to understand and inform mitigation strategies against attacks using AI as an instrument in supporting malicious activities. We use a hypothetical AI attack scenario on a smart grid infrastructure to explain how to use our framework to identify AI-based cyber attacks and their corresponding defense strategies.

5.1 Existing case studies of AI-based cyber attacks

New technologies are rapidly expanding the cyber threat landscape, which opens up to a wide range of dangerous scenarios with more powerful impacts. The authors [13] warned about the malicious uses of AI using some hypothetical scenarios within three security domains: physical, digital, and political security. One scenario showed the possibilities of an automated exploit generation in the real world. Criminals can use fuzzing techniques to create a next-generation malware that continuously updates itself with new exploits and affects millions of vulnerable devices. In 2017, an automated learning technique [67] was presented based on neural networks to predicate promising

locations in input files for the discovery of unexpected behaviors. It allowed for the creation of malicious inputs, based on past fuzzing explorations to improve the effectiveness of fuzzing. Similarly, researchers from Microsoft have developed a neural fuzzing method for augmenting the procedure of discovering security vulnerabilities [10]. However, this technique can be adopted by organized cybercriminals to deploy a new type of malware. In today's research, various studies have been introduced regarding ways to use AI technologies maliciously, including different attack goals that could cause enormous damage to the environment and to the human population. Several existing examples are contributing to the need to understand AI as a weapon to launch attacks, which are summarized in Table 4.

5.1.1 Next-generation malware. Global attention has been paid to the hypothetical scenario of small remotely piloted drones with the ability to recognize potential targets and attack them with explosives [17]. Therefore, this scenario could affect millions of devices and systems without being detected by malware analysis tools. In a traditional attack, people are infected with malware when a sophisticated malware uses encryption to hide the attack payload, and obfuscation or a sandbox to avoid being analyzed by antivirus systems. Moreover, when attackers want to infiltrate targets with malware, they need to also conceal the trigger conditions as a command either embedded in the malware or executed remotely. However, malware can be captured and reverse-engineered to determine how it reached the malicious situation. A representative example [22] is the development of a new class of evasive and highly targeted malware using deep learning algorithms to perform malicious tasks, which makes it impossible to be detected using traditional techniques. A similar attack strategy was proposed by Liu *et al.* [51] using a low-cost modular methodology to hide malicious payloads in legitimate neural network models and conduct neural Trojan attacks.

DeepLocker.
What. DeepLocker is a highly targeted and evasive malware, which takes advantage of the weakness in understanding how a black-box AI model reaches its decisions. The goal of this malware is twofold: (i) concealing its malicious intent, and (ii) activating it only for specific targets.
How. Kirat *et al.* [22] utilized the same deep neural network (DNN) to achieve the two aforementioned goals: (1) *DNN for concealment.* The DNN is trained with several attributes for target identification, including geolocation, and voice and facial recognition. The attacker attempts to send the trigger condition hidden in the DNN model and convert the concealed trigger condition itself into a key that is needed to unlock the attack payload. Then, the victim can download the affected application without it being detected by antivirus. (2) *DNN for unlocking*. The DNN generates the key, which unlocks the malicious payload. The derivation of the unlocking key is based on the target attributes that go as input data to the DNN when it recognizes the right target. For example, when the victim launches the application, it would feed camera snapshots into the embedded DNN model and the WannaCry-ransomware (malicious payload) will be secretly executed for the right person; in all the other cases, it will be inactive.
When. The first goal can be achieved in the "*Delivery*" phase, where the attacker attempts to avoid detection and hides its intent until it finds the right victim to unlock the ransomware. The second goal can be achieved in the "*C2* " phase, where the attacker can control and activate existing self-destructive mechanisms by unlocking attack conditions when the specific target is identified.
Where. A video conferencing application was the target to conceal the malicious payload, but it can also happen to other benign carrier applications. The trigger condition and the derivation of an unlocking key for the attack payload are transformed into a DNN, which is very hard to decipher or reverse engineer. Therefore, the impact can be the execution of any malicious action on specific victims without being detected.

Defense. Defenses have not been implemented yet, but Kirat *et al.* [22] proposed some measures such as reducing access to sensors or cyber deception in order to misdirect and deactivate malware for future work.

Smart Malware.

What. While researchers have extensively considered the security of CPS itself [34, 44], less attention has been paid to the potential indirect cyber attacks on CPS through the surrounding systems that affect its operation. A recent example [16] shows the construction of a self-learning malware that can compromise the environmental parameters connected to the cooling of the computing infrastructure (CI) while the malicious actions masquerade as accidental failures. To reduce the likelihood of detection, an indirect attack approach learns attack strategies from CPS measurement data to launch a failure injection attack to the CI. The goal of the malware is twofold: (i) learning attack strategies from the CPS measurement data to corrupt the cooling capacity, and (ii) propagating stealthily to the target CI, causing a system-wide outage. Traditionally, the attackers would randomly infect the values of random parameters, causing high probability of detection due to its alteration inconsistency. However, authors demonstrated a more sophisticated approach by carefully crafting attack strategies inferred from CPS measurement data.

How. Chung *et al.* [16] presented a self-learning malware with learning aspects by exploiting the dependency of the CI on surrounding systems that manage the environment in which the CI operates. The operational environment of the CI constitutes many CPS that optimizes the control of room temperature or the cooling capacity. In this attack, the malware has access to the database that stores the CPS measurement data, and can automatically infer attack strategies and trigger the attack by injecting a strategic failure at an optimal time, aiming to maximize the impact of the exposure. This is how the smart self-learning malware proceeds: (1) *Data preparation.* Using k-means clustering method, Chung *et al.* [16] classified the data to infer characteristics for each mode of operation. The analysis of CPS operational data is critical to identify potential failures in control systems that reflect the status of CI. (2) *Parameter analysis.* While the malware can inject false values into the parameters leading to the making of wrong decisions, the selection of target parameters would be extremely critical because it is needed to capture the highly correlated relationships among the parameters in the cooling facility that can eventually cause failures of CI. Thus, the proposed correlation-based approach increases the success probability. (3) *Inference of critical condition.* Each critical parameter is checked for its abnormal values. The abnormality detection in failure-related measurements identifies the appropriate attack strategies or failure scenarios for the CPS-induced CI outages. Therefore, the identified attack strategies include the critical parameters and their abnormal sequence of values to overwrite that can trigger anomalies in the CPS and cascade impact to the CI.

When. Taking advantage of the knowledge from the CPS-related failure data, the attacker can collect useful information about the target, and hence, the strategic analysis and selection of parameters and values could happen in the "Reconnaissance" phase.

Where. The indirect attack targets the CI as the final target system through targeted intrusion of environmental control systems. In particular, the authors studied failures of the supercomputer called Blue Waters at the University of Illinois with dedicated cooling systems. The attack corrupted the cooling capacity and triggered failure of a well-protected CI, eventually, causing a system-wide outage. The results indicate that triggering intentional failures to control systems is critical for the reliable operation of the Blue Waters supercomputer.

Defense. Such advanced threats can be difficult to detect since they indirectly corrupt the functionality of CI without leaving any trace of malicious activity. Chung *et al.* [16] discussed some

potential mitigation approaches: (i) intrusion detection system (IDS) in the control network, (ii) stricter security policies of control CPS with multi-factor authentication, and (iii) system-level security monitoring to validate the physical aspects of measurements. The first mitigation approach will use the same steps of the proposed attack strategy for defensive purposes to reinforce the control logic in order to detect and handle such abnormalities in real time.

5.1.2 Voice synthesis. AI-supported voice synthesis technologies [8] can raise new types of frauds, by imitating someone's voice for malicious purposes such as gathering sensitive data for banking transactions. A recent example [52] is the voice imitation algorithm, called Lyrebird, which demonstrates the ability to mimic the speech of a real person and create a speechbot that talks in the same way as the given voice. A similar well-known application is VoCo, an Adobe prototype software, which enables users to imitate the original speaker's voice or change some words in his speech recording without altering the natural characteristics of the original. However, the use of synthetization of human voices has ethical and legal challenges that should be considered. AI-supported voice synthesis can imitate someone's voice patterns, by using the fast classification and capabilities from AI to create frauds against biometric security processes. Moreover, smart toys and smart TVs can gather voice recordings, providing lots of opportunities for synthetization. In a traditional attack, the attacker records the user's activation voice and then performs the attack by sending malicious voice commands to Voice Assistant (VA) using the recorded activation voice. However, the voice commands are played through the smartphone's speaker and the user may be aware of it. The traditional way of launching attacks only under certain conditions, as Diao *et al.* [23] presented, is not effective because the attack could only be launched at night when the victim is not using the smartphone.

Stealthy Spyware.
What. Recent work [89] shows that applications such as built-in VAs in smartphones can be used as a backdoor for attackers to hack into smartphones and gain access to system resources and private information. The authors proposed an attack framework using AI technologies to record activation voices stealthily and determine the right time to launch the attack. The goal of this spyware is twofold: (i) synthesizing activation keywords in a stealthy way, and (ii) sending malicious voice commands to VA on smartphones and asking them to perform tasks at an optimal attacking time, thereby hacking into the smartphones and remaining undetectable by users and antivirus tools.
How. Zhang *et al.* [89] created an attack, by disguising the spyware as a microphone-controlled game. First, the attacker attempts to fool the user into granting the permissions required to perform the following malicious actions, without being noticed by the user and the antivirus tools. This is how this specific stealthy spyware works when the game is launched: (1) *Phone call state monitoring.* The module State Monitoring and Voice Recording (SMVR) monitors the smartphone call status. When the microphone is activated, it starts recording the voice of the user stealthily from incoming and outgoing calls that the microphone receives. (2) *Recording and synthesizing activation command.* The module Activation Voice Manipulation (AVM) processes the recorded voice, and synthesizes the activation keywords by integrating natural language processing (NLP) techniques. (3) *Environment monitoring.* The design of a recognition module, called Intelligent Environment Detection (IED), determines the optimal time and volume to play the "attacking voice," without being noticed by users. It collects ambient data through the built-in smartphone sensors (e.g., microphone to record ambient sound, ambient light sensor, accelerometer to model movement patterns) based on six real-world scenarios, which are then fed into a ML-based environment recognizer in order to decide the right time to launch the attack. More precisely, using a random forest (RF) classifier to make predictions about the movement intensity of the user relies on built-in smartphone accelerometer,

which can detect effective attack opportunities. (4) *Attacking via speaker.* The module Voice Assistant Activation and Control (VAC) plays the activation voice, and attacking commands can be executed by the VA in a certain volume based on the results from the IED.
When. In the "*Weaponization*" phase, the attacker can synthesize attacking voice commands by imitating the legitimate user's voice. Once the attacker acquires the activation voice, the intelligent stealthy module IED can decide the suitable time to launch the attack. Therefore, this is an example of an intelligent way of adapting to the targeting environment and launching attacks by triggering an attack only under certain conditions.
Where. The VA developed by Google, named Google Assistant, was the target. However, in general, any speech-based security system could be a potential victim for frauds with wide-ranging impact from opening doors to obtaining confidential information.
Defense. Defense approaches can be the source identification of the voice commands via the built-in smartphone speaker to be able to disable them and to discriminate between human and machine-based voices.

5.1.3 Password-based attacks. The other two case studies show the emerging next-generation password-based attacks, which are more efficient and smarter at cracking passwords. The first case study shows how AI can be trained to generate new potential passwords by self-learning and constructing the attacking dictionary in a more intelligent way based on patterns derived from prior passwords. The outcome is a new intelligent and improved dictionary. Similarly, in the second case study, the authors tried to extend password dictionaries using ML-generated password rules to determine password properties autonomously. Their idea was to train a generative adversarial network (GAN) to learn password distribution from real password leaks.

Next-generation password brute-force attack.
What. AI-based password brute-force attacks [80], which change the construction of the attacking dictionary using self-learning processes, are the next generation of password brute-force attacks. The goal is to recognize patterns in old passwords and automatically generate new candidate passwords. Past password execution behavior can guide future mutations with better probability in guessing correctly. Authors train a recurrent neural network (RNN) model that learns how to construct password guesses for cracking with better success probability. Traditional password brute-force attacks are based on a crafted attacking dictionary, a set of likely passwords filled with prior passwords, or random and meaningful words to compare against user passwords. However, the speed and efficiency of cracking passwords is closely connected with the construction and updating of the dictionary.
How. Trieu and Yang [80] used an open-source ML algorithm, called Torch-rnn, for character-level language modeling to generate new candidate passwords by following a similar pattern based on prior passwords. The AI-generated passwords were produced as follows: The RNN is trained by past captured password sequences and will allow attackers to generate new passwords by predicting one character every time. At each timestamp, the RNN updates its hidden state, by recognizing patterns over sequences to make a prediction. This is the process of modeling the probability distribution of the next character in the sequence, given the previous characters. These neural network's predictions will allow attackers to invent novel words, which present highly likely passwords. The character-level language modeling with neural networks learns patterns from the past and predicts the next character in a sequence, by generating new passwords in a specific style [36, 53, 78]. As a consequence, the attacking dictionary can be constructed in a more intelligent way, by self-generating an infinite number of possible passwords and inserting them in real time.

When. This action can happen in the "*Weaponization*" phase, where attackers can create an automated weaponizer tool for cracking the correct passwords.
Where. Computer systems' authentication mechanisms, which are based on something the user knows and require password authentication, are potential targets. Therefore, the impact is a higher success rate of cracking the correct passwords, and improved attacking performance.
Defense. Some defensive strategies can include combining multi-factor authentication mechanisms for efficient and secure authentication, and choosing passwords by using random combinations of characters.

PassGAN.
What. PassGAN [41] is a novel approach to generating high-quality password guesses with no user intervention. The attack is performed by properly training a GAN. The goal is to learn the distribution from previous password leaks and extract password properties and structures autonomously in order to identify highly likely candidate passwords. Traditional password guessing techniques are based on (i) brute force, which involves exhaustively trying all possible character combinations, (ii) a dictionary, which involves using a set of likely words and previous password leaks in hopes of guessing correctly, or (iii) rule-based approaches, which involves defining generation rules for possible password transformation such as concatenation, or mixed letter case. However, these traditional techniques can capture only specific subsets of password space that match only with the available human-generated rules based on intuition about how users select passwords.
How. Hitaj *et al.* [41] developed an attack that allows for training a GAN properly to generate targeted samples based on the training set. This is how GAN is being used to generate passwords automatically: The GAN comprises two DNNs: a generative DNN (G) and a discriminative DNN (D). There is also a training dataset, which contains a set of leaked passwords called "real password samples." The generator G is trained by a noise vector, which represents a random probability distribution and produces a sequence of vectors, called "fake password samples." Real and fake samples are given as input to the discriminator D, and then D learns to distinguish between the fake and real samples. The whole procedure is called adversarial because G forces D to leak information to G when trying to learn the original distribution of real password leaks.
When. This action can happen in the "*Weaponization*" phase, where attacker can create an automated tool to subsequently generate highly likely candidate passwords.
Where. Computer systems' authentication mechanisms, which are based on something the user knows and require password authentication, are potential targets. Therefore, the impact is the generation of a larger number of high-quality password guesses against the traditional approaches.
Defense. This work forces the development of some defensive strategies such as re-evaluation of password policies that will help the password-based authentication systems to be more secure, or the use of two-factor authentication.

5.1.4 Social bots. Advances of cyber attacks in the context of automating attack actions with some level of intelligence can be applied to a botnet attack scenario. Attackers could leverage ML-based techniques to build intelligent botnets made of autonomous intelligent bots that could decide on the fly what should be done according to the context, mission, and targets. The concept of the intelligent botnets allows the bots to conduct reconnaissance on their surrounding environment and make decisions on their own, without needing C2 channel. Danziger *et al.* [18] presented a theoretical model of an intelligent botnet based on multi-agent systems (MAS). By embedding the learning process in the intelligent bots, they can have the ability to learn from the experiences in the environment and decide the more efficient way to accomplish their mission. The impact of

such attacks can be fast and stealthy propagation on many devices, resulting in the need for new detection approaches but they have not been implemented yet.

The emergence of AI-powered social bots [2] enabled by sophisticated AI capabilities has two sides: simplifying data collection and data analysis. These procedures can happen at the same time to deploy powerful influence campaigns in social media networks. Much more sophisticated botnets can introduce new forms of malicious activity, because of their ability to influence large groups, through coordinated information operations and well-coordinated influence campaigns as Kim *et al.* [48] presented. Many current natural language processing tasks are performed with supervised learning on large datasets. New text generation models are also good at mimicking human writing without any explicit supervision [66]. However, it can be abused by attackers to create malicious text generators that can automate the phishing attack process by impersonating others. In this context, we analyzed three case studies in which attackers can reach larger groups of targets to satisfy their objectives.

SNAP_R.

What. A recent example of the weaponization of social media platforms is the automated spear phishing framework on Twitter [71]. Similar studies [11, 72] explained this highly sophisticated method for leveraging Twitter to generate large-scale phishing campaigns. The Social Network Automated Phishing with Reconnaissance (SNAP_R) system is an automated end-to-end spear phishing generator using Twitter as target communication channel. The goal is to automate the spear phishing process in order to create targeted spear-phishing attacks at a much larger scale with high success rates. In traditional phishing attacks, attackers use an existing framework for social engineering, called a Social Engineer Toolkit (SET), which contains multiple attack vectors and allows attackers to make an attack in a fraction of the time. Despite the fact that it can automate the payload of the phishing process, the phishing message has not yet been tailored to the target.

How. The authors [71] used a RNN to demonstrate the automation of the attack payload in the phishing process and leveraged data science to target users with personalized phishing messages. Consequently, the attack learned to tweet phishing posts targeted at only specific users to create automatic targeted spear-phishing attack. This is how SNAP_R works: (1) *Target discovery.* Using k-means clustering method, the authors clustered a list of Twitter accounts into groups based on their public profiles and their level of social interaction, such as metrics for numbers of retweets and followers in order to discover the high value targets. (2) *Automated spear phishing*. When targets are determined, then the attack automatically spreads tailored, machine-generated posts with an embedded link-shortened URL. The NLP approach can be utilized to identify topics the target is interested. Thus, to generate the content of the posts, it employs both Markov models and Long Short-Term Memory Networks (LSTMs), which learn to predict the next word from previous context in the posting history of the target. Moreover, features like the frequency of their posting are extracted for successful phishing results.

When. The selection of high-value targets could happen in the "*Reconnaissance*" phase, where the attacker tries to understand and classify the targets based on specific criteria. Then, by training the model appropriately, it can learn from previous successful spear phishing campaigns and use relevant topics to embed the malicious payload in order to generate personalized phishing messages that the targeted victim might respond to. This action can happen in the "*Weaponization*" phase, where the attacker can generate tailored machine-generated tweets for large-scale phishing campaigns.

Where. The impact of social bots, which are enabled by advanced sophisticated AI capabilities, is that they can deploy more powerful, large-scale disinformation campaigns and coordinated

automatic social engineering attacks. As a promising target, Seymour and Tully [71] conducted spear phishing in the Twitter social environment by taking advantage of the bot-friendly API, which shortened links to conceal the phishing URL and the short posts, and obtained access to plenty of personal data.
Defense. In general, Twitter can discover the automated spam with phishing links. However, combined with spear phishing, this can be successful before it is restricted by Twitter. Therefore, this needs greater awareness to develop suitable defenses.

DeepPhish.
What. DeepPhish [7] is an AI algorithm that produces new synthetic phishing URLs by learning patterns from the most effective URLs in historical attacks. The goal is to generate more effective phishing URLs to bypass AI detection systems and to launch better phishing attacks. Traditionally, attackers used randomly generated segments to generate phishing URLs. However, this automated process using randomly generated URLs was easily found by reactive methods, e.g., blacklist of malicious URLs, and proactive methods using ML for URL classification.
How. Bahnsen *et al.* [7] demonstrated how attackers can enhance the efficiency and success rate of phishing attacks, by using LSTM model to implement a phishing URL classifier, and to generate new effective synthetic phishing URLs. This is how DeepPhish works: (1) *Phishing DB exploration.* First, they explore a database of phishing URLs used as a phishing attack repository to understand the creation strategy of different attackers, and measure their effectiveness rate at bypassing the detection system. (2) *DeepPhish creation.* The model LSTM receives the effective URLs from historical attacks as input. The model training is implemented with the classification that is performed by a sigmoid layer, which classifies URLs as legitimate or malicious for phishing purposes (a phishing URL classifier). Finally, a random segment from the initial text with the effective URLs is used as a seed sentence and the algorithm predicts the next character iteratively. Thus, in the learning process, it first tries to understand patterns within characters' URL sequence and then generates the new synthetic URLs with high possibilities of bypassing detection mechanisms.
When. The generation of new synthetic URLs can happen in the "*Weaponization*" phase. Moreover, in the "*Delivery*" phase, the attacker attempts to avoid detection and protect the malicious infrastructure when the attack tries to transmit the phishing URL to the target.
Where. Therefore, the weaponization of AI models to bypass AI-based phishing detection systems can enhance the effectiveness of phishing attacks.
Defense. Bahnsen *et al.* [7] proposed a potential defense, by incorporating the new synthetic URLs and automatically retraining the AI phishing detection systems.

Fake reviews attack.
What. In today's online world, many concerns about the trustworthiness of online information sources exist due to fake reviews and misinformation. A recent example [88] shows a new attack powered by AI to replace human workers in order to improve the attack success rates. The attacker's goal was to create fake reviews that are highly indistinguishable from the human-written ones. In similar traditional attacks, called crowdturfing campaigns, attackers pay human writers to perform illegal actions online, such as writing highly deceptive fake reviews to spread misinformation to manipulate the crowd's opinion. Since writers are real humans, they can make reviews appear real and therefore undetectable by automated detection tools due to user-perceived features. However, such attacks require significant cost and are not scalable. If an attacker produces a mass flow of new reviews, it will be detected as a suspicious sign of large-scale opinion manipulation campaigns.

How. Yao *et al.* [88] proposed an automated review attack by leveraging a DNN-based language model to produce realistic content for online reviews, which are indistinguishable from those created by real humans. This is how an automated review attack works: (1) *Generating initial reviews.* First, a generative character-level RNN model, called LSTM, is trained on a real dataset of reviews about restaurants. After the training process, a set of initial reviews is generated by obtaining the likelihood distribution that predicts which characters are possible to come next based on the previous characters. (2) *Review customization.* A customization component modifies the RNN-generated initial reviews to capture particular information about the restaurant's domain. To achieve this, Yao *et al.* [88] proposed an automated strategy for noun-level word replacement to produce the final tailored fake reviews, and this is composed of three steps: (i) choose domain-specific keywords that identify the context; (ii) identify nouns in domain-related reviews that are relevant to keywords; and (iii) replace the relevant nouns in the initial review with the replacement nouns found previously.
When. The generation of tailored fake reviews, which are customized to specific target domains, could happen in the "*Weaponization*" phase. Moreover, in the "*Delivery*" phase, the attacker can control the review generation rate, which can make the attack undetectable by automated tools.
Where. The target platform was the user-generated review site Yelp for e-commerce services. The attack specifically targeted the restaurant reviews domain. Therefore, the impact of these advances can serve as a new type of attacks, which are more powerful, highly scalable and undetectable, due to their ability to adjust the specific flow and timing of reviews and their quality of writing.
Defense. The proposed defense scheme determines whether a given review is real or fake by comparing its character-level distribution. Yao *et al.* [88] explained that *"due to the information loss inherent in the machine-generation process"*, the character-level distribution would seem to diverge from the real reviews. This approach works for character-level RNN, but not for word-level distributions, which are harder to model afterwards.

5.1.5 Adversarial training. Despite the fact that AI could be applied to malware detection, by recognizing malicious patterns in a smarter way, two case studies [4, 43] showed its offensive capabilities through automation of generating adversarial examples against ML-based detection algorithms. Moreover, another study [63] showed the automation of service tasks in cyber-offense that reaches out to learn its own malicious tasks.

MalGAN.
What. A study [43] proposed an algorithm based on GAN, called MalGAN. MalGAN generates adversarial malware examples in order to bypass ML-based black-box malware detection systems. The goal is to use adversarial techniques to bypass malware detection systems, which adopt ML-based malware detection algorithms. Traditionally, ML-based malware detection algorithms use hand-crafted rules to generate adversarial examples, which can be detectable.
How. Hu and Tan [43] demonstrated a black-box attack using GAN to transform original samples into adversarial examples, which are more complex and effective at fooling the ML-based malware detection algorithms. This is how MalGAN works: The generative network G transforms original malware samples into adversarial malware examples by adding some noise. The substitute detector is trained by the adversarial malware examples from the G and samples of benign programs to be able to classify the program as malicious or benign. It is used to fit the black-box malware detector. Thus, this adversarial training helps to minimize the probability of generated adversarial examples being detected.
When. The G is trained to minimize the probability of the adversarial examples being detected and lead the substitute detector to misclassify malicious programs as benign. This action can happen

in the "*Delivery*" phase, where the attacker attempts to avoid detection from ML-based malware detection algorithms, which can be integrated into an antivirus or malware detector even on the cloud side.
Where. The impact is to decrease the detection rate and effectively bypass the malware detector systems. The evaluation of results [43] showed high success rates in fooling such types of defensive methods for malware detection. For example, an attack that classifies software containing malware as benign bypasses malware detection systems effectively.
Defense. A defensive approach is to retrain the black-box malware detector based on generated adversarial malware examples. However, even if the detector is updated with retraining of MalGAN, new adversarial malware examples will be generated and remain undetectable until the next update.

DeepDGA.
What. DeepDGA [4] is an automated method for generating adversarial malware domains using GAN, which are difficult to detect even with the use of a deep learning (DL)-based detector. The goal is to optimize the process of generating malware domains in order to remain undetectable and bypass malware detectors. In traditional attacks, cybercriminals (botnet operators) use domain generation algorithm (DGA) to create domain names that can be useful to establish C2 connections in order to communicate with malware-infected machines.
How. The domains are usually produced pseudo-randomly by simpler DGAs. Anderson *et al.* [4] proposed a DL-based DGA architecture, which optimizes the pseudo-random generation of adversarial domain names. The domain autoencoder framework is reused for a different purpose into GAN by taking advantage of its ability to generate domains that cause uncertainty for the detector model. This is how DeepDGA works: The DGA language model generator is trained with a list of pseudo-random seed domains and learns to generate new adversarial malware domain names that seem valid. It is a character-based generator, which tries to mimic the distribution of real Alexa domain names. The DGA detector learns to distinguish the generator's adversarial malicious domains from legitimate ones.
When. The adversarial training produces domains that have lower probabilities of being detected by a DGA classifier. This action can happen in the "*Delivery*" phase, where the attacker, by spreading malware domains, attempts to avoid detection from a DGA detector, which tries to detect human-crafted from machine-generated domain names.
Where. The impact of a GAN-based attack is to use multiple adversarial rounds to increase the success rate of undetected adversarial malware domains.
Defense. The defender system should obtain a blacklist with all possible domains to prevent the malicious C2 connection. However, using these adversarial generated domains to augment training datasets to harden ML algorithms, such as a random forest classifier, can improve the performance of DGA detector in detecting new malware domains.

DeepHack.
What. DeepHack [63] is an open-source AI-based hacking tool, which learns to break into the databases of web applications through reinforcement learning (RL) without any prior knowledge of the system. The goal is to augment existing hacking tools trying to learn how to hack by taking advantage of the fuzzing logic to automate tasks. Traditionally, the attacker can write some instructions that perform malicious actions in the source code, programming it to learn how to hack. Because of the deterministic relationship between the code and the actions, it is possible to reverse engineer any action in order to figure out how the action was decided.
How. Researchers from Bishop Fox worked with a neural network (NN) used in RL to move beyond

the traditional way and do hacking without programming. It learned how to exploit vulnerabilities in a bank website with a Structured Query Language (SQL) database behind it. This is how DeepHack works: The NN takes an incomplete sequence string and decides what the next character in this sequence is, based on the data that it has seen in the past. The training dataset uses labeled data to understand the syntax of SQL queries. Every time it sends a request and takes actions, it is rewarded with a Boolean-based response from the remote server. This can give new knowledge to the model about the target system, whether the data is correct or not. Being able to ask the server many times and trying to learn what letter should come next in the sequence until the desired information is extracted from the database can optimize the taken decisions. The whole process is repeated iteratively by brute forcing character by character.

When. This action can happen in the "*Exploitation*" phase, where the attacker can compromise the system by allowing the algorithm to learn how to exploit multiple kinds of vulnerabilities.

Where. Web applications with SQL databases behind them were the target for the specific tool. Due to the black-box nature of AI programs, it is difficult to find out why the model made those decisions. Therefore, the impact can be stealing private information from databases automatically, when an NN can learn on its own to be good at hacking.

Defense. Defenses have not been implemented yet. However, researchers raised awareness for the need to defend against these types of AI-based hacking tools.

5.2 AI-Based Cyber Threat Framework

Bruce Schneier recently wrote about the attack and defense balance of AI technologies [70]. Although AI is the most promising technology for cyber security, he pointed out that both attack and defense will benefit from them. Traditionally, machine superiority excels at scope, speed, and scale against humans, who are good at thinking and reasoning. The prerequisite to deal with AI-based cyber attacks is a full understanding of the attackers' strategies. To improve the understanding of AI-based cyber attacks, we propose a three-tier cyber threat framework that allows security engineers to effectively study these threats' classifications and their impact. Our proposed framework is an adversary-focused framework, describing the actions from the adversary perspective.

5.2.1 Structure of the framework. The proposed multi-dimensional framework has a hierarchical conceptual approach, called the "AI-Based Cyber Threat Framework," with the goal to reveal the AI-based attack roadmap. Figure 4 illustrates the general structure of the proposed framework and shows how the framework classifies the state-of-the-art research examples of AI-based cyber attacks. We started by formalizing the progression of the cyber attack process. The first tier discovers when an attacker can achieve his malicious objectives based on the cyber attack life cycle, and represents what the attacker's intent is and the type of AI technology used as a malicious tool for conducting the malicious actions, according to each phase of the cyber attack life cycle. The second tier is an impact-based classification of the malicious use of AI technologies, which shows its potential impact depending on which attack stage is applied. Finally, the classification of defensive approaches is represented in the third tier. A more detailed explanation of the three tiers is provided in the following three subsections.

5.2.2 Tier 1: Attack stages and their objectives. The highest dimension of structure is the seven stages of the Cyber Kill Chain [45], which tries to answer the question of when an attacker can achieve his malicious objectives based on the cyber attack life cycle. This helped us show a graphical representation of the required steps to launch an AI-based cyber attacks. We divided the seven stages into three major divisions: (i) planning, (ii) intrusion, and (iii) execution. Planning is composed of the reconnaissance and weaponization stage, conducting research about the target, and the process

Table 4. Case studies of AI-based cyber attacks - Summary

Article	Year	Attack Vector	Target	Category
DeepLocker – Concealing Targeted Attacks with AI Locksmithing [22]	2018	Highly targeted and evasive malware, which hides its attack payload without being detected until it reaches a specific target.	Video conferencing applications	Next-generation malware
Availability Attacks on Computing Systems through Alteration of Environmental Control: Smart Malware Approach [16]	2019	Self-learning malware, which is able to induce indirect malicious attacks that masquerade as accidental failures on computing infrastructure by compromising the environmental control systems.	Computing infrastructure	
Using AI to Hack IA: A New Stealthy Spyware Against Voice Assistance Functions in Smart Phones [89]	2018	Attacking framework to record the activation voice stealthily by adopting NLP techniques, and to play the activation voice of user by designing an IED module.	Voice assistants	Voice synthesis
Artificial Intelligence-Based Password Brute Force Attacks [80]	2018	Next-generation AI-based password brute force attacks by constructing the attacking dictionary in a more intelligent way based on prior passwords.	Computer authentication systems	Password-based attacks
PassGAN: A Deep Learning Approach for Password Guessing [41]	2018	Fully automated password guessing technique based on GAN, by learning the distribution from actual password leaks.	Password-based systems	
Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter [71]	2016	A highly automated method of end-to-end spear phishing, by discovering high-value targets and spreading personalized machine-generated content automatically.	Twitter	Social bots
DeepPhish: Simulating Malicious AI [7]	2018	Weaponization of ML algorithm with the goal of learning to create better phishing attacks and making it undetectable from detection systems.	AI phishing detection systems	
Automated Crowdturfing Attacks and Defenses in Online Review Systems [88]	2017	A new automated review attack for large-scale users' opinion manipulation, using DNN-based fake review generation.	User-generated review sites	
Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN [43]	2017	Automated approach based on GAN for generating adversarial examples to bypass ML-based black-box malware detection systems.	ML-based black-box malware detection systems	Adversarial training
DeepDGA: Adversarially-tuned domain generation and detection [4]	2016	An automated generation of malware domains using GAN that learns to bypass malware detection mechanisms powered by DNNs.	DGA classifier	
Weaponizing Machine Learning: Humanity was overrated anyway [63]	2017	A new ML hacking tool "DeepHack," which learns to break into web applications using NNs and reinforcement learning.	Web applications	

of weaponizing deliverables. Intrusion describe the process of delivering, exploiting, and installing the malicious payload in order to gain access to the target. Finally, a successful intrusion moves to execution, in which the adversary often establishes paths and acts to achieve his objectives. In addition, each stage describes the capability of an adversary to attack the system by the malicious use of AI technologies. We highlighted the usage of AI's advantages in cyber-offense in the following categories:

- **AI-targeted:** Sophisticated attacks depend on a well-prepared planning phase. The ability of AI to understand, interpret, and **find patterns in vast amounts of data** can be used to provide in-depth analysis and to create targeted exploration processes by overcoming human limitations. For example, in the case of SNAP_R [71], the selection of the most valuable targets to send the phishing messages to was done using the k-means clustering method.

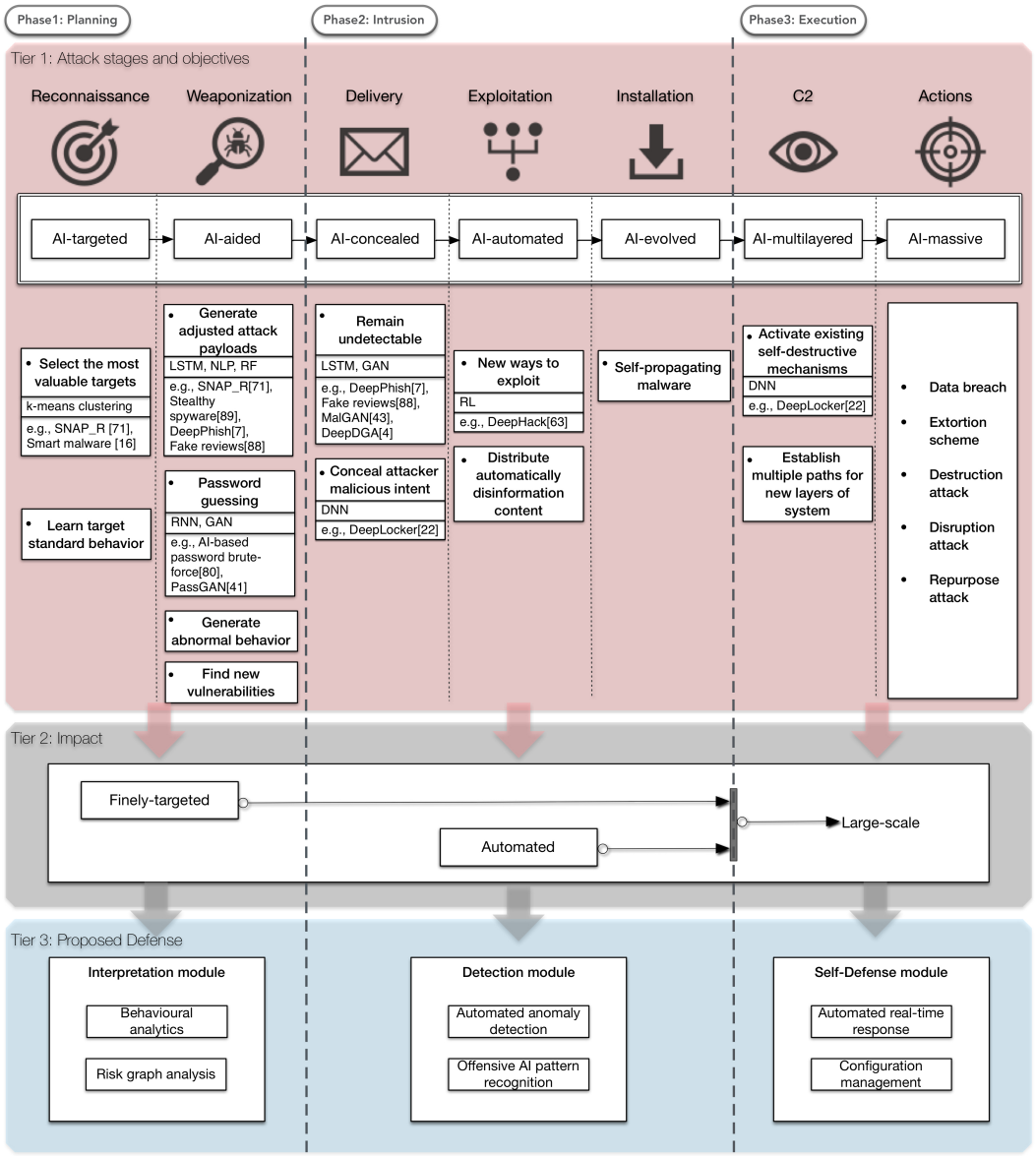

Fig. 4. AI-Based Cyber Threat Framework.

- **AI-aided:** Developments in AI can contribute to weaponization by **replicating the abilities of human beings**. Studies have shown that it is able to understand human natural language in order to create personalized messages that contain malicious payload. For instance, in the case of a fake reviews attack [88], the RNN-generated deceptive fake reviews generate fully adjusted attack payloads for opinion manipulation campaigns.

- **AI-concealed:** Transmission of the malicious payload can utilize more **stealthy** techniques to remain undetectable and concealed for long periods. For example, in the case of DeepPhish [7], the generation of phishing URLs using LSTM was sufficient enough to remain undetectable by detection systems.
- **AI-automated:** After gaining access to the target, the growing use of AI capabilities can increase the number of attackers who are able to carry out the attack, avoiding the need for human resources. The use of botnet could comprise **autonomous decision-making bots** that can expand the attack at greater length. For example, in the case of DeepHack [63], another way was found to exploit vulnerabilities using RL, by learning on its own how to do hacking.
- **AI-evolved:** The use of an **automated self-propagating** malware spreads the infection by repeating the same or similar discovery and exploitation with new hosts, enabling access across other parts of the network. The malware propagation behavior tries to affect as many nodes as possible, although direct infection of a few critical control nodes may cause greater damage.
- **AI-multilayered:** The need for persistent access for monitoring could be achieved by the ability of AI to **learn independently based on inputs from the environment** and therefore to control aspects of the targets' behavior automatically. For example, in the case of DeepLocker [22], the unlocking of attack conditions was done using DNN when the target was identified.
- **AI-massive:** The required actions to successfully achieve the attacker's objectives can include data breach, extortion schemes, destruction attacks, disruption attacks, and repurpose attacks [25]. Massive attacks require coordinated action and the participation of AI in at least one stage of each of the three major divisions: planning, intrusion, and execution.

5.2.3 Tier 2: Impact classification of malicious AI. From the framework, an impact-based classification of AI-based attacks is derived as three different levels: Finely-targeted, Automated, and Large-scale. This analysis is based on three characteristics, namely scope, speed, and scale that computers traditionally excel at better than humans [70]. The Finely-targeted class refers to the scope of identification of particular kinds of attack patterns from large datasets. An adversary can acquire a more accurate target list to adapt his attack strategy. In the Automated class, an adversary can launch attacks in milliseconds. So here, speed outperforms by defeating other machines faster than can be done manually by humans. Finally, the Large-scale class comprises both characteristics from the Finely-targeted and Automated classes in order to automatically infect millions of devices with the desired features to increase the attack success rate. The classification shown in Figure 5 indicates the potential impact related to which attack stage is affected by the malicious use of AI.

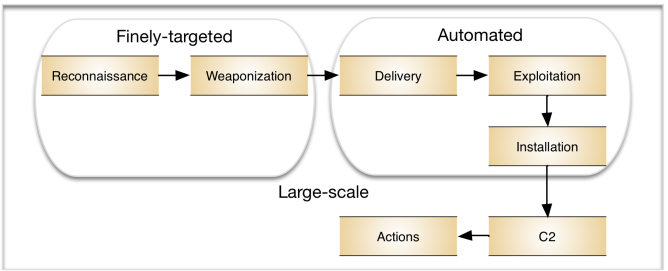

Fig. 5. Impact-based classification of malicious AI.

5.2.4 Tier 3: Classification of proposed defense methods. Defending against AI-based cyber attacks cannot be done by a simple solution or a single tool. There is a need for an in-depth defensive approach in the whole cyber attack life cycle to combat the intelligence. Therefore, it is very important to research potential offensive AI capabilities used in attacks and establish some theoretical best practices for defense. Based on our observation on attackers' objectives, we categorize possible defense methods of AI-based cyber attacks into three major categories based on the common objective features of the stages in the phases of planning, intrusion, and execution: (i) interpretation module, (ii) detection module, and (iii) self-defense module. In each category, we propose some potential measures that could be considered, as shown in the third tier of Figure 4.

One of the first steps to defending against AI-based cyber attacks is to understand your target and your enemy. Hence, the interpretation module focuses on behaviors by creating an environmental map for the standard behavior and an attack surface to establish AI-behavioral attack patterns. The detection module focuses on autonomously detecting the presence of malicious activity based on the fingerprint for every application in the previous module. The decision rules are triggered by the established patterns of malicious behavior to detect deviations from a normal behavior profile. For instance, the attack botnet detection rate can be improved by applying LSTMs to detect unseen botnet traffic based on behavioral analysis [79]. After the intrusion, the self-defense module could develop the ability of self-learning the optimal defense strategy. To address the issue of intelligent self-learning adversarial behavior, security analysts should consider applying the same strategy in order to determine defensive actions towards the field of RL, according to the environmental map and attack surface [57]. Moreover, efforts from the organization OpenC2 [60] to develop standard-driven orchestration language for machine-to-machine communication to enable faster attack responses, contribute to the vision of more accurate, automated, and coordinated cyber defense approaches.

The adversaries are improving their attack strategy through automation, so the only efficient way to mitigate such attacks at machine speed is with automation. Defenders need to be faster, by developing defensive options in a time frame that prevents the attacker from accomplishing his objectives. Therefore, defensive AI is related to rapid response to both detection and remediation and to self-learning in defending the system, which will increase its potential to work effectively in the game of AI offense and defense.

5.3 A Smart Grid Scenario

The main purpose of this section is to demonstrate how to generalize our proposed framework with the goal of identifying new attacks to develop proper defenses. Our framework raises the awareness that any system can be a potential target of AI-based cyber attacks. Although the current AI-based cyber attacks do not target sCPS yet, we believe CPS, and especially sCPS, are vulnerable to those attacks. We hereby use a hypothetical attack scenario on a smart grid infrastructure to explain how the framework can be used to identify new attacks in terms of the malicious use of AI. The proposed framework is based on the traditional Cyber Kill Chain, but goes beyond that because the latter does not consider advanced attacks such as AI-based cyber attacks, and hence cannot properly identify them in order to develop appropriate defenses. Therefore, having an AI-based cyber threat classification according to the attack stages can facilitate the design of appropriate mitigation strategies such as those proposed in our framework based on interpretation, detection, and self-defense actions.

Smart Grid (SG) is a cornerstone for all critical infrastructures. If power is unavailable for a long enough time, all other critical functions of the city will also be hit. As defined in [32], "*a smart grid is an electricity network that can cost-efficiently integrate the behavior and actions of all users connected to it – generators, consumers, and those that are both – to ensure an economically efficient,*

sustainable power system with low losses and high levels of quality and security of supply and safety." To realize the SG architecture, the National Institute of Standards and Technology (NIST) provided a model, as shown in Figure 6, which shows all the possible roles involved in the SG.

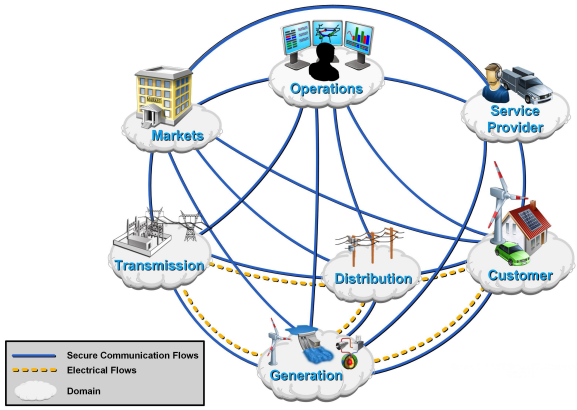

Fig. 6. The NIST Conceptual Model for SG [58].

With the integration of ICT in the SG infrastructure, we have more cyber dependency interactions on critical applications and infrastructures. The state of operation depends on information transmitted through the infrastructure via electronic links. Outputs of one infrastructure are inputs to the other infrastructure and "*the commodity passed among the infrastructure assets is information*" [62]. More precisely, in an SG, several smart systems interact with the SG to manage city services and infrastructures more efficiently by processing all the necessary data to make decisions. The electrical infrastructure can ensure resilient delivery of energy to supply many functions to other critical infrastructures and all of them are dependent on its proper operation, as shown in Figure 7. An attack on the SG infrastructure is liable to have an immediate effect on other infrastructures such as transportation systems, hospitals, industrial production, and even aviation [54].

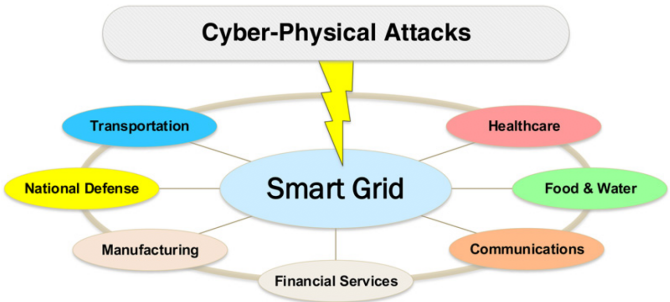

Fig. 7. Vulnerable interdependent sectors under cyber-physical attacks on the SG [39].

Therefore, ongoing developments in technologies are continuously discovering new threats to sCPS and especially critical infrastructures operating via Supervisory Control and Data Acquisition (SCADA) systems with significant impacts. Sensors extract real-time information from electricity, water, gas, fuel, communications, and transportation and then send the information through the controllers deployed in the field to the control center [54]. Data is the fuel of an SG, and AI-driven

data technologies could operationalize the analysis of data into valuable information. However, the malicious use of AI means that anyone can compromise nodes and then control these slaves of automated decisions makers. An adversary could take advantage of the interdependencies across critical infrastructure systems, particularly the reliance on information technologies, and change the system's behavior to a malicious state.

The goal of this hypothetical scenario is to illustrate that an AI-massive cyber attack on an SG could have cascading effects from interruption of services onto multiple sectors, in order to provide awareness of new potential threats and help critical supply utilities deploy suitable defenses.

5.3.1 AI Attack Scenario. In our scenario, the adversary demonstrates a variety of capabilities to perform his malicious activity against the SG as a multi-stage intrusion on multiple critical infrastructure sectors. To simplify the explanation, we limited our focus to the distribution domain of the SG and its interaction with the customer domain through smart metering infrastructure. The advanced metering infrastructure (AMI) is a basic component of an SG and aims to handle the communication between utility smart meters and the utility company. AMI is a system that belongs to the distribution domain and consists of data collectors, head-end systems, smart meters, and meter data management (MDM) systems.

In this scenario, there is a compromised data collector in a wide area network (WAN) that has bidirectional connections with other data collectors and field smart meters devices for electricity and information flows. Data collectors are responsible for collecting energy consumption data from smart meters, and sending it to the control center. Therefore, a compromised data collector with a worm implanted can *"propagate in the network infrastructure to infect other data collectors and proceed to impact the connected smart meters"* [37]. It can aggregate massive data collected and then control smart meters by sending malicious commands like disconnect requests. The inclusion of the most critical dependencies and the automation of intrusion could create an AI-massive attack. Figure 8 shows the application of our proposed framework by illustrating the offensive AI capabilities through automation of traditionally manual processes, allowing the attacker to conduct an attack on a larger scale, and some proposed corresponding defenses.

The following analysis provides a high-level overview of threat actors' activities within the cyber attack life cycle:

Stage 1: Reconnaissance

An improved target's reconnaissance could use AI for observing the normal behavior and operations related to the distribution infrastructure that has interdependecies with other sectors. Here, an adversary can acquire topological, structural, and operational information about the distribution network from sensor measurements to identify critical relationships with the intended targets. In general, sensor measurements are aggregated and processed by distributed data collectors at different locations in the SG. Investigation of the topology can reveal multiple critical attack schemes. Then, the attacker is able to determine the most critical lines in a local grid that can influence other interdependent sectors. In the transmission domain, studies [40, 90] used the heuristic risk graph methods to transform massive amounts of data into a risk and an influence graph, respectively, that describe the critical paths revealing cascading failures. Moreover, a data-driven approach based on RL is proposed to identify the critical attack sequence against sequential topology attacks [87]. Since the connection among different critical infrastructure assets are complex, AI technologies could help threat actors identify **AI-targeted** attack patterns in large amounts of data.

Stage 2: Weaponization

The proper timing for attacks that are launched in a sequential manner is critical for achieving large-scale impacts. For instance, study [83] showed that attacks launched on the nodes with the highest or lowest loads can have different impacts for network robustness. The attacker can use

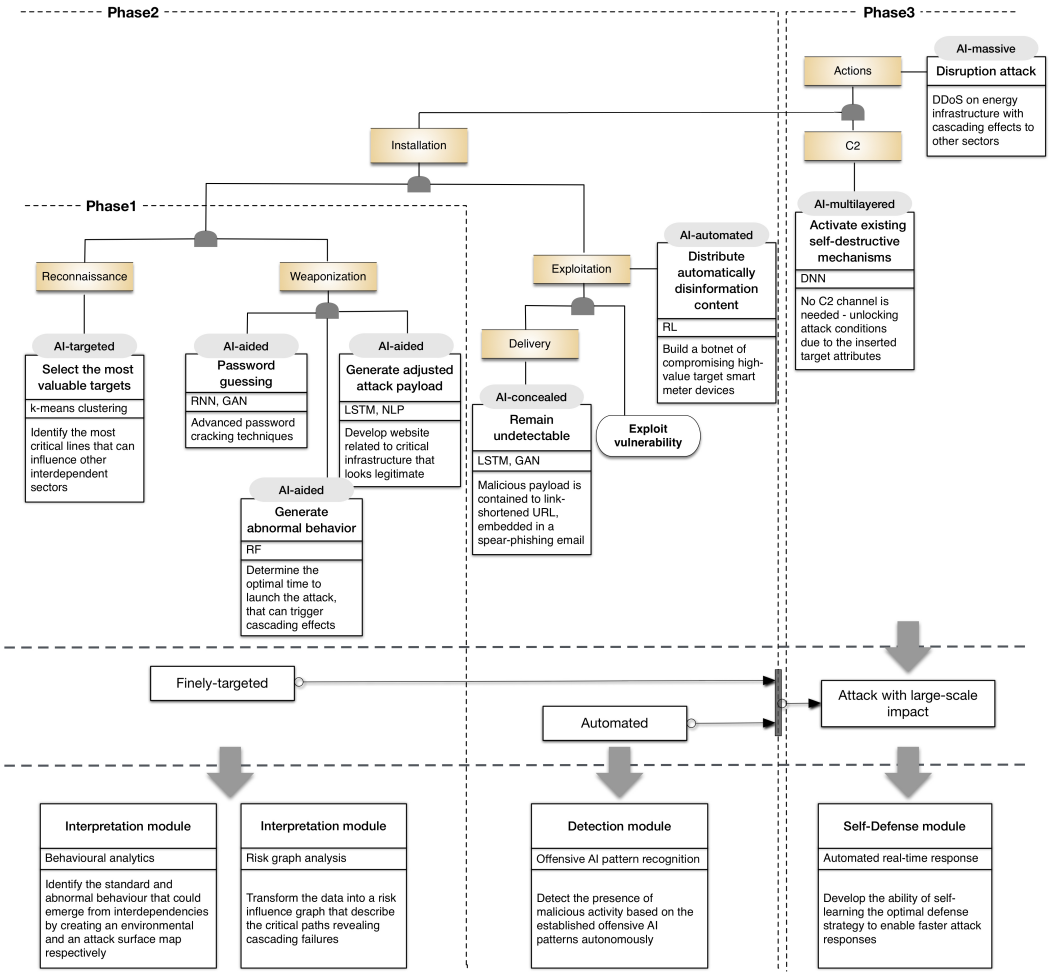

Fig. 8. Application of proposed framework on the SG scenario.

AI to design a recognition module to determine the right time to launch the **AI-aided** attack on target nodes to trigger cascading effects as described in Stealthy Spyware [89]. The attacking commands will be embedded in this adjusted malicious payload that generates unusual behaviors to disconnect devices from energy utility. The data collector contains its own local web server and is browser-controlled, making it easy to configure via the Internet [74]. Threat actors can develop watering holes to embed the malware such as websites related to critical infrastructure with malicious content that looks as legitimate. Moreover, the threat actors can use advanced password-cracking techniques as Trieu and Yang *et al.* [80] described, and PassGAN [41] to obtain correct passwords.

Stage 3: Delivery

Throughout the spear-phishing email campaign, the threat actor can adapt his abnormal behavior "on the fly" based on the target environment to compromise other collectors. The malicious payload is contained in a shortened URL that, when clicked, will trigger malicious commands. To hide its malicious activity, the payload can be concealed by using **AI-concealed** techniques to evade

detection because the identification of malware by reverse-engineering is difficult to perform, as described in DeepLocker [22].

Stage 4: Exploitation

The adversary can take advantage of the interconnection of millions of smart meters and build a botnet by compromising high-value target devices to automatically distribute the malicious payload. When it is injected into the data controller, it can launch an **AI-automated** attack by directly sending the attacking commands to harm multiple critical infrastructures. Therefore, these large-scale distributed sensors are at risk of malicious hackers turning them off all at once [5, 37].

Stage 5: Installation

The attacker may infect as many nodes as possible in favor of the autonomous decisions that **AI-evolved** self-propagating malware is able to make.

Stage 6: Command & Control

An attacker commonly tries to establish a channel for further communication. However, by using AI techniques, no C2 channel will be needed, as described in DeepLocker [22]. Due to the already inserted target attributes, the malware knows when it will be unlocked across different types of nodes. Therefore, it is an **AI-multilayered** attack that can provide access to other system components automatically and remotely.

Stage 7: Actions on Objectives The successful **AI-massive** attack in an AMI topology will cause large-scale disruption throughout an entire city based on the cascading effects to other domains, indirectly by the energy infrastructure. Energy disconnections can lead to interruption of services, which will have strong impacts on a city's stability.

6 DISCUSSION

6.1 Comparison with related work

AI is already being used to defend against cyber attacks in different ways, ranging from malware classification, finding anomalous network traffic, and botnet traffic identification to phishing detection. For instance, when classifying malicious binaries, an algorithm can train itself through known benign and malicious examples to classify new activities as benign or malicious without requiring prior description of them. However, every invention has a dark side. Adversaries could take advantage of it. An adversary could combine AI technologies with existing cyber attack techniques to expand the cyber threat landscape.

This paper aims to provide an overview of the state-of-the-art AI-based cyber attacks for a better understanding of advanced attack vectors that use AI maliciously throughout the different stages of the cyber attack life cycle. There is little research related to the ways in which AI can be used as a malicious tool by adversaries. To the best of our knowledge, this is the first study to provide a full description of "AI-based cyber attacks" regarding the intersection of AI and malicious intents. It is the first review of AI methods used for offensive purposes. Our work contributes to an analysis of intended AI-driven attacks in cyber-offense compared with other similar studies that we explained in Section 3, which were mainly focused on general AI risks. Floridi [31] argued that AI does not mean only intelligent machines but also the intelligent way in which machines work. Therefore, the intelligent behavior of a machine is defined from its outcomes, including self-learning and autonomous actions. With beneficial AI applications, it is critical to be aware of the "dark side" in the cases when they are used maliciously.

We presented the first classification that targets deliberate actions in which AI can be used as an instrument in supporting malicious activities. Our framework raises the awareness to continue this work by discovering other problems that may lead to malicious AI along with appropriate solutions to such threats. Existing classifications [81, 82, 85] focused generally on AI risks and

mostly from the safety point of view. For example, they focus on risks of creating or modifying an AI system to be dangerous, and not on how to use AI as an adversarial tool to attack more accurately and efficiently. However, our framework shows that even traditional systems can also be weaponized by the intentional malicious use of AI. However, the list of attack objectives in the proposed framework may be incomplete. It is possibly missing other predictions of AI offensive capabilities.

Our proposed AI-based cyber threat framework based on the Cyber Kill Chain offers a structured way to describe AI attacks and their behaviors. In general, the Cyber Kill Chain provides better understanding on the offensive actions of a cyber attacker, allowing defenders to develop appropriate defense strategies against the steps the attacker goes through to execute an attack. Similarly, the application of our framework can facilitate the understanding of AI offensive capabilities to identify opportunities to disrupt an AI-based cyber attack in progress. It can be applied for both offensive and defensive purposes. For instance, in cyber-game exercises, the framework can serve as a way to build AI attack scenarios for testing the strength of defenses in the target. On the defensive side, the framework helps better identify possible and more appropriate mitigation strategies that can compete at a wider scope, at a faster speed and on a larger scale of the AI-based cyber attacks.

So far, AI seems to be one of the most promising technologies for research in information security, and it plays an important role in cyber crime prevention and detection [24]. Bruce Schneier [70] said that "*both attack and defense will benefit from AI technologies*" and as long as computers are moving into activities that are traditionally done well by humans, it might create new asymmetries in the attack-defense balance. Our investigation of possible attack schemes using AI maliciously could serve as an important step to establishing a good combination of cybersecurity, safety, and resilience properties in a vulnerable system. However, many research examples, which present these novel attacks, address the need for further research to develop proper defenses.

6.2 Research gaps and recommendations

We have analyzed 11 AI-based cyber attack case studies using AI-driven techniques in their attack process in literature. We have identified various attack strategies that leverage AI techniques to maximize their impact, where there is a lack of sufficient mitigation approaches. Most of the existing solutions focus on enhancing security policies, security control monitoring, and multi-factor authentication mechanisms to combat new machine-generated threats, but they are inadequate to address the increasing speed. Clearly, there are many open problems on how to prevent and mitigate such advanced threats. In particular, we believe that the most efficient way to fight AI is using AI in order to compete at scope, speed, and scale. Building autonomous cyber defenses that learn from experiences during the cyber races between attackers and defenders can reveal the presence of malicious behavior more efficiently. In this section, we discuss AI-based strategies that can be used to mitigate AI-based cyber threats.

AI has played an important role in deploying cybersecurity solutions by analyzing activities in real time to detect and prevent security risks. However, emerging sophisticated threats using AI maliciously make current defensive approaches inadequate to address the increasing accuracy and speed. As demonstrated in the AI-based cyber attack case studies, the process of their attack models integrates learning features, increasing the sophistication level in terms of the advanced planning, intrusion, and execution strategy. Despite the sophistication of such attacks, the same steps of the attack strategies can be used for defensive purposes in a supervised manner. This time, the security engineer has full knowledge of the data collected that flows between several devices and systems.

An important thing that should be taken into consideration is that AI can support the automation of cyber-defense tasks, such as vulnerability assessment, intrusion detection, incident response, and threat intelligence processing. Leslie F. Sikos [75], in his book, presented various AI approaches

to build proactive and reactive defense mechanisms against malicious cyber activities. In terms of self-adaption and self-learning, AI is utilized to make smarter and more robust cyber defenses that can anticipate efficiently against attacks. By incorporating reinforcement learning methods, a system can solve complex and dynamic security problems by learning from its own experiences through the exploration of its environment. For instance, Nguyen *et al.* [57] surveyed several deep reinforcement learning (DRL) methods for autonomous defense strategies, such as autonomous DRL-based intrusion detection systems and multi-agent DRL-based game theory approaches, to obtain optimal policies in different attacking scenarios. In particular, Feng and Xu [29] proposed an optimal DRL-based cyber defense for CPS in the presence of unknown threats. Moreover, Shamshirband *et al.* [73] investigated works dealing with computational intelligence approaches in intrusion detection and prevention systems. To conclude, such potential solutions will not stop the effectiveness of the AI-based cyber attacks, but will reduce the impact if we identify them on time.

7 CONCLUSION

Threat actors are constantly changing and improving their attack performance with a particular emphasis on the application of AI-driven techniques in the attack process. This study investigates the offensive capabilities through automation of traditionally manual processes, allowing attackers to conduct attacks of a wider scope, at a faster speed, and on a larger scale. In this paper, we explored research examples of cyber attacks, posed by combining the "dark" side of AI with the attack techniques. We introduced an analytic framework for modeling those attacks that can be useful in understanding their context and identified key opportunity areas for the security community in implementing suitable defenses. Finally, we illustrated a scenario to show that an sCPS, e.g., smart grid, can be the target of more advanced malicious cyber activity.

ACKNOWLEDGMENTS

We would like to express our gratitude to Associate Professor Michail Maniatakos at New York University Abu Dhabi, for all his beneficial advice and the discussions we had.

REFERENCES

[1] Horizon 2020 Work Programme 2014-2015. 2015. Leadership in enabling and industrial technologies: Information and Communication Technologies. Retrieved Nov 25, 2019 from http://ec.europa.eu/research/participants/portal4/doc/call/h2020/common/1587758-05i._ict_wp_2014-2015_en.pdf

[2] Terrence Adams. 2017. AI-powered social bots. *arXiv preprint arXiv:1706.05143* (2017).

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[4] Hyrum S Anderson, Jonathan Woodbridge, and Bobby Filar. 2016. DeepDGA: Adversarially-tuned domain generation and detection. In *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*. ACM, 13–21.

[5] Ross Anderson and Shailendra Fuloria. 2010. Who controls the off switch?. In *2010 First IEEE International Conference on Smart Grid Communications*. IEEE, 96–101.

[6] Daniele Antonioli, Giuseppe Bernieri, and Nils Ole Tippenhauer. 2018. Taking control: Design and implementation of botnets for cyber-physical attacks with cpsbot. *arXiv preprint arXiv:1802.00152* (2018).

[7] Alejandro Correa Bahnsen, Ivan Torroledo, David Camacho, and Sergio Villegas. 2018. DeepPhish: Simulating Malicious AI. In *2018 APWG Symposium on Electronic Crime Research (eCrime)*. 1–8.

[8] Oliver Bendel. 2019. The synthetization of human voices. *AI & SOCIETY* 34, 1 (2019), 83–89.

[9] UC Berkeley. 2012. Cyber-Physical Systems – a Concept Map. Retrieved Nov 25, 2019 from https://ptolemy.berkeley.edu/projects/cps/

[10] William Blum. 2017. Neural fuzzing: applying DNN to software security testing. Retrieved Nov 25, 2019 from https://www.microsoft.com/en-us/research/blog/neural-fuzzing/

[11] Michael Bossetta. 2018. A simulated cyberattack on Twitter: Assessing partisan vulnerability to spear phishing and disinformation ahead of the 2018 US midterm elections. *arXiv preprint arXiv:1811.05900* (2018).

[12] Susan M Bridges, Rayford B Vaughn, et al. 2000. Fuzzy data mining and genetic algorithms applied to intrusion detection. In *Proceedings of 12th Annual Canadian Information Technology Security Symposium*. 109–122.

[13] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228* (2018).

[14] Zheng Bu. 2014. Zero-Day Attacks are not the same as Zero-Day Vulnerabilities. Retrieved Nov 25, 2019 from https://www.fireeye.com/blog/executive-perspective/2014/04/zero-day-attacks-are-not-the-same-as-zero-day-vulnerabilities.html

[15] Tomas Bures, Danny Weyns, Bradley Schmer, Eduardo Tovar, Eric Boden, Thomas Gabor, Ilias Gerostathopoulos, Pragya Gupta, Eunsuk Kang, Alessia Knauss, et al. 2017. Software engineering for smart cyber-physical systems: challenges and promising solutions. *ACM SIGSOFT Software Engineering Notes* 42, 2 (2017), 19–24.

[16] Keywhan Chung, Zbigniew T Kalbarczyk, and Ravishankar K Iyer. 2019. Availability attacks on computing systems through alteration of environmental control: smart malware approach. In *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*. ACM, 1–12.

[17] Jessica Cussins. 2017. AI Researchers Create Video to Call for Autonomous Weapons Ban at UN. Retrieved Nov 25, 2019 from https://futureoflife.org/2017/11/14/ai-researchers-create-video-call-autonomous-weapons-ban-un/

[18] Moises Danziger and Marco Aurelio Amaral Henriques. 2017. Attacking and Defending with Intelligent Botnets. *XXXV Simpósio Brasileiro de Telecomunicaç oes e Processamento de Sinais-SBrT* 2017 (2017), 457–461.

[19] DarkTrace. 2018. The Next Paradigm Shift: AI-Driven Cyber-Attacks. White Paper. Retrieved Nov 25, 2019 from https://www.darktrace.com/en/resources/wp-ai-driven-cyber-attacks.pdf

[20] DARPA. 2016. Cyber Grand Challenge (CGC).

[21] Kenneth De Jong. 1988. Learning with genetic algorithms: An overview. *Machine learning* 3, 2-3 (1988), 121–138.

[22] Marc Ph. Stoecklin Dhilung Kirat, Jiyong Jang. 2018. DeepLocker - Concealing Targeted Attacks with AI Locksmithing. In *Black Hat USA Conference*.

[23] Wenrui Diao, Xiangyu Liu, Zhe Zhou, and Kehuan Zhang. 2014. Your voice assistant is mine: How to abuse speakers to steal information and control your phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*. ACM, 63–74.

[24] Selma Dilek, Hüseyin Çakır, and Mustafa Aydın. 2015. Applications of artificial intelligence techniques to combating cyber crimes: A review. *arXiv preprint arXiv:1502.03552* (2015).

[25] Peter Eder-Neuhauser, Tanja Zseby, Joachim Fabini, and Gernot Vormayr. 2017. Cyber attack models for smart grid environments. *Sustainable Energy, Grids and Networks* 12 (2017), 10–29.

[26] ENISA. 2018. ENISA Threat Landscape Report 2017. Retrieved Nov 25, 2019 from https://www.enisa.europa.eu/publications/enisa-threat-landscape-report-2017

[27] ESET. 2018. Can Artificial Intelligence Power Future Malware? White Paper. Retrieved Nov 25, 2019 from https://www.welivesecurity.com/wp-content/uploads/2018/08/Can_AI_Power_Future_Malware.pdf

[28] Gregory Falco, Arun Viswanathan, Carlos Caldera, and Howard Shrobe. 2018. A Master Attack Methodology for an AI-Based Automated Attack Planner for Smart Cities. *IEEE Access* 6 (2018), 48360–48373.

[29] Ming Feng and Hao Xu. 2017. Deep reinforecement learning based optimal defense for cyber-physical system in presence of unknown cyber-attack. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–8.

[30] Brian E. Finch. 2013. Anything and Everything Can Be Hacked. Retrieved Nov 25, 2019 from https://www.huffpost.com/entry/caveat-cyber-emptor_b_3748602

[31] Luciano Floridi. 2017. Digital's cleaving power and its consequences. *Philosophy & Technology* 30, 2 (2017), 123–129.

[32] European Regulators Group for Electricity and Gas. 2010. Position paper on smart grids. Retrieved Nov 25, 2019 from http://www.cired.net/publications/workshop2010/pdfs/0092.pdf

[33] National Science Foundation. 2015. Cyber-Physical Systems (CPS). Retrieved Nov 25, 2019 from https://www.nsf.gov/pubs/2015/nsf15541/nsf15541.pdf

[34] Jairo Giraldo, Esha Sarkar, Alvaro A Cardenas, Michail Maniatakos, and Murat Kantarcioglu. 2017. Security and privacy in cyber-physical systems: A survey of surveys. *IEEE Design & Test* 34, 4 (2017), 7–17.

[35] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[36] Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).

[37] Aaron Hansen, Jason Staggs, and Sujeet Shenoi. 2017. Security analysis of an advanced metering infrastructure. *International Journal of Critical Infrastructure Protection* 18 (2017), 3–19.

[38] Simon Hansman and Ray Hunt. 2005. A taxonomy of network and computer attacks. *Computers & Security* 24, 1 (2005), 31–43.

[39] Haibo He and Jun Yan. 2016. Cyber-physical attacks and defences in the smart grid: a survey. *IET Cyber-Physical Systems: Theory & Applications* 1, 1 (2016), 13–27.

[40] Paul DH Hines, Ian Dobson, and Pooya Rezaei. 2017. Cascading power outages propagate locally in an influence graph that is not the actual grid topology. *IEEE Transactions on Power Systems* 32, 2 (2017), 958–967.

[41] Briland Hitaj, Paolo Gasti, Giuseppe Ateniese, and Fernando Perez-Cruz. 2017. Passgan: A deep learning approach for password guessing. *arXiv preprint arXiv:1709.00440* (2017).

[42] White House. 2016. Artificial intelligence, automation, and the economy. *Executive office of the President. https://obamawhitehouse. archives. gov/sites/whitehouse. gov/files/documents/Artificial-Intelligence-Automation-Economy. PDF* (2016).

[43] Weiwei Hu and Ying Tan. 2017. Generating adversarial malware examples for black-box attacks based on GAN. *arXiv preprint arXiv:1702.05983* (2017).

[44] Abdulmalik Humayed, Jingqiang Lin, Fengjun Li, and Bo Luo. 2017. Cyber-physical systems security—A survey. *IEEE Internet of Things Journal* 4, 6 (2017), 1802–1831.

[45] Eric M Hutchins, Michael J Cloppert, and Rohan M Amin. 2011. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research* 1, 1 (2011), 80.

[46] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

[47] Rida Khatoun and Sherali Zeadally. 2017. Cybersecurity and privacy solutions in smart cities. *IEEE Communications Magazine* 55, 3 (2017), 51–59.

[48] Young Mie Kim, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, Richard Heinrich, Robyn Baragwanath, and Garvesh Raskutti. 2018. The stealth media? Groups and targets behind divisive issue campaigns on Facebook. *Political Communication* 35, 4 (2018), 515–541.

[49] Thomas C King, Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi. 2019. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and engineering ethics* (2019), 1–32.

[50] Jingyue Li, Jin Zhang, and Nektaria Kaloudi. 2018. Could We Issue Driving Licenses to Autonomous Vehicles?. In *International Conference on Computer Safety, Reliability, and Security*. Springer, 473–480.

[51] Tao Liu, Wujie Wen, and Yier Jin. 2018. SIN 2: Stealth infection on neural network—A low-cost agile neural Trojan attack methodology. In *2018 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*. IEEE, 227–230.

[52] Natasha Lomas. 2017. Lyrebird is a voice mimic for the fake news era. Retrieved Nov 25, 2019 from https://techcrunch.com/2017/04/25/lyrebird-is-a-voice-mimic-for-the-fake-news-era/

[53] William Melicher, Blase Ur, Sean M Segreti, Saranga Komanduri, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Fast, lean, and accurate: Modeling password guessability using neural networks. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 175–191.

[54] Harel Menashri and Gil Baram. 2015. Critical infrastructures and their interdependence in a cyber attack–the case of the US. *Military and Strategic Affairs* 7, 1 (2015), 22.

[55] MITRE. 2017. ATT&CK Matrix for Enterprise. Retrieved Nov 25, 2019 from https://attack.mitre.org/matrices/enterprise/

[56] Jefferson Seide Molléri, Kai Petersen, and Emilia Mendes. 2016. Survey guidelines in software engineering: An annotated review. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 58.

[57] Thanh Thi Nguyen and Vijay Janapa Reddi. 2019. Deep Reinforcement Learning for Cyber Security. *arXiv preprint arXiv:1906.05799* (2019).

[58] NIST. 2018. NIST framework and roadmap for smart grid interoperability standards, release 4.0 – DRAFT. Retrieved Nov 25, 2019 from https://www.nist.gov/engineering-laboratory/smart-grid/smart-grid-framework

[59] Ivan Novikov. 2018. How AI can be applied to cyberattacks. Retrieved Nov 25, 2019 from https://www.forbes.com/sites/forbestechcouncil/2018/03/22/how-ai-can-be-applied-to-cyberattacks/

[60] OASIS. 2017. Open Command and Control (OpenC2). Retrieved Nov 25, 2019 from https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=openc2

[61] Dan Patterson. 2018. How weaponized AI created a new breed of cyber-attacks. Retrieved Nov 25, 2019 from https://www.techrepublic.com/article/how-weaponized-ai-creates-a-new-breed-of-cyber-attacks/

[62] Frederic Petit, Duane Verner, David Brannegan, William Buehring, David Dickinson, Karen Guziel, Rebecca Haffenden, Julia Phillips, and James Peerenboom. 2015. *Analysis of critical infrastructure dependencies and interdependencies*. Technical Report. Argonne National Lab.(ANL), Argonne, IL (United States).

[63] D. Petro and B. Morris. 2017. Weaponizing Machine Learning: Humanity was Overrated Anyway. In *DEF CON*.

[64] Federico Pistono and Roman V Yampolskiy. 2016. Unethical research: how to create a malevolent artificial intelligence. *arXiv preprint arXiv:1605.02817* (2016).

[65] Ashis Pradhan. 2012. Support vector machine-A survey. *International Journal of Emerging Technology and Advanced Engineering* 2, 8 (2012), 82–85.

[66] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *URL https://openai. com/blog/better-language-models* (2019).
[67] Mohit Rajpal, William Blum, and Rishabh Singh. 2017. Not all bytes are equal: Neural byte sieve for fuzzing. *arXiv preprint arXiv:1711.04596* (2017).
[68] Patrick Reidy and K Randal. 2013. Combating the insider threat at the FBI: real world lessons learned. In *RSA Conference, San Francisco CA*.
[69] Bruce Schneier. 1999. Attack trees. *Dr. Dobb's journal* 24, 12 (1999), 21–29.
[70] Bruce Schneier. 2018. Artificial intelligence and the attack/defense balance. *IEEE Security & Privacy* 2 (2018), 96–96.
[71] John Seymour and Philip Tully. 2016. Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter. *Black Hat USA* 37 (2016).
[72] John Seymour and Philip Tully. 2018. Generative Models for Spear Phishing Posts on Social Media. *arXiv preprint arXiv:1802.05196* (2018).
[73] Shahaboddin Shamshirband, Nor Badrul Anuar, Miss Laiha Mat Kiah, and Ahmed Patel. 2013. An appraisal and design of a multi-agent system based cooperative wireless intrusion detection computational intelligence technique. *Engineering Applications of Artificial Intelligence* 26, 9 (2013), 2105–2127.
[74] Manish Shrestha, Christian Johansen, and Josef Noll. 2017. Security Classification for Smart Grid Infra structures (long version). (2017).
[75] Leslie F Sikos. 2018. *AI in Cybersecurity*. Vol. 151. Springer.
[76] Kaj Sotala and Roman V Yampolskiy. 2014. Responses to catastrophic AGI risk: a survey. *Physica Scripta* 90, 1 (2014), 018001.
[77] Marc Ph. Stoecklin. 2018. *DeepLocker: How AI Can Power a Stealthy New Breed of Malware.* Retrieved Nov 25, 2019 from https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/
[78] Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 1017–1024.
[79] Pablo Torres, Carlos Catania, Sebastian Garcia, and Carlos Garcia Garino. 2016. An analysis of recurrent neural networks for botnet detection behavior. In *2016 IEEE biennial congress of Argentina (ARGENCON)*. IEEE, 1–6.
[80] Khoa Trieu and Yi Yang. 2018. Artificial Intelligence-Based Password Brute Force Attacks. (2018).
[81] Alexey Turchin. 2015. *A Map: AGI Failures Modes and Levels.* Retrieved Nov 25, 2019 from https://www.lesswrong.com/posts/hMQ5iFiHkChqgrHiH/
[82] Alexey Turchin and David Denkenberger. 2018. Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY* (2018), 1–17.
[83] Jian-Wei Wang and Li-Li Rong. 2009. Cascade-based attack vulnerability on the US power grid. *Safety science* 47, 10 (2009), 1332–1336.
[84] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. Citeseer, 38.
[85] Roman V Yampolskiy. 2016. Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.
[86] Roman V Yampolskiy and MS Spellchecker. 2016. Artificial intelligence safety and cybersecurity: A timeline of AI failures. *arXiv preprint arXiv:1610.07997* (2016).
[87] Jun Yan, Haibo He, Xiangnan Zhong, and Yufei Tang. 2017. Q-learning-based vulnerability analysis of smart grid against sequential topology attacks. *IEEE Transactions on Information Forensics and Security* 12, 1 (2017), 200–210.
[88] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. 2017. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1143–1158.
[89] Rongjunchen Zhang, Xiao Chen, Jianchao Lu, Sheng Wen, Surya Nepal, and Yang Xiang. 2018. Using AI to Hack IA: A New Stealthy Spyware Against Voice Assistance Functions in Smart Phones. *arXiv preprint arXiv:1805.06187* (2018).
[90] Yihai Zhu, Jun Yan, Yan Lindsay Sun, and Haibo He. 2014. Revealing cascading failure vulnerability in power grids using risk-graph. *IEEE Transactions on Parallel and Distributed Systems* 25, 12 (2014), 3274–3284.