ELEKTROTEHNIŠKI VESTNIK 91(3): 73-94, 2024
OVERVIEW SCIENTIFIC PAPER

# A safer future: Leveraging the AI power to improve the cybersecurity in critical infrastructures

**Mojca Volk**

*Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška cesta 25, 1000 Ljubljana, Slovenija*
*E-pošta: mojca.volk@fe.uni-lj.si*

**Abstract.** In the intricate landscape of the cybersecurity, critical infrastructures represent the most vital systems underpinning the societal and economic wellbeing, with their disruption or incapacitation having potentially catastrophic consequences. The increasing complexity, digitalization and interconnectedness of these systems have rendered them susceptible to a broad spectrum of risks challenging the existing paradigm of the safety and security. Thus, securing critical infrastructures against the escalating cybersecurity threats has become an essential yet extremely challenging endeavour. In the light of these considerations, the paper offers a deeper understanding of the dynamic, adaptive and intelligence-driven approaches in the cyber defence that leverage the AI power, thus representing a transformative innovation with a potential to redefine the security strategies and frameworks in critical infrastructures. The cybersecurity threats and vulnerabilities are addressed and the existing and emerging approaches and best practices in the sector-specific intrusion detection and prevention systems and deception technology are investigated, followed by an in-depth study of AI applications in the cyber defence. This includes the current approaches and early best practices complemented by a discussion on advanced topics, such as explainable and adversarial AI. Finally, guidelines are drafted to inform and provide guidance on the introduction of the AI applications for the cyber defence purposes in critical infrastructures.

**Keywords:** Critical infrastructure (CI), cybersecurity, artificial intelligence (AI), deception technology

### Varnejša prihodnost: Izboljšava kibernetske varnosti kritičnih infrastruktur s pomočjo umetne inteligence

Kritične infrastrukture predstavljajo v kontekstu kibernetske varnosti vitalne sisteme, ki gradijo temelje družbene in ekonomske blaginje. Zaradi naraščajoče kompleksnosti, digitalizacije in medsebojne povezanosti so ti sistemi vedno bolj dovzetni za širok spekter tveganj, uspešni kibernetski napadi nanje pa lahko povzročijo katastrofalne posledice. Posledično je zaščita kritičnih infrastruktur in njihova odpornost na intenzivno rastoč spekter kibernetskih groženj danes bistvenega pomena, obenem pa predstavlja vedno bolj kompleksen in večstranski izziv. V luči teh premislekov si ta članek prizadeva vzpostaviti globlje razumevanje dinamičnih, prilagodljivih in z inteligenco podprtih pristopov v kibernetski obrambi, ki za svoje delovanje izkoriščajo moč umetne inteligence. Slednja predstavlja transformativno inovacijo s pomembnim potencialom za redefinicijo varnostnih strategij v kritičnih infrastrukturah. Članek naslovi področje kibernetskih tveganj in ranljivosti kritičnih infrastruktur in razišče nove pristope in primere iz prakse na področju sektorsko-specifičnih sistemov za zaznavanje in preprečevanje vdorov ter tehnologij zavajanja. Podrobno analizira možne aplikacije umetne inteligence v okviru kibernetske obrambe, vključno z ilustracijo prvih primerov iz prakse in diskusijo odprtih raziskovalnih vprašanj, kot sta na primer razložljiva in sovražna umetna inteligenca. Na podlagi vzpostavljenega razumevanja nato predlaga izhodišča, ki usmerjajo in informirajo vpeljavo aplikacij umetne inteligence za kibernetsko obrambo v kritične infrastrukture.

**Ključne besede:** Kritična infrastruktura, kibernetska varnost, umetna inteligenca, tehnologije zavajanja

*Received 22 April 2024*
*Accepted 2 June 2024*

## 1 INTRODUCTION

As the modern world harnesses the benefits of digitalization and emerging technologies, nations and people are becoming increasingly dependent on a safe and resilient operation of critical infrastructures (CI), which in turn have taken over a fundamental and irreplaceable role in sustaining the functioning of modern societies. CIs are the backbone of essential services, providing the necessary foundation for the economic and social security, public health, and safety of a nation [1]. CIs include physical and virtual systems, assets, networks and services, and encompass a diverse range of sectors, including energy infrastructures, transportation systems (airports, highways, and railways), water supply and waste management facilities, communication systems, healthcare facilities, financial institutions, emergency services, and defence. The increasing complexity, digitalization and interconnectedness have rendered them susceptible to a broad spectrum of risks, challenging the existing paradigm of the safety and security. Safeguarding the CIs' resilience and security has become paramount, making them a prime focus for research and innovation.

The cybersecurity emerges as a particularly daunting challenge within this context. The progressive digitalization and integration of CIs with the information systems and emerging technologies create space for an ever evolving and increasingly complex landscape of the cybersecurity vulnerabilities. The expanse of the attack surface grows with each newly adopted ICT technology, such as Industrial Control Systems (ICS), Unmanned Aerial Vehicles (UAVs), autonomous systems, Internet of Things (IoT), as well as the advanced technologies such as Artificial Intelligence (AI) [1][2]. The cyber threats exhibit a diverse range of actors, motivations, and attack vectors. State-sponsored adversaries, criminal organizations, hacktivists, and even insider threats have demonstrated their ability to exploit vulnerabilities and launch targeted cybersecurity attacks against networks, systems, and personnel. The ramifications of the cybersecurity attacks can be far-reaching, ranging from data breaches and information theft to disruption of command-and-control systems, malfunctions of the logistics, manipulation of CI, and even the compromise of the national security. Although in place, the traditional cyber defence mechanisms are often outpaced by the agility of the cyber threats that evolve continuously, thereby necessitating a more dynamic, adaptive, and intelligence-driven approach to safeguard these essential systems.

In the light of these considerations, AI presents a transformative innovation with a potential to redefine the security strategies and frameworks for CI. The AI’s ability to analyse large volumes of data at an unprecedented speed enables the identification of the potential cyber threats before they materialize. The Machine Learning (ML) algorithms in particular can evolve in response to past incidents and emerging threats and provide predictive insights that human operators may not discern. Automation of monitoring and maintenance operations using AI can significantly diminish the likelihood of the human error and oversight and reduce the window of opportunity for the cyber attackers to exploit vulnerabilities.

The paper offers a deeper understanding of the vulnerabilities and threats CI is exposed to. It summarizes the cybersecurity attack types and provides illustrative examples of major incidents observed in the past decades. It provides a review of the cyber defence technologies, methods, best practices and strategies as observed in different CI types, focusing on sector-specific applications, followed by an in-depth investigation of the necessary yet challenging introduction of AI to the cyber defence. It focuses on the early AI best practices and illustrative examples and discusses advanced topics and emerging research avenues. This knowledge is gathered in order to analyse and establish an understanding of the types of threats the CI is exposed to through the adoption of AI, and to draft a guidance for CI operators on mitigation and protection possibilities.

The remainder of the paper is organized as follows. Section 2 discusses the CI cybersecurity landscape and the types of cyber attacks on CI. Section 3 provides a review of the applicable cyber defence strategies and technologies illustrated with known best practices from different CI sectors. Section 4 delves deeper into the adoption of AI in the CI cyber defence, focusing on technological aspects and early real-world examples. Section 5 discusses advanced topics and outstanding research challenges. Section 6 details the guidelines drafted based on the current knowledge and best practices to inform and steer CI operators in introducing AI applications for the CI cyber protection purposes. Section 7 draws conclusions of the presented work.

## 2 CI CYBERSECURITY THREAT LANDSCAPE

Compared to the cyber attacks observed in the IT systems, the cyber attacks targeting CIs exhibit certain complexities and consequences, likely to be contributed to the prevailing trend in CI of the converging operational technology (OT) and traditional information technology (IT) environments (see Figure 1). Such infrastructures, including power plants, transportation systems and water treatment facilities, rely on specialized systems for their operation, such as Supervisory Control and Data Acquisition (SCADA), Industrial Internet of Things (IIoT) and ICS systems. These systems merge the legacy and modern technologies to manage physical processes in the infrastructure, and have not always been designed with the cybersecurity in mind, making them particularly vulnerable to attacks that could lead not only to data breaches, but also to a physical damage and disruption of essential services [3]. This includes the use of insecure protocols and interfaces with a lack of encryption and insufficient authentication measures, insufficient OT network monitoring, absence of the network segmentation, software security issues, such as Windows and Linux operating system vulnerabilities, or outdated equipment, lack of the access control in real-time OT solutions, invisibility of the devices, etc. [1]. Secondly, in addition to the IT-OT convergence, modern CIs are progressively adopting the state of the art and emerging technologies, such as mobile communication networks, cloud infrastructure and IIoT, leading to an increased interconnectedness and exposure of critical services and capabilities. As a result, CIs themselves are becoming increasingly interconnected. This further expands the attack surface and amplifies the potential for the cascading failures and devastating damage. Energy-related CIs are specifically illustrative of this vulnerability where a failure of a smart grid can cause outages, failures as well as physical and virtual damage in almost all other CIs. Such multi-faceted exposure of CIs provides ample opportunities for the attackers to exploit vulnerabilities, particularly in parts of the infrastructure that provide a real-time control and monitoring of CI to maintain its efficiency, stable operation, safety and reliability, thus having the capability to cause severe damage or disruption of critical services [2].

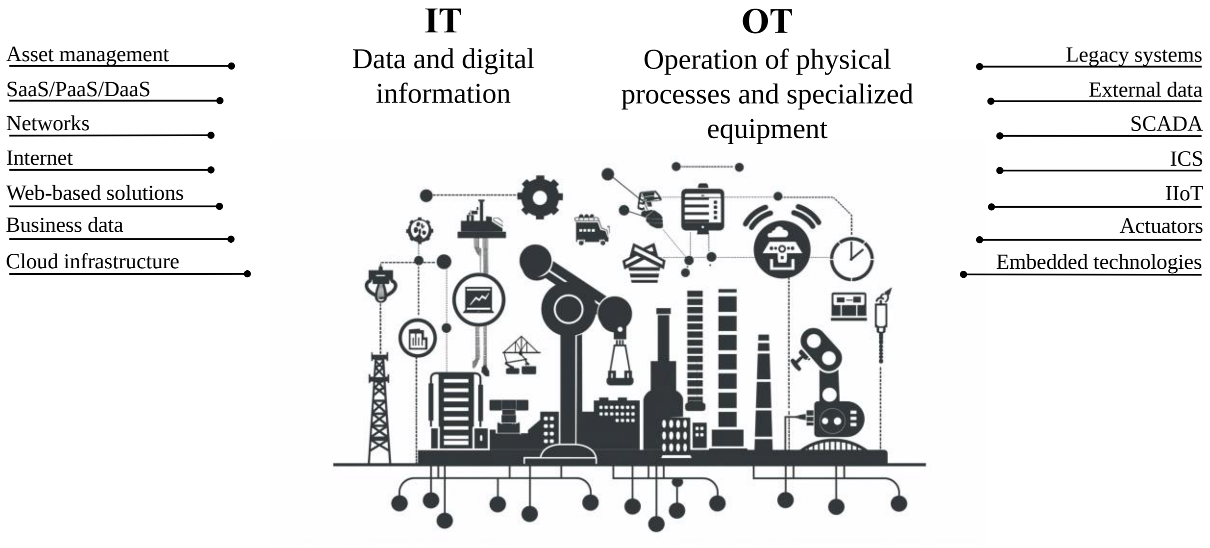

Figure 1. Integration of the IT and OT environments in CI.

In consequence, there is a progressive increase in the number and variety of the cyber attacks on CI, which is highly concerning. Both the number and sophistication of the cyber attacks targeting the OT systems in particular are fast advancing [4][5] and each individual CI sector is continuously challenged by a plethora of emerging threats specific to their domain. The technologies and methods used to execute a cyber attack on CI are diverse and multifaceted. The prevailing threats include the IT and OT malware that typically exploit network security vulnerabilities and social engineering to gain access and propagate through critical control capabilities, advance persistent threats (APTs), insider threats, nation-state attacks, and ransomware. Ransomware in particular is one of the earliest forms of the cyber attacks, which has been closely followed by malware. Both types of the cyber attacks have been observed for decades and through time, the attack methods and mechanics have steadily progressed in terms of invasiveness, evasiveness, and sophistication.  The APT attacks are specifically concerning in the context of CI. They are prevailingly nation-state sponsored and target the data theft and tampering the control and management capabilities, particularly in the energy CI. A general trend can be observed about a typical progression of such an APT attack, i.e. the cyber kill chain process that involves an initial attack to gain access to CI through phishing, insider attack or supply chain attack, followed by a deployment of malware to implement sabotage actions, and finally data exfiltration. Types of the attackers and their incentives include cybercrime groups interested in financial gains, state-sponsored groups pursuing espionage and disruption fuelled by geo-political events, and hacktivists. The most common types of the attacks, their goals and mechanics are summarized in Table 1 [6][7].

According to the European Union Agency for Cybersecurity (ENISA), 2022 and 2023 have seen a notable escalation in the cybersecurity attacks on CI, both in terms of the variety and number of incidents as well as their consequences, with ransomware and Distributed Denial of Service (DDoS) attacks representing over 50 % of all the detected cyber attacks [6]. The Center for Strategic and International Studies (CSIS) [8] maintains a report about the significant cybersecurity incidents since 2006 targeting government agencies, defence and high-tech companies or inducing economic losses of over one million US dollars. A simple statistical analysis of the reported incidents confirms the concerning growth in the number of the cyber attacks on CI (see Figure 2).

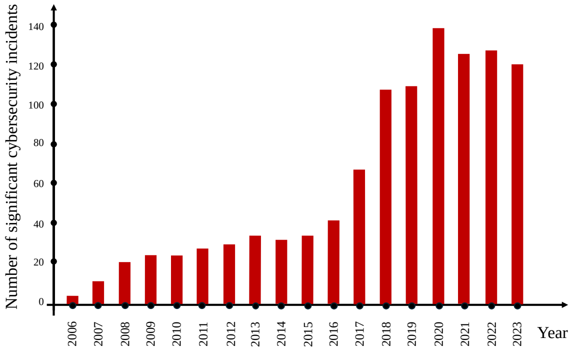

Figure 2. Number of the significant cybersecurity incidents since 2006 [8].

Table 1: Types of cyber attacks, their mechanics and goals

| Attack type | Mechanics and goals |
|---|---|
| Ransomware | A cybersecurity attack where the attackers take control of the target's assets (e.g. encryption of the files containing sensitive data) and demand a ransom in exchange for the return of the asset's availability. |
| Malware | A cyber attack that involves the use of a malicious software or firmware designed to damage, disrupt or gain an unauthorized access to the systems that will have an adverse impact on the integrity, confidentiality, or availability. Also known as a malicious code and malicious logic. Examples include viruses, trojan horses, worms and spyware. |
| Social engineering | Malicious activities that attempt to exploit the human error or human behaviour with the objective of gaining access to information or services. It relies on various forms of manipulation, including phishing, pretexting, baiting and scareware, with the goal to trick victims into making mistakes, handing over sensitive information, visiting websites, granting access to systems or services, or perform other types of actions that compromise the security. |
| Threats against data | Malicious activities aimed at stealing, altering or destroying digital information classified as sensitive, confidential or protected. The attacks can be classified in two basic groups, i.e. the data breach where the attempt is to deliberately gain an authorized access and release data, and data leak where the attempt is to cause events, such as a human error or misconfiguration that can consequently cause an unintentional loss or exposure of data. The primary consequences of such attacks include privacy breaches, financial losses and damage to reputation. Man-in-the-middle attacks fall within this category. |
| Distributed Denial of Service (DDoS) | A well-known and prevailing type of cyber attacks attempting to compromise the availability of systems, services, data or other resources, by exhausting the resources or overloading the components of the network infrastructure. Attackers typically use a network of hijacked resources to launch the assault. |
| Threats against availability | Intentional or unintentional disruption causing Internet outages, blackouts and shutdowns of censorship. This can happen because of government-directed shutdowns, massive natural events such as earthquakes or cyclones, as well as incidents such as power outages, cyber attacks, technical failures, or military actions. The frequency and diversification of the threats continues to proliferate, resulting in a significant monetary loss to national economies. |
| Information manipulation | Cyber attacks where a false or misleading information is deliberately spread or a genuine information is altered to deceive, mislead, or influence individuals or systems. This can involve an altering digital content, creating fake news, or manipulating the data to compromise decision-making processes, public opinion, or the integrity of the information systems. The attacks rely mostly on non-illegal behaviours that can cause potentially negative impacts on values, procedures or political processes. False data injection is one type of the attack of this group. |
| Supply chain attack | Targets less-secure elements in the supply network to compromise the final product or organization. By infiltrating a trusted vendor or component, attackers exploit these relationships to distribute malware or gain an unauthorized access to sensitive systems and data |

A detailed threat landscape analysis reveals that a significant number of the cybersecurity incidents are reported across a variety of the CI sectors, including government infrastructures, defence, healthcare, communications, energy, banking and finance, and transport, with all types of attacks represented. Two very relevant resources in this respect are MITRE ATT&CK for ICS framework that serves as a live online common industry lexicon managed by the MITRE Corporation that documents the tactics and techniques of attacks on the OT systems through eleven categories [9], and the recently released NSA Elitewolf, a Github repository containing various ICS/SCADA/OT focused signatures and analytics made available for the IC operators to identify and detect a potentially malicious cyber activity in their OT environments [10]. Following a simple keyword search applied to the report provided by CSIS [8], the volume of the reported attacks per a specific CI sector is presented in Figure 3, where the majority of the reported incidents target government infrastructures and services, followed by defence and energy sectors. Interestingly, compared to the statistics reported in [1], the government-related CIs have only recently emerged as a major target, which can be attributed to the current worldwide and regional geopolitical tensions.

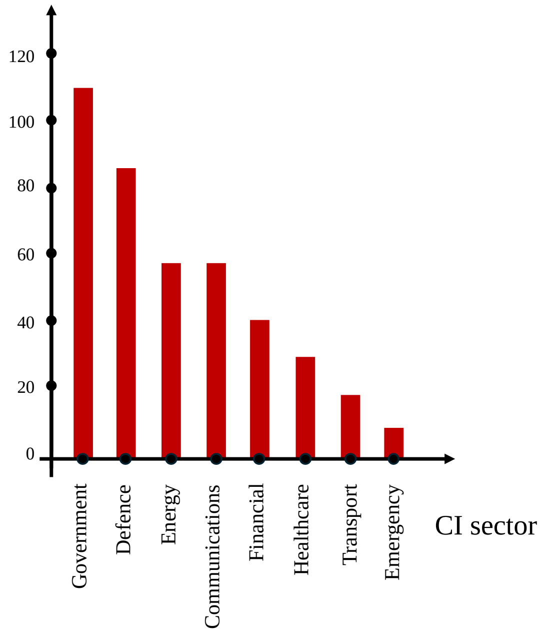

Figure 3. Number of the significant cybersecurity incidents since 2006 [8].

Some of the most well-known incidents attacking energy infrastructures were the 2021 Colonial Pipeline ransomware attack [11], causing a significant financial loss and public distress, and the Ukraine Power Grid Attacks in 2015 and 2016 using BlackEnergy 3 malware that led to disruptions to the Ukrainian power grids, causing widespread energy outages [12]. Two other notorious incidents are the 2010 Stuxnet attack on the

Iranian nuclear facilities, a highly sophisticated malware attack that was believed to have targeted the uranium enrichment infrastructure and was suspected to be nation-state sponsored [13], and the 2020 SolarWinds supply chain attack where a malicious software was installed on a major software upgrade, resulting in more than 18.000 affected businesses [4]. Table 2 provides a review of the prevailing types of the cyber attacks encountered in specific types of CI with other examples of real-world incidents. The timeline of the incidents is presented in Figure 4. It must be noted, however, that the provided statistics and the selected examples are not representative of the entire volume of cyber attacks on CI, partially due to the exclusion of smaller-scale incidents and due to the scarcity of incident reports for the security and privacy reasons, in particular in the most devastating or security-sensitive cases.

The illustrated threats against CI and escalating cybersecurity concerns in general necessitate a specialized approach with a comprehensive visibility of the entire infrastructure and more sophisticated cyber defence mechanisms to effectively prioritize and manage the known and suspected vulnerabilities, as discussed in the next section.

Table 2: The prevailing types of the cyber attacks on CI with examples of the well-known attack incidents

| CI attack type and target | Examples |
|---|---|
| **OT MALWARE**<br>Malware specifically designed to target the OT systems to gain control of management functions, reconfiguration of control capabilities, execution of DoS or data theft/exposure in an OT system. | Stuxnet [13] – changes the control configuration/capabilities by exploiting OS vulnerabilities to affect the PLC operation. It was used in the 2010 attack on the Iranian nuclear infrastructure.<br>Havex [14] – exploits through remote access capabilities or compromised supply chain by compromising the installation SW to gain access to and report information about specific servers in the OT system.<br>Industroyer (Crashoverried) [15] – targets industrial control protocols to directly control switching and circuit devices, wipe data and execute DoS on specific devices. Used in the 2016 (Industroyer) and 2022 (Industroyer 2) attacks on the Ukrainian power system.<br>TRITON [16] – accesses and reprograms safety instrumented system controllers causing the controller to enter a failed safe state, automatically shutting down the industrial process.<br>BlackEnergy [5] – exploits the MS Office and remote access vulnerabilities. Used in the 2016 Ukrainian power grid attack. |
| **ADVANCED PERSISTENT THREAT (APT)**<br>A sophisticated attack campaign in which the intruder establishes a long-term presence within the system. The initial infection vector exploits a compromised supply chain and/or social engineering attacks, possibly combined with other decoy tactics, such as DDoS, to install malware enabling a remote access. This is followed by an expansion to critical parts of the system causing an illicit access to the sensitive data or even malfunctions, and finally extraction when sensitive data is exported from the system using again a range of distraction tactics, e.g. a DDoS. | SolarWinds Orion APT [17] – an APT attack in 2020 exploiting a supply chain attack by infiltrating a malicious code into a network monitoring and configuration management software suite to install a backdoor into the system, followed by privilege escalation and user impersonation to penetrate further into the system and perform different types of attacks and compromises.<br>A41APT [18] – an ATP attack exploiting the SSL-VPN vulnerabilities and installing malware (SodaMaster, P8RAT and FYAnti) to deploy a remote management tool. Used to target multiple industries, including for example the Japanese manufacturing industry in 2010. |
| **INSIDER THREAT**<br>Malicious actions taken by individuals within an organization that can compromise the CI availability and security. Insider threats can either be accidental or intentional. | Stradis Healthcare [19] – during the COVID-19 pandemic in 2020, a former employee accessed the company shipping system and deleted shipping data, causing delays in the delivery of the personal protective equipment.<br>Pegasus Airlines [20] – personally identifiable information exposed in 2022 as a result of a cloud misconfiguration by a system administrator.<br>South Georgia Medical Center [21] – patient test results, names, and birth dates were leaked in 2021 by a former employee who used a USB stick to download the exposed data. |
| **NATION-STATE ATTACK**<br>A cyber attack carried out by a state-sponsored actor against another government or a private organization, the goal of which is espionage, disruption, destruction or political message. It exploits a variety of methods, such as phishing, DDoS, malware and ransomware. | Russian cyberattack on Ukraine [22] in Feb. 2022 caused disruptions to broadband satellite internet access services by disabling modems that provided communication via Viasat Inc's KA-SAT satellite network. Malware AcidRain was believed to be used and a probable goal was to disrupt Ukrainian command and control during the invasion. |
| **RANSOMWARE**<br>An attack using a custom platform-specific SW designed to encrypt, lock or exfiltrate data. Most ransomware attacks use e-mails as the delivery method and target PCs/workstations and exploit vulnerabilities of Windows OS. Supply chain ransomware is a specific subgroup of attacks where ransomware is distributed through trusted SW distribution methods (e.g. SW updates). | Colonial pipeline attack [23][21] – used DarkSide RaaS and caused a shortage of the gas supply for customers by compromising the management computer system through compromised credentials for a legacy VPN.<br>JBS USA [21]– was based on REvil RaaS and caused a shutdown of beef manufacturing operations in 2021.<br>Maersk [21] – was attacked in 2017 using NotPetya malware, which exploited the EternalBlue Windows vulnerability and spread via backdoor in MeDoc SW, and locked the system used to operate shipping terminals all over the world.<br>WannaCry attack on NHS [24] – carried out in 2017 by exploiting a Microsoft security vulnerability on PCs, leading to a network closure and blockage of essential services in NHS, including e.g., ambulance handover process, transfer of CT/MR scans and chemo orders, etc. |

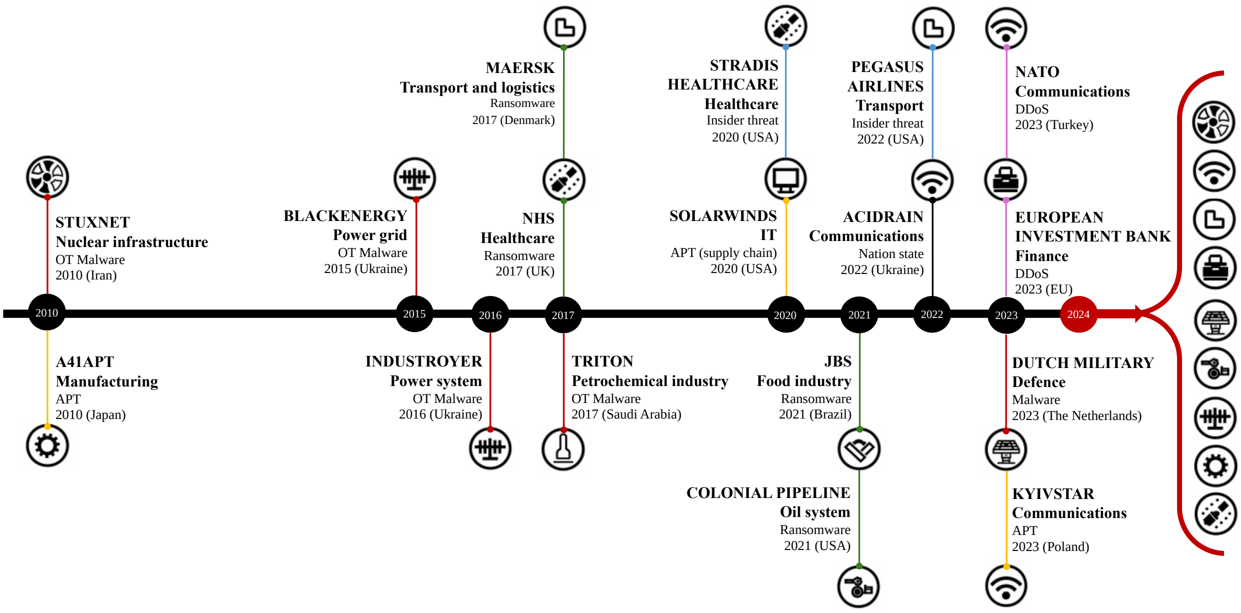

Figure 4. Timeline of major cyber incidents targeting CI.

## 3 Cyber defence strategies and technologies in CI

The complexity and persistence of the CI cyber threats has recently led to establishment of multi-layered and integrated cyber defence strategies that focus on resilience of the infrastructure and services by combining a multitude of complementary approaches and technologies. Cybersecurity advisory organizations, such as the US National Institute of Standards and Technology (NIST) and Cybersecurity and Infrastructure Security Agency (CISA), as well as evidence from the best practices encourage the use of proactive and adaptive approaches relying on a real-time detection and assessment, continuous monitoring and intelligence-driven analysis to identify, detect, protect against, respond and profile specific cyber attacks. In this respect, a combination of passive and active cybersecurity measures should be considered when establishing a trusted and robust security system able to protect the CI, crucial data, and the user privacy, and specialized deception technology should be employed.

### *3.1 Adaptive CI cybersecurity strategy*

Figure 5 depicts the stages of an adaptive cybersecurity strategy in reference to the timing of a cybersecurity incident. The respective aims and methods of individual stages, and typical systems and tools implementing them are the following [25][26].

The first stage, prediction, identifies the most probable cyber attacks, targets and attack methods in advance, i.e., before the incident occurs, becomes apparent or causes negative effects. Prediction relies on trend analysis, threat intelligence and historical data to predict probable attack vectors and targets. It comprises also a risk assessment to identify vulnerabilities and prioritize the threats, with a focus on critical assets that if compromised would have a most significant impact on the public safety and services.

The second stage focuses on the prevention of an attack by securing the infrastructure from external cyber attacks in order to avoid the occurrence of any damage or loss. The prevention is focused on the measures directly blocking a cyber attack or creating conditions that install limits or prevent the attack from succeeding, e.g., securing the infrastructure (firewalls, antivirus and anti-malware SW, encryption etc.), training employees, and implementing robust security policies and procedures. This entails implementation of robust cybersecurity frameworks designed for IT and specific OT environments [5], such as the IEC 62443 series and NIST SP 800-82 for securing the ICS and OT systems, including the network security, access control, and incident response, ISO/IEC 27001 for information security management systems (ISMS), including the OT protection guidelines, ENISA OT Cybersecurity Recommendations, including threat intelligence, network security, and incident response, and sector-specific regulation, e.g., NERC CIP standards for the energy sector, as well as adherence to industry standards to establish and enforce guidelines and best practices designed to protect information systems against cyber threats. This stage incudes implementation of a supply chain security to prevent infiltration through third parties, robust backup and redundancy strategies ensuring that all critical data and systems can be quickly restored in the event of a cyber incident, minimizing the downtime and operational impact, and adoption of the Zero Trust

Architecture to minimize internal and external risks by adopting approaches, such as the least-privileged access, micro-segmentation and continuous monitoring of the network activity, to prevent an unauthorized access and data breaches as well as to limit and a contain cross-contamination, in particular from the IT to the OT parts of the infrastructure. The development and maintenance of an incident response and recovery plan is also part of the prevention stage, specifying procedures for responding to the cybersecurity incidents (roles and responsibilities, communication plans, recovery procedures for systems and services restoration after an incident). Regular red teaming exercises and penetration testing should also take place to simulate the cyber attacks and test the effectiveness of the adopted security measures in a controlled environment. The last but not the least, this stage entails establishment of the training and awareness programmes for employees and collaboration capabilities with government agencies for threat intelligence sharing and coordinated responses to threats.

The third stage takes place after an incident occurrence. It focuses on identification of an ongoing attack that can no longer be prevented in order to establish a timely awareness and initiate appropriate response procedures. It entails implementation of continuous monitoring and layered defence capacities for a fast anomaly detection, monitoring of suspicious activities, thus providing a timely awareness of potential issues and mitigation of any potential incidents. This includes combinations of several different complementary approaches and techniques, including intrusion detection systems (IDS), security information and event management (SIEM) systems, and anomaly detection tools, as well as other advanced active threat intelligence and cyber protection technologies, such as intrusion protection systems (IPS) and research and operational honeypots.

Finally, the fourth stage implements a response and recovery. It relies on the incident response techniques, methods and solutions designed to take specific actions in an attempt to contain an ongoing cyber attack and mitigate and manage its consequences. The response can be focused on the threat eradication, system recovery to a normal operation, and forensic analysis for learning and future security strengthening purposes.

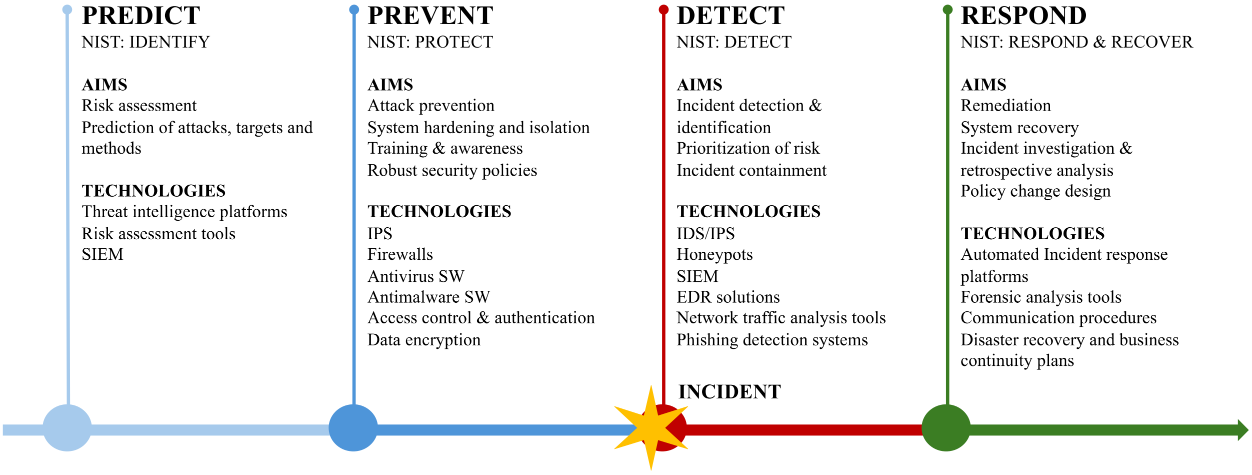

Figure 5. Stages of the adaptive cybersecurity.

## 3.2 *Cyber defence technologies*

Today, there is a broad range of approaches and dedicated technologies available for the detection, monitoring, profiling and behavioural analytics of cyber attacks as well as deterrence and active engagement with the threat itself. They can be broadly categorized as passive and active cybersecurity measures. The passive measures, largely found in the traditional approaches, comprise a set of the security practices, tools, and technologies that are reactive by nature, work in the background or are part of a layered defence strategy and operate without a direct intervention during an attack. They include firewall configurations, antivirus and antimalware software, encryption, and intrusion detection systems (IDS) [27][28]. However, the passive methods are becoming increasingly inefficient against sophisticated and adaptive cyber attacks, where more proactive and dynamic defence approaches are required. Thus, the emerging strategies are increasingly leveraging the capabilities of active and dynamic actions to detect, respond to and mitigate threats through a direct hands-on interaction with the threat as it occurs. This includes intrusion prevention systems (IPS), red team exercises and penetration testing, ethical hacking, incident response teams and comprehensive threat hunting activities such as the Cyber Kill Chain and MITRE ATT&CK framework [27][28][29].

The IDS and IPS systems, collectively referred to as IDPS, serve as the first line of the defence against a wide range of cyber threats, including external hacking attempts and insider threats, providing capabilities to identify and mitigate potential security threats before they can impact the system integrity and functionality. The IDS systems monitor the infrastructure for malicious activities and policy violations, raising alerts whenever a known threat or a suspicious activity potentially indicating a new type of attack is detected or whenever an unauthorized activity deviates from the established policies. They are categorised as the network-based IDS when designed for monitoring an incoming network, and host-based IDS when focused on individual device monitoring, and as signature-based IDS when relying on predefined patterns of the known threats to identify attacks, anomaly-based IDS when ML and statistical modelling are used to identify deviations from a normal behaviour, and specification-based IDS combining the benefits of the signature and anomaly-based IDS approaches by manually specifying the behavioural characteristics of an attack [3][30]. The IPS systems further extend the IDS detection capabilities by taking predefined actions in real-time to prevent the exploitation of vulnerabilities. This includes actions to report, block, suspend or reset suspected malicious activities, e.g., termination of malicious processes, blocking of suspicious IP addresses and rerouting the malicious traffic, or modifying the firewall rules to enhance the security posture. The measures installed by IDPS can collectively help identify vulnerabilities and proactively tackle ongoing attacks in real-time, and provide the capabilities and assets for detection, mitigation, monitoring and management of the cybersecurity incidents. There are numerous implementations and extensive research is underway on the IDPS capabilities and technologies for different types of CI, in particular in the smart energy and more generally in the IIoT and ICS sectors. Some selected examples include an IDS for autonomous distributed IoT systems [31], a ML-based IPS for unmanned aircraft systems [32], and a distributed IDS for the SCADA systems in smart grids [33].

The deception technology, i.e. honeypots, honeynets and other forms of digital decoys, introduces an additional layer of defence in a dynamic and adaptive security environment [5]. In general, the cyber deception protects networks by creating uncertainties and complexities for the attackers, thus increasing the costs and risks associated with their activities. The scientific basis for these technologies is their ability to mimic real systems designed to mislead the attackers into engaging with the decoys, thereby revealing their presence and providing an opportunity to observe their tactics and intentions without compromising the actual resources [34][35]. A honeypot is a technology that complements and expands the field of operation of the IDPS systems to improve the detection of the zero-day attacks in signature-based IDS/IPS systems, and to support the operation of the anomaly-based IDPS systems towards a more accurate detection [36]. There are two specific groups of the honeypots. The research honeypots are implemented in an isolated manner and separately from actual CI, deliberately exposing interesting systems, services and capabilities and thus setting traps for cyber attackers in order to observe the attack characteristics and collect data, which is used for a detailed analysis, profiling and planning of further defence measures. The production or in-network honeypots, on the other hand, enhance security procedures in an actual infrastructure [36]. They are embedded directly inside CI that is being protected and serve as active decoys, luring the attackers away from the actual resources, thus providing a real-time protection as well as intelligence collection for the analysis and profiling purposes [37]. Honeypots can also be categorized as low-interaction, medium-interaction, or high-interaction honeypots, depending on the scope and complexity of their design and their capabilities to interact with and adapt to the attacker activities in real-time [5][36]. The current research is focused on the development of the sector-specific honeypots for specialized systems, such as IIoT, ICS and CPS [38], complemented with advanced cyber intelligence tools capable of delivering actionable insights and decision support while minimizing the cognitive burden imposed on its users, e.g. tailored visualizations, cyber attack modelling with behavioural analytics, and deep learning techniques [37][39]. The interconnectedness of the OT and IT systems in CI allows for exploitation of a broad range of the available general IT-oriented honeypots, such as Dionaea [40] for the attack and malware detection, SSH/Telnet honeypot Cowrie [41], and Honeytrap solution [42][43]. Research on the honeypots specialized for specific CIs is also underway, but on a much smaller scale. Moreover, a detailed review of the available literature reveals that the vast majority of the reported CI honeypot experiments is conducted on the public Internet infrastructure or within university research environments, whereas only a small portion takes place in actual CIs or simulation environment thereof. Examples of the specialized sector-specific honeypots for CIs along with the available evidence about practical experiments for particular vertical sectors are presented in Table 3. The collected examples are primarily in the smart energy, water management and smart factory domains, whereas the application of the honeypots in other sectors is either less extensively addressed or predominately concerned with the IoT cybersecurity on the device level, e.g. of medical devices in healthcare or of autonomous vehicles in transportation.

The integration of the passive and active cyber defence measures and deception technology creates dynamic and adaptive cybersecurity capabilities and is representative of a synergistic approach installing a comprehensive and effective cybersecurity strategy. However, CIs continue to be particularly exposed and impacted by the increasing scale and progressive sophistication of cyber attacks. Thus, further advances are required towards an even more dynamic, intelligent, and potentially more resilient approach to the CI security. To do so, the emerging strategies are increasingly leveraging the capabilities of AI, which has found several applications to predict, identify, and neutralize the cybersecurity threats and provide a new level of threat intelligence [65]. The opportunities and challenges of introducing AI into the CI cyber defence are addressed in more detail in the next section.

Table 3: Review of the specialized sector-specific honeypots in CIs

| Honeypot | Mimicked systems/services | Demonstrated sector/CI |
|---|---|---|
| Conpot [5] (basic and extended versions) | ICS/SCADA systems simulation incl. ICS protocols and ICS/SCADA PLCs (e.g., Siemens S7-200 PLC) | Amazon AWS [44]<br>Smart grid [45]<br>Scheider Electric PowerLogic ION6200 smart meter [46]<br>Siemens S7-200 PLC simulation in an electric power plant [47] |
| IEC104 honeypot [36] | ICS/SCADA systems | Smart energy systems (electric power installations, power grids) |
| DiPot [48] | ICS systems | ICS systems generally |
| HoneyPLC [49] | ICS systems; support for a wide range of PLC models and suppliers | Amazon AWS, PLC systems generally |
| HoneyVP [50] | ICS systems | ICS systems generally |
| GridPot [51] | SCADA CPS honeynet framework | Smart energy system [52] |
| CryPLH [53] | ICS systems (S7-300 Siemens PLC) | ICS systems generally |
| HoneyPhy [54] | CPS honeypot (generic analogue thermostat and the DNP3 protocol) | ICS systems generally |
| iHoney [55] | ICS infrastructure | Water treatment plant |
| XPOT [56] | PLC honeypot (Siemens S7 314C-2) | ICS systems generally |
| TrendMicro[57] | ICS honeypots | Factory environment |
| Cyber CNI [58] | CPS honeypot (incl. PLC, SCADA, HW devices) | Industry 4.0 factory emulation |
| GasPot [59] | Veeder Root Guardian AST simulation (tank gauge) | Gas pipelined infrastructure |
| SHaPe [60] | Electric power substation IED simulation | Power systems |
| Wilhoit [61] | ICS honeypots mimicking water pressure station | Water treatment infrastructure |
| Murillo [62] | ICS plant simulation (water tanks, sensors, actuators, and PLC devices) | Water treatment infrastructure |
| MimePot [63] | ICS components and control processes simulation | Water distribution PoC implementation |
| NeuralPot [64] | ICS industrial environment simulation | ICS systems generally |

# 4 AI APPLICATIONS IN CI CYBERSECURITY

The integration of AI into the cybersecurity protocols represents a paradigm shift from a static defence to a dynamic, intelligent, and potentially more resilient approach to securing CIs. AI, and ML in particular, provide capabilities for recognizing patterns and predicting future events based on a prior experience, thereby preventing or detecting and even responding to potentially malicious activities [25]. At present, the ML algorithms are commonly used for intrusion detection, classification, prediction and prevention, automated incident response, malware detection and analysis, anomaly detection, zero-day attack prediction, as well as for advanced analytics and other intelligent applications in the threat intelligence that operate based on extraction of insights from the cybersecurity data [25][34][66][67].

According to NIST, AI has been recently recognized as a major enabler in all stages of the cybersecurity, i.e. identification, protection, detection, reaction and defence against cyber attacks. A summary of the possible uses and applications of AI/ML in different cybersecurity stages is provided in Table 4 [34][67]. We hereafter focus more thoroughly on the adoption of AI in the cyber defence mechanisms of IDPS and deception technologies as well as complementary capabilities in the detection and response operations, such as predictive intelligence. However, AI is expected to empower all stages of the cyber protection in the technological, managerial and procedural aspects, including the AI-based security by design and micro-segmentation, automated AI-based vulnerability identification and assessment, AI-assisted penetration testing, AI-based cybersecurity infrastructure investments optimization, in-depth AI forensics, etc. [65].

Table 4: AI applications in the adaptive cybersecurity

| Stage | AI/ML applications (examples) |
|---|---|
| PREDICT | Prediction of new attack vectors based on identification of trends, anomalies and potential new threats, and behaviour analytics, using historic and current datasets from various sources |
| PREVENT | Identification and blocking of potentially malicious activities, automation of security measures and configurations, e.g. through using automated network security policy management tools, advanced antivirus software, and adaptive authentication systems |
| DETECT | Enhancements of traditional (signature-based) detection capabilities through pattern recognition indicative of threats, e.g., zero-day attacks and sophisticated malware |
| RESPOND & RECOVER | Automated incident analysis for threat prioritisation and future response optimizations<br>Advanced forensics |

## 4.1 AI learning approaches

In its broadest sense, AI can be referred to as a computer system with a human-like intelligence, the capabilities of which include the ability to reason, learn, solve problems, self-correct, and interpret the natural language [30][66]. Herein, the ML technologies represent algorithms that have the ability of making decisions or predictions by learning from data instead of being explicitly programmed, by automatically creating analytical models in the concrete domain of application [25]. Depending on the learning type, the ML approaches can be classified into supervised learning, unsupervised learning, deep learning (DL), generative learning, and reinforcement learning, as well as combinations thereof, i.e. semi-supervised learning that combines supervised classification and unsupervised clustering methods, transfer learning where a pre-trained model is applied to a new classification task of a related problem, federated learning that takes place across several independent decentralised datasets, and ensemble learning that combines multiple learning algorithms either sequentially or in parallel for improving the resulting predictive performance. Presently, the supervised learning is the most frequently used approach in the cybersecurity applications. However, it suffers from two significant drawbacks. Firstly, the traditional supervised learning is capable of identifying only pre-defined features or parameters [3]. In response, other ML approaches are currently considered to overcome the feature extraction issues. DL, for example, has the ability to directly train on the original data without feature extraction and as a result it is able to detect nonlinear relationships, and is therefore specifically useful for the detection of the previously unknown attacks on CI [3]. Secondly, the supervised learning requires annotated training datasets, which must be recent, representative, high-quality and containing relevant features. Thus, a choice of the model depends on the learning properties, quality of the available cybersecurity data and on the effectiveness of the learning algorithm. Studies of the AI/ML-based intrusion detection solutions for the IoT and CPS systems [67][116][117][118] for example demonstrate varied levels of the effectiveness in using different models, i.e., decision trees, random forests and K-Nearest Neighbours perform well, while deep learning, MLP, Naïve Bayes and Logistic Regression show a lower performance, and expectedly fusion methods outperform the basic classifier models. Thus, each particular AI/ML application requires targeted studies and careful selection of the most appropriate model.

## *4.2 AI training datasets*

The availability, quality and recency of the training datasets is a crucial challenge in the AI/ML-based cyber defence [65]. A closer examination shows that most of the available datasets are outdated and thus unable to support the AI algorithms in establishing understanding of the most recent cyber attack patterns [131]. Also, the sufficiently broad real-world cybersecurity datasets for CIs are scarce, which is partially due to the privacy, regulatory and legal limitations, e.g., in healthcare, or even explicit requests from the infrastructure operators, associated with sensitive nature of such data [132]. Moreover, some CIs are subjected to further sector-specific challenges associated with the data availability. In defence, for example, rare or even hypothetical events, or out-of-bounds inputs are features rather than dataset anomalies [133]. Also, some applications require highly varied scenarios with all possible combinations of attributes that cannot be realistically captured in the original data, i.e. in an AI-assisted UAV-based visual reconnaissance application that cannot take place in all possible environments and flight conditions. To overcome the dataset scarcity problem, the few-shot learning approach is an emerging direction where a few malicious samples, such as zero-day attacks, are collected in realistic settings [69]. Another approach is the use of synthetic datasets in place of real-world datasets [134][135]. It allows to generate highly diverse or even novel datasets, fine grain control of data attributes, and automatic annotation or data labelling where necessary, which is particularly appropriate for CIs that require training datasets comprising unusual or rare events, or a broad variety of possible scenarios. The scarcity of high-quality datasets is further exacerbated also by the fact that the current AI practice predominately relies on isolated uses of individual datasets, which in addition to the availability issues stems from the poor understanding of the relationships between individual datasets. The research shows that the same limited choice of the available datasets have been used in numerous studies on the cyber attack detection mechanisms [65], i.e., datasets DARPA'98 [137], KDD'99 [138], NSL-KDD [139], and CIC-IDS2017 [140]. The AI training approaches based on a successful fusion of multiple datasets are thus an emerging research topic. Some of the well-known and used cybersecurity datasets relevant in the context of CI are summarized in Table 5.

## *4.3 AI applications in cyber defence*

AI has many applications in IDPS, deception technologies and incident response systems. In IDSP, the signature-based systems suffer from their inability to detect new attacks [69]. For example, sophisticated malware uses concealment techniques to reprogram itself after each consecutive attack iteration, thus successfully preventing the detection based on the attack signature [67]. This shortcoming is overcome in the anomaly-based IDPS that has capabilities to detect new types of attacks, but the approach consequently suffers from false positives, i.e. normal traffic patterns wrongly recognized as deviations [70]. To overcome these challenges, AI enhances the network-based IDPS systems with advanced capabilities for an automated and intelligence-driven detection of novel threats and further reduction of false alarms resulting from misclassification of a normal behaviour [66][67]. A range of the AI-based capabilities is applied for different purposes, such as anomaly detection by analysing traffic patterns and payloads, detection of encrypted threats by analysing flow properties, outlier detection by means of unsupervised clustering, etc. In the host-based IDPS, ML is used e.g., for malware classification and detection of malicious system state changes. AI is used in IDPS also for response and containment purposes, e.g., to adaptively modify policies in real time in order to block malicious traffic, prioritize and contain host endpoint attacks, or even apply an adaptive system-wide response with successive strategy refinements after each iteration [71]. Such strategies are particularly beneficial in response to zero-day attacks where an adaptive response is crucial.

Table 5: Some of the well-known cybersecurity datasets

| Dataset | Characteristics |
|---|---|
| DARPA'98 [137] | The 1998 DARPA Intrusion Detection Evaluation Dataset consisting of an off-line evaluation using network traffic and audit logs collected on a simulation network, and of real-time evaluation that took place in the AFRL network test bed identifying attack sessions in real time during normal activities. |
| KDD’99 Cup [138] | Most widely used dataset with 41 features attributes and class identification. It distinguishes between four categories of attacks: DoS, remote-to-local (R2L) intrusions, user-to-remote (U2R) intrusions, and PROB and conventional data. |
| NSL-KDD [139] | Updated KDD’99 Cup with removed redundant records to avoid skew. |
| CAIDA’07 [141] | A 2007 dataset with anonymized traces of the recorded DDoS attack traffic. |
| ISCX’12 [142] | A dataset containing network traffic generated in a real-world physical test environment, containing centralized botnets. |
| CTU-13 [143] | A botnet traffic dataset containing 13 separate malware captures, including botnet, normal, and background traffic. |
| UNSW-NB15 [144] | A dataset containing 49 features and roughly 257.700 records, which represent 9 different forms of attacks, including DoS. |
| CIC-IDS2017 [140] | An intrusion detection evaluation dataset containing benign and most common attacks, and the results of a network traffic analysis with labelled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack. |
| CIC-DDoS2019 [145] | A dataset containing common DDoS attacks. It includes also the results of the network traffic analysis with labelled flows (time stamp, source, and destination IPs, source and destination ports, protocols and attack). |
| UNSW-NB15 2015 [146] | A dataset containing an hour of traffic that represents 9 types of the major attacks – fazer, shellcode, backdoor, DoS, exploit, generic, reconnaissance, analysis, worm. |
| CSE-CIC-IDS2018 [147] | A comprehensive dataset comprising various classes of attacks. Six attack scenarios were used: bruteforce (dictionary password matching), heartbleed (SSL/TLS vulnerability), botnet, DoS, DDoS, WEB application attack. |
| WUSTL-IIOT-2018 ICS SCADA cyber security dataset [148] | A dataset prepared with a SCADA system, emulating real-world industrial systems, and focusing on reconnaissance attacks (port scanner, address scan attack, device identification attack, device identification attack – aggressive mode, exploit). |
| ADFA 2013 [149] | A dataset intended for IDS evaluation, containing server traffic for Ubuntu Linux 11.04 with Apache 2.2.17 and PHP 5.3.5, FTP, SSH, MySQL 14.14, and TikiWiki software. It includes traces of network attacks: Hydra-FTP, Hydra-SSH, Adduser, Java-Meterpreter, Meter-preter, Webshell. |
| Bot-IoT Dataset 2018 [150] | A large IoT data set containing various attack types, including DDoS, DoS and service scanning. The included types of the IoT devices are weather station, intelligent refrigerator, lamps with motion sensors, remote garage doors, and intelligent thermostat. |
| IoT-23 [151] | A datatset containing 20 malware captures executed in IoT devices, and 3 captures for benign IoT devices traffic. |

The analysis of examples of AI-based IDS for CIs reveals a varied use of the ML learning approaches, as summarized in Table 6. The predominant category of ML applications in IDS falls within the scope of the supervised learning approaches. Examples of the use of the ML classification applications include e.g., IDS for advanced metering infrastructure in smart grids using a decision tree for anomaly detection [72], Android malware detection in the context of IoT using support vector machine classification [73], anomaly-based IDS using a lightweight logic regression model for network security improvement and reduction of the human involvement in the botnet detection [74], IDS for a gas pipeline infrastructure using K-Nearest Neighbour [75], and application of fuzzy logic, neural networks, and support vector machines to improve the false-alarm problem and detection of different attack types, such as DDoS [76]. Examples of ensemble learning applications include network-based IDS with a network intrusion prediction based on random forest and support vector machine using multiple decision trees [77], a random-forest based man-in-the-middle attack detection for SCADA IoT systems [78], and IDS for IoT platform integration using a combination of a random forest and a neural network [79]. Network-based intrusion detection using an association rule-mining approach [80] and belief-rule-based association rule with the ability to handle the various types of uncertainties in IoT environments [81] are examples of the rule-based applications. A more recent category of AI-assisted IDS utilizes deep learning. The examples include SCADA IDS using a Genetically Seeded Flora feature optimization technique merged with Transformer Neural Network [82], an optimized Back-Propagation Neural Network for SCADA intrusion detection in water treatment systems [83] intrusion detection in CPS using a LSTM-based recurrent neural network [84], a smart grid IDS solution using an Autoencoder-Generative Adversarial Network for attack detection [85], identification of cyber attacks in IIoT using recurrent neural networks and artificial neural networks [86], and IDS using CNN for detection of man-in-the-middle attacks on military-grade Robot Operating System [87]. In the generative learning category, IDS solutions using auto-encored based approaches are available for malware and intrusion detection [88][89], and intrusion detection based on deep belief networks [90]. Semi-supervised learning is also employed, e.g., for fog-based attack detection using the ELM-based Semi-supervised Fuzzy C-Means method for cloud/fog-computing in IoT environments [91], a deep Feed Forward Neural Network as a classifier with a deep autoencoder for anomaly detection in SCADA systems [82], and for false data injection attacks detection in smart grids using autoencoders and generative adversarial network [92]. A model-free reinforcement learning approach was used for online cyberattack detection in smart grid [93], and inverse reinforcement learning for anomaly detection based on sequential data in safety-critical environments [94].

Examples of AI-powered IPS solutions include an IPS for unmanned aircraft system incorporating customized Threat Analysis and Risk Assessment (TARA) and dynamically applied prevention rules for the detected attacks using deep learning [95], a network-based IPS based on self-organizing incremental neural network and support vector machine for industrial applications [95], and an IPS based on game theory for Cyber Physical Systems (CPS) using reinforcement learning [96]. Other generic IPS examples include automatic incident characterization using ML to assign severity of the incident [97] and solutions for AI-based alert triage based on alert grouping in NIDPS using unsupervised clustering algorithms [98] and alert prioritization using auto-encoders [99].

Table 6: ML approaches in IDPS for specific CIs

| Learning approach | Examples of applications in specific CIs |
|---|---|
| Supervised - classification | Anomaly detection in smart grids [72]<br>Android malware detection in IoT [73]<br>IDS for gas pipeline infrastructure [75]<br>NIPS for industrial applications [95] |
| Supervised - ensemble learning | Man-in-the-middle attack detection for SCADA IoT systems [78]<br>IDS for IoT platform integration [88] |
| Deep learning | SCADA intrusion detection in water treatment systems [82]<br>Intrusion detection in CPS [84]<br>Attack detection in smart grids [85],<br>Identification of cyber attacks in IIoT [86]<br>Detection of man-in-the-middle attacks on military-grade Robot Operating System [87]<br>IPS for unmanned aircraft system [32] |
| Semi-supervised learning | Attack detection in fog computing IoT [91]<br>False data injection attacks detection in smart grids [92] |
| Reinforcement learning | Online cyber attack detection in smart grid [93]<br>Anomaly detection in safety-critical environments [94]<br>IPS for CPS systems [96] |

In the deception technology generally and the honeypots specifically, AI is employed for two principal purposes, i.e. to improve the adaptive behaviour capabilities [100], and to implement retrospective analysis. A variety of ML techniques has been proposed for adaptive behaviour capabilities in honeypots, including e.g., the use of reinforcement learning for concealment purposes and increased engagement [101] and for improving emulation capabilities [102], and more recently by using various types of Markov chains, e.g., to increase the number of commands from an attack sequence [103]. Following the attack data collection, different ML approaches are employed for analytics purposes in order to implement attack classification and modelling, with the majority of approaches relying on supervised and unsupervised learning, e.g., for DDoS identification [104], and for training dataset preparation [105]. The AI-assisted honeypot solutions include a honeynet for enhanced IoT botnet detection rate using logistics regression and cloud computing [106], and a production honeypot DeepDig [107] that uses ML for attacker profiling and adaptability. Other approaches use ML for anti-detection, i.e. reinforcement learning, and for zero-day DDoS attack prevention [103]. Honeypots Heliza [108] and RASSH [109] utilize reinforcement learning to implement interactivity during attacks, e.g., allowing and blocking commands and substituting messages. Another practical example of the deception technology besides the AI-powered honeypots includes the use of AI for generating decoy text files and deliberate manipulation of comprehensibility of real documents protected using genetic algorithm [110].

There are several other relevant AI cybersecurity application directions underway in the context of the AI-assisted cybersecurity in CIs that either incorporate or complement and extend capabilities of IDPS and deception technology. For example, the use of AI is extensively examined also for predictive intelligence in order to support capabilities to predict in advance the type, intensity and targets of an intrusion. Examples include the use of DL for network intrusion alert forecasting based on specific targets or malicious sources [111][112] and malware attack prediction based on recurrent neural networks [113]. Also, for security monitoring purposes, AI is considered to support and extend capabilities for security threats identification and investigation through data analysis and intel presentation. Some selected illustrative examples include e.g., SIEM for the detection, normalisation and correlation of cyber attacks and anomalies in smart grids [114], and a cyber attack detection system for ICS [115].

## 5 ADVANCED TOPICS AND FUTURE RESEARCH DIRECTIONS

Despite the obvious benefits, the AI-based IDPS and deception technology solutions suffer from several major challenges. The two prominent ones in the context of CI are explainability, measured in terms of the utilized model being interpretable, and robustness which represents the stability of the model against adversarial attacks. Available research demonstrates that there is no one approach that exhibits superiority in both aspects [71] and each represents a relevant emerging research direction.

## *5.1 Explainable AI*

Explainability is an essential aspect in the AI-based security applications [67], i.e., the transparency of the algorithm functioning that reveals how and based on what facts the initial conclusion was made. The current AI algorithms lack transparency in their decision-making process [65], thus suffering from lack of acceptance and trustworthiness. The issue of interpretability or black-box AI is a well-known dilemma where there is a trade-off between the prediction accuracy and explainability of the model, which is particularly challenging for the AI applications with high security requirements such as in CIs. The challenge has recently led to a new research field on explainable AI (XAI) to achieve improved transparency, e.g., devising explainable IDS systems (X-IDS) [70]. Numerous approaches are considered for the implementation of the XAI, including feature importance ranking, local explanations focused on a specific datapoint or prediction, rule-based learning implementations and the use of inherently interpretable models, i.e. linear and logistic regression, visual interpretation of the model attention, and explainability layers and user interfaces to support exploration of the model internal representations in a human-readable form. Despite its potential, however, XAI is a nascent research direction. Some of the currently known challenges include the introduction of additional complexity, reduced prediction accuracy, and human-AI interfacing for understandable explanations.

## *5.2 Adversarial AI*

Even though designed to improve the robustness of the cyber defence, the AI and ML algorithms incorporated in IDPS and deception technologies are attractive cybersecurity targets themselves, creating a whole new attack surface in CIs [119][120]. Cyber attackers target vulnerabilities of the AI applications as part of their attack strategies, e.g. offensive cyber operations using synthetic images, adversarial data manipulation etc., while at the same time they progressively leverage the AI capabilities to enhance their attack techniques and improve the defence avoidance [25][134]. This represents a new dimension in the threat landscape that the defence mechanisms must recognize, acknowledge and manage [152][153].

Table 7: Categories of adversarial attacks on AI

| Attack | Description |
|---|---|
| Poisoning attack (training) | Modifies training data to get a desired outcome at inference time. This allows the attacker to create backdoors in the model where an input with the specified trigger will result in a particular output. |
| Evasion attack (inference) | Based on adversarial inputs, the attacker elicits an incorrect response from a model. Typically, malicious inputs are indistinguishable from normal data. Evasion attacks can be targeted, where the malicious input is designed in a way to produce a specific classification, or untargeted where any incorrect classification is attempted. |
| Functional extraction (inference) | In this type of the attack, the model is iteratively queried in order to build a functionally equivalent model. Also known as a reverse engineering or model extraction attack. |
| Inversion attack (inference) | Attack that recovers sensitive information about training data, either fully or partially (properties, attributes). |
| Prompt injection attack (inference) | In this type of the attack, malicious prompts are injected into the model to cause an unintended behaviour, typically in the form of ignoring the original instructions and instead following the adversary instructions. The attack applies to the LMM models with complex inputs. |
| Cyber attack | Various cyber attacks targeting the AI infrastructure and its components instead of the model itself, e.g., API keys, data servers etc. |

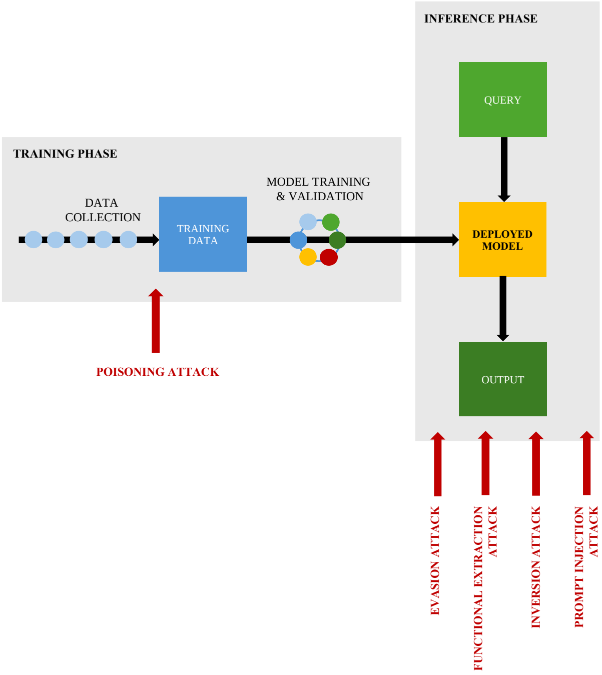

Figure 6. Types of adversarial the AI attacks in training and inception stages.

Fundamentally, the adversarial attacks pursue three general types of objectives, i.e., reduced availability, integrity violation and compromised privacy [154], with several different attack types (see Table 7). Attacks can be classified into two major groups according to the time of their occurrence, i.e. adversarial attacks during the AI training phase, and attacks in the inception phase, i.e., when an already trained model is tested, verified and deployed (see Figure 6). The adversarial attacks in the training phase are categorized as poisoning attacks, in which false or misleading data is injected into the training data, causing the model to learn incorrect patterns or behaviours. For example, to induce unpredictable, incorrect or even false predictions, the attacks utilize data perturbation, i.e., they slightly modify the input data, which leads to faulty model predictions. This can result in severe consequences in the context of services in CI. A model can thus be falsely trained to interpret an image of a tank as a civilian vehicle [133]. Poisoning was demonstrated in [121] for the case of an AI-based drugs dosage prescription solution where malicious insertion of 8% of the erroneous data to the AI algorithm caused a 75% change in the prescribed drug dosage in 50% of the patients. Poisoning attacks are specifically problematic because they cannot be detected until the trigger is activated, and even then, the deviations from the expected behaviour can be minimal yet sufficient to cause damage.

The evasion attacks on the other hand take place in the inference phase, where the model inputs are manipulated (infected or falsified), inducing misclassification of an already trained model. In this type of the attacks, an incorrect response is elicited from the deployed model using adversarial inputs, which are typically indistinguishable from normal input data, causing an incorrect classification. The well-known examples from the transportation and autonomous driving domain include a successful misclassification of stop signs using altered images of the stop sign with stickers resembling graffiti [122], and autonomous vehicle camera image perturbations inserting road markings causing the vehicle to steer to the reverse traffic lane [156].

Table 8: Adversarial AI attack scenarios in CI

| **Type of infrastructure** | **Potential adversarial AI attack scenarios** |
|---|---|
| Smart energy systems | Perturbation of the input data to the AI algorithms used in energy management systems (e.g., falsified weather data) inducing an incorrect demand prediction, causing instabilities or even blackouts. |
| Transportation infrastructure | Input image perturbations inducing an incorrect classification causing a wrong interpretation of road signs and signatures from the vehicle camera and other sensor readings, resulting in dangerous driving behaviour (e.g., using wrong lanes or falsely interpreting road signs). Tampering of the traffic conditions data (traffic jams, accidents) to induce incorrect congestion predictions leading to suboptimal rerouting decisions resulting in congestions and gridlocks. |
| Water treatment infrastructure | Tampering with the input sensor data used by AI algorithms managing chemical dosing, causing public health hazards because of improper treatment levels. |
| Healthcare | Exposure of sensitive health data through stealing and reconstruction of the data from AI models deployed in smart health systems with access to patient records. Input image perturbation used by AI-based diagnostics tools causing misdiagnosed patients and incorrect treatment decisions. |
| Financial services | Injection of crafted transaction data that mimics normal user behaviour to AI-based fraud detection systems causing approval of malicious transactions with direct financial consequences. |

The third category that also takes place in the inference phase are privacy-related attacks, where an attacker reconstructs the model or hijacks the data the model was trained on by analysing the black-box model, potentially causing an exposure of confidential data in the training dataset or the model itself [157]. Health- and medical-related CIs are prone to such attacks, possibly leading to an exposure of highly sensitive patient data [152]. The mechanics of individual adversarial AI attacks are represented in Figure 7.

An important observation based on the research of the relevant literature and other public sources is that there are currently no widely reported adversarial AI attacks on real CIs that have been confirmed. The documented cases are largely experimental, with demonstrations primarily taking place in experimental environments. Potential attack scenarios in specific CIs are summarized in Table 8. However, the lack of reported real-world incidents does not diminish the concern given the increasing integration of the AI systems into CI and the field requires a further research attention. Protection against the adversarial AI is essential. It includes the application of appropriate pre-emptive measures suitable for the CIs and AI applications therein. The current approaches primarily constitute detection methods, such as real-time input monitoring, and robustness methods that comprise e.g. resistant training approaches and improved model rigidity to adversarial attacks. Adversarial training is employed where the training data incorporates examples of attack methods. The applicability of the approach, however, is highly dependent on the models it can target e.g., in IDPS the tree-based algorithms are subject to such adversarial training technique in order to improve their robustness [123], and on the realism and recency of the modelled attacks in a particular CI environment [124]. The Adversarial Threat Landscape for Artificial-Intelligence Systems knowledge base provided by MITRE (MITRE ATLAS) is a relevant resource in this respect providing adversary tactics and techniques against the AI-enabled systems based on real-world attack observations and realistic demonstrations from AI red teams and security groups [158]. Furthermore, once in progress, cyber attacks on AI are very difficult to detect because of the explainability problem, where further research is also required.

## *5.3 Emerging governance*

To address the AI explainability and robustness problems and manage the deceptive consequences of cyber attacks on AI, numerous standardization and certification frameworks are underway globally. This includes ISO/IEC NP TR 24029-1 addressing the assessment of the robustness of the neural networks [125] and the ISO/IEC TR 5469:2024 on functional safety and AI systems [126], the OECD AI principles [127], the G20 AI Principles, the World Economic Forum ten AI Government Procurement Guidelines, and the UNESCO Recommendation on the ethics of AI [128]. In the EU, the principal frameworks include the EU Cybersecurity Act establishing the framework for the cybersecurity standards and certification procedures for the digital technologies and services available in the EU, which among others mandates the EU Agency for Network and Information Security (ENISA) to guide the finalization of the national certification programmes in the EU member states [129], and the recently enforced EU AI Act [130] which provides categorization of the AI applications according to their level of risk, establishes a regulatory frameworks similar to the GDPR for the data protection, and requires high-risk AI applications to undergo rigorous audits. Several other regional and national regulatory frameworks are also underway addressing the security, transparency, trustworthiness and ethics of AI. This landscape is expected to undergo further developments in the near future in order to arrive at a more robust, transparent, ethical, trustworthy and acceptable AI capable of serving the domains in question and the society as a whole.

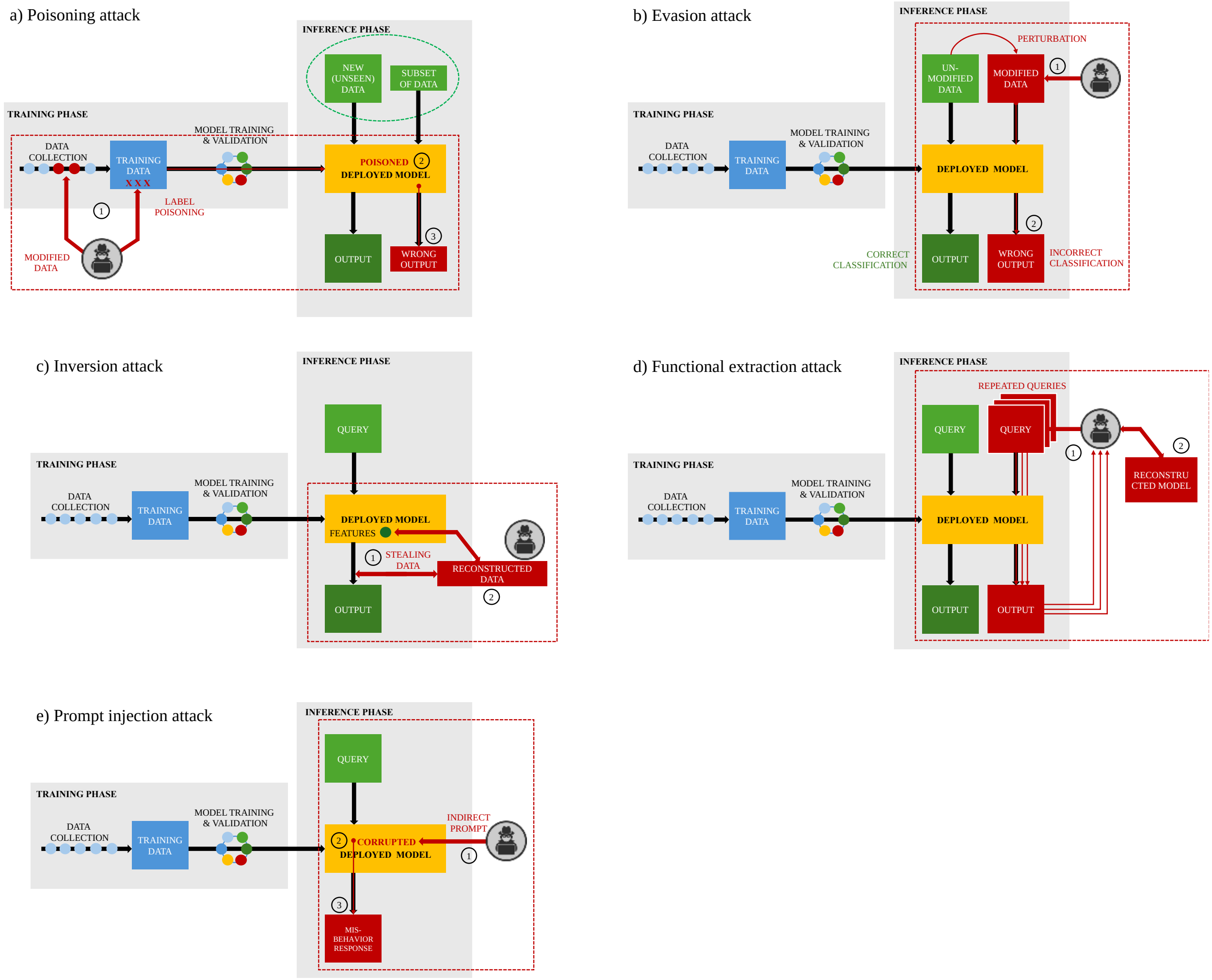

Figure 7. The mechanics of the adversarial AI attacks

## *5.4 Other research topics*

There are several further trends emerging in the scope of the research on advanced cyber defence and the use of AI within the CI purview. Convergence of the blockchain technologies and AI is one such direction, aimed to establish a decentralized and verifiable approach to the cybersecurity with tamper-evident and immutability controls. Another comprehensive topic is quantum computing, promising to deliver a new generation of encryption capabilities through quantum-resistant cryptography as well as other emerging areas of application, such as the use of the quantum algorithms for an enhanced AI-based threat detection and response.

Ethical AI is emerging as a prominent and challenging research field, currently faced with several crucial considerations [7]. This includes discriminatory AI through training data bias inheritance that raises ethical concerns associated with fairness and justice, and the ethics of the AI applications in warfare and offensive cyber operations generally which, if not appropriately controlled, can unintendedly lead to devastating consequences, including collateral damage or conflict escalations. The research community is further concerned also with the robustness of the AI and the possibility of being repurposed for malicious activities, such as AI weaponization. There are also ethical concerns and objections with respect to the privacy and informed use of AI in the cybersecurity, as well as several other challenges associated with the socio-economic and legal impacts, including accountability in decision-making and AI-induced unemployment. In conclusion, the identification and discussion of the ethical issues and value conflicts involved in cybersecurity in relation to CI and the adoption of the AI applications are fundamentally important in assisting further guidance.

## 6 GUIDANCE ON THE ADOPTION OF AI FOR THE CI CYBER DEFENCE

The preceding sections demonstrate the extensive opportunities of AI in providing enhancements to the cyber defence capabilities. However, the implementation of reliable, resilient and trustworthy AI applications into CI is in a nascent phase lacking sufficient best-practice examples and guidance about the most appropriate approaches. Thus, a set of guidelines is drafted hereafter to inform and guide an integrated and strategic approach to an AI-powered secure CI. The guidelines draw from key findings in relevant scientific literature, and from available advisory resources on securing IT and OT in CI [159][160] and on the introduction and securing AI in such environments [161].

In the process of securing the introduction of an AI system into CI, four fundamental perspectives can be distinguished, i.e. conducting a thorough risk assessment and alignment with the general security practice of CI and vulnerability management procedures, securing and hardening the IT/OT environment the AI system is introduced into, adoption of specific measures and technologies to install secure and hardened AI specifically, and continuous maintenance of appropriate knowledge capacities. Each of the identified perspectives is an essential element of the CI overall security posture, entailing a range of the possible approaches requiring a further consideration and validation through best practices and experience to be gained in the next stages of the AI-based CI evolution.

Risk assessment – The risk assessment is an essential initial phase of the AI introduction, including the definition of the AI use cases and identification of the vulnerabilities and impacts, followed by a risk prioritization in the alignment with the CI risks management strategy.

Securing the CI environment – AI is considered an IT system and will thus be deployed in the IT parts of CI. The security and resilience must be planned and installed in both the IT and OT parts of CI and security measures must be in place for the deployment of the AI specifically as well as for general robustness of the environment. This includes hardening through the adoption of the security-by-design and Zero Trust principles, strict oversight over remote access and Internet connections, using also publicly available resources such as Shodan to discover Internet-accessible OT devices, and monitoring, management and possibly removal of any non-vital remote access. A secure network architecture must be implemented using a combination of the approaches and techniques, i.e. demilitarized zones (DMZs), firewalls, sandboxing, and network segmentation to protect the IT and OT parts specifically from a direct exposure to the Internet wherever applicable. Secure SW management should be implemented, including SW updates and patches to minimize the exposure through the known vulnerabilities. Network hardening must be addressed specifically, by securing remote access through virtual private network, encryption and multifactor authentication, traffic filtering and the use of geo-blocking where appropriate etc. Cyber defence capacities should be installed by using the IDPS capabilities and other targeted cyber protection and defence solutions, both in the IT and OT parts of IC. These approaches, measures and technologies apply to CI irrespective of the AI introduction.

Securing AI – Dedicated capabilities and measures for the AI system are specifically required, during preparation and acquisition, deployment and operation. In preparation of the AI deployment, supply chain security must be instilled for any part of the AI system provided externally. Also, secure software maintenance practices should be adopted, such as the use of cryptographic mechanisms and digital signatures for the AI system validation, secure SW storage and versioning. Prior and during the deployment, hardening of the boundaries between the IT environment and the AI system should be installed along with an implementation of access control and instalment of privileged access only, and identifying and securing data sources and sensitive AI data using encryption at rest and secure communication protocols in transit. Testing and validation of externally acquired AI models should be conducted in a secure development environment prior to its deployment into production. Testing of the AI system for the robustness, accuracy and potential vulnerabilities prior to deployment as well as after any subsequent modifications should also be implemented. Advanced measures, such as adversarial training, should also be considered. If an AI system exposes application programming interfaces, they should be secured through authentication and authorization and the use of secure protocols. Penetration testing and audits should be considered by external experts to detect any vulnerabilities that have not been detected internally. Once the system is deployed, strong access control should be employed for access to the AI model to prevent any tampering, e.g., by using a role- or attribute-based access control. The access protection must be specifically focused on the protection of model weights. An automated anomaly detection, analysis and response capabilities for the AI system should also be considered to identify and react to any possible cybersecurity incident. This entails active an AI behaviour monitoring to detect unauthorized changes and access and inference attempts, as well as any further security posture updates as new threats emerge. Finally, capabilities for a manual inspection and mitigation should be in place in order to prevent overreliance on the AI system and instil its augmentation instead.

Awareness, training and knowledge capacity building – User training and awareness as well as advanced knowledge building should be instilled and maintained at all times to minimize the human error and support a secure and trustworthy AI operation maintenance in CI. This includes knowledge about the security principles generally as well as about specific topics, such as secure password management, secure data handling, phishing prevention, etc. Advanced knowledge should be maintained and improved on a continuous basis in the essential AI-related areas, such as awareness about the current and emerging threat landscape, explainability, ethics and adversarial robustness.

The presented guidelines are drafted in order to provide a general orientation in crucial aspects of the use of AI for the cybersecurity in the CI environments. However, in order to fully exploit the potential of the AI-enhanced cyber defence, the presented guidance must be further complemented with knowledge, capacities and procedures to install a continuous evolution in all relevant topics, existent and emerging guidance on the security, standardization and certification of CI, as well as sustained collaboration and communication with the relevant cybersecurity advisories and the research community to understand and instil the latest technological advancements and practices and thus maintain a cutting-edge posture of the AI-enabled CI cyber defence.

## 7 Conclusion

The AI adoption marks a significant advancement towards enhancing the cyber defence capabilities in response to the increasingly sophisticated threats in CI. This paper shows that the dynamic, adaptive, and intelligence-driven AI applications not only fortify the defence mechanisms but also propel the development of more resilient critical infrastructures. However, the discussed approaches and capabilities introduce complexities and considerations, particularly in terms of the resilience, robustness, explainability, ethical use, and the need for robust AI governance frameworks. In an attempt to harness its potential while prioritizing safety and trustworthiness, the effectiveness of the AI systems for the cyber defence must be continuously assessed against the emerging CI threats and vulnerabilities, carefully considering also the aspects of dual use, and a balance should be sought between the advancements in innovation and their ethical application into practice. Last but not least, emphasizing collaboration across sectors and a continuous and thoughtful consideration of best practices as they emerge will be of the utmost importance in establishing future pathways for this particular sector.

## Acknowledgements

The research received support from the Slovenian Research and Innovation Agency, the Ministry of Defence of Slovenia, and the Government Information Security Office of Slovenia, through the research project “Cyber security of defence systems and critical infrastructures” (grant nr. V2-2378), and the research program “Decentralized Solutions for the Digitalization of Industry and Smart Cities and Communities” (grant nr. P2-0425).

## Literature

[1] Riggs, H., Tufail, S., Parvez, I., Tariq, M., Khan, M. A., Amir, A., ... & Sarwat, A. I. (2023). Impact, vulnerabilities, and mitigation strategies for cyber-secure critical infrastructure. *Sensors*, 23(8), 4060.

[2] Stevens, T. (2020). Knowledge in the grey zone: AI and cybersecurity. *Digital War*, 1(1), 164-170.

[3] Sakhnini, J., Karimipour, H., Dehghantanha, A., & Parizi, R. M. (2020). AI and security of critical infrastructure. *Handbook of Big Data Privacy*, 7-36.

[4] OT Cybersecurity: The Ultimate Guide. (2022, April 25). *Industrial Defencer*. Retrieved February 12, 2024, from https://www.industrialdefender.com/blog/ot-cybersecurity-the-ultimate-guide.

[5] Mesbah, M., Elsayed, M. S., Jurcut, A. D., & Azer, M. (2023). Analysis of ICS and SCADA Systems Attacks Using Honeypots. *Future Internet*, 15(7), 241.

[6] Lella, I., Ciobanu, C., Tsekmezoglou, E., Theocharidou, M., Magonara, E., Malatras, A., & Svetozarov Naydenov, R. (2023). ENISA threat landscape 2023: July 2022 to June 2023.

[7] Viganò, E., Loi, M., & Yaghmaei, E. (2020). Cybersecurity of critical infrastructure. The Ethics of Cybersecurity, 157-177.

[8] Significant Cyber Incidents. (2023, May 7). *Center for Strategic & International Studies (CSIS)*. Retrieved March 23, 2024, from https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents.

[9] *ICS Matrix*. (n.d.). MITRE ATT&CK. Retrieved March 27, 2024, from https://attack.mitre.org/matrices/ics/.

[10] *ELITEWOLF*. (n.d.). Github. Retrieved March 27, 2024, from https://github.com/nsacyber/ELITEWOLF.

[11] Easterly. (2023, May 7). The Attack on Colonial Pipeline: What We’ve Learned & What We’ve Done Over the Past Two Years. *Cybersecurity and Infrastructure Security Agency (CISA)*. Retrieved March 23, 2024, from https://www.cisa.gov/news-events/news/attack-colonial-pipeline-what-weve-learned-what-weve-done-over-past-two-years.

[12] Cyber-Attack Against Ukrainian Critical Infrastructure. (2021, July 20). *Cybersecurity and Infrastructure Security Agency (CISA)*. Retrieved March 23, 2024, from https://www.cisa.gov/news-events/ics-alerts/ir-alert-h-16-056-01.

[13] Falliere, Falliere, O. Murchu, & Chien. (2010, November). W32.Stuxnet Dossier. In *Symantec Security Response*. Retrieved March 19, 2024, from https://www.wired.com/images_blogs/threatlevel/2010/11/w32_stuxnet_dossier.pdf.

[14] ICS Focused Malware. (2021, July 20). *Cybersecurity and Infrastructure Security Agency (CISA)*. Retrieved March 23, 2024, from https://www.cisa.gov/news-events/ics-advisories/icsa-14-178-01.

[15] CRASHOVERRIDE Malware. (2017, July 25). *Cybersecurity and Infrastructure Security Agency (CISA)*. Retrieved March 23, 2024, from https://www.cisa.gov/news-events/ics-alerts/ics-alert-17-206-01.

[16] TRITON Malware Targeting Safety Controllers. (2017, December 22). In *National Cyber Security Center*. Retrieved March 19, 2024, from https://www.ncsc.gov.uk/information/triton-malware-targeting-safety-controllers.

[17] Advanced Persistent Threat Compromise of Government Agencies, Critical Infrastructure, and Private Sector Organizations. (2021, April 15). *Cybersecurity and Infrastructure Security Agency (CISA)*. Retrieved March 23, 2024, from https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-352a.

[18] APT10: sophisticated multi-layered loader Ecipekac discovered in A41APT campaign. (2021, March 30). *Securelist by Kaspersky*. Retrieved March 23, 2024, from https://securelist.com/apt10-sophisticated-multi-layered-loader-ecipekac-discovered-in-a41apt-campaign/101519/

[19] Aver. (2021, January 13). An employee, fired. *Kaspersky Daily*. Retrieved March 23, 2024, from https://www.kaspersky.com/blog/fired-insider/38381/.

[20] Pryimenko. (2023, September 13). 7 Examples of Real-Life Data Breaches Caused by Insider Threats . *Ekran*. Retrieved March 23, 2024, from https://www.techtarget.com/searchsecurity/tip/The-biggest-ransomware-attacks-in-history.

[21] K. Pratt. (2023, September 13). The 10 biggest ransomware attacks in history. *TechTarget*. Retrieved March 23, 2024, from https://www.techtarget.com/searchsecurity/tip/The-biggest-ransomware-attacks-in-history.

[22] Case study: VIASAT. (2022, June). *CyberPeace Institute*. Retrieved March 23, 2024, from https://cyberconflicts.cyberpeaceinstitute.org/law-and-policy/cases/viasat.

[23] Winberg. (2021, June 2). Analysis of top 11 cyber attacks on critical infrastructure. Retrieved March 23, 2024, from https://www.england.nhs.uk/long-read/case-study-wannacry-attack/.

[24] NHS England business continuity management toolkit case study: WannaCry attack. (2023, April 21). In *NHS England*. Retrieved April 1, 2024, from https://www.england.nhs.uk/long-read/case-study-wannacry-attack/.

[25] Sarker, I. H. (2023). Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects. *Annals of Data Science*, 10(6), 1473-1498.

[26] National Institute of Standards and Technology. (2024, February 26). *The NIST Cybersecurity Framework (CSF) 2.0* (NIST CSWP 29). Retrieved April 18, 2024, from https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf

[27] Ntingi, N., Duvenage, P., du Toit, J., & von Solms, S. (2022, June). Effective Cyber Threat Hunting: Where and how does it fit?. In *European Conference on Cyber Warfare and Security (Vol. 21, No. 1, pp. 206-213).*

[28] Vargas Martínez, C., & Vogel-Heuser, B. (2018). Towards industrial intrusion prevention systems: A concept and implementation for reactive protection. *Applied Sciences*, 8(12), 2460.

[29] Alanazi, S. S., Alanazi, A. A. (2022). Knowing the Unknown: The Hunting Loop. *Int. j. adv. appl. sci*., 1(9), 8-19.

[30] Santoso, F., & Finn, A. (2023). An In-Depth Examination of Artificial Intelligence-Enhanced Cybersecurity in Robotics, Autonomous Systems, and Critical Infrastructures. *IEEE Transactions on Services Computing*.

[31] Al-Hamadi, H., Chen, R., Wang, D. C., & Almashan, M. (2020). Attack and defense strategies for intrusion detection in autonomous distributed IoT systems. *IEEE Access*, 8, 168994-169009.

[32] Schermann, R., Ammerer, T., Stelzer, P., Macher, G., & Steger, C. (2023, October). Risk-Aware Intrusion Detection and Prevention System for Automated UAS. In *2023 IEEE 34th International Symposium on Software Reliability Engineering Workshops (ISSREW)* (pp. 148-153). IEEE.

[33] Mohan, S. N., Ravikumar, G., & Govindarasu, M. (2020, October). Distributed intrusion detection system using semantic-based rules for SCADA in smart grid. In *2020 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)* (pp. 1-5). IEEE.

[34] Callegari, C., Forti, A. C., D'Amore, G., de la Hoz, E., Santamaria, D. E., García-Ferreira, I., & López-Civera, G. (2016, July). An Architecture for Securing Communications in Critical Infrastructure. In *DCNET* (pp. 111-120).

[35] Yarali, A., & Sahawneh, F. G. (2019, December). Deception: Technologies and strategy for cybersecurity. In *2019 IEEE International Conference on Smart Cloud (SmartCloud)* (pp. 110-120). IEEE.

[36] Grigoriou, E., Liatifis, A., Grammatikis, P. R., Lagkas, T., Moscholios, I., Markakis, E., & Sarigiannidis, P. (2022, July). Protecting IEC 60870-5-104 ICS/SCADA systems with honeypots. In *2022 IEEE international conference on cyber security and resilience (CSR)* (pp. 345-350). IEEE.

[37] Husák, M., Jirsík, T., & Yang, S. J. (2020, August). SoK: contemporary issues and challenges to enable cyber situational awareness for network security. In *Proceedings of the 15th International Conference on Availability, Reliability and Security* (pp. 1-10).

[38] Mashima, D. (2022). Mitre att&ck based evaluation on in-network deception technology for modernized electrical substation systems. *Sustainability*, 14(3), 1256.

[39] Kiennert, C., Ismail, Z., Debar, H., & Leneutre, J. (2018). A survey on game-theoretic approaches for intrusion detection and response optimization. *ACM Computing Surveys (CSUR)*, 51(5), 1-31.

[40] *Dionaea*. (n.d.). Github. Retrieved March 27, 2024, from https://github.com/DinoTools/dionaea).

[41] *Cowrie*. (n.d.). Github. Retrieved March 27, 2024, from https://github.com/cowrie/cowrie.

[42] *Advanced Honeypot framework*. (n.d.). Github. Retrieved March 27, 2024, from https://github.com/honeytrap/honeytrap)

[43] Kuskov, Kuzin, Shmelev, Makrushin, & Grachev. (2017, June 19). Honeypots and the Internet of Things. *Securelist By Kaspersky*. Retrieved April 18, 2024, from https://securelist.com/honeypots-and-the-internet-of-things/78751/.

[44] Jicha, A., Patton, M., & Chen, H. (2016, September). SCADA honeypots: An in-depth analysis of Conpot. In *2016 IEEE conference on intelligence and security informatics (ISI)* (pp. 196-198). IEEE.

[45] Pliatsios, D., Sarigiannidis, P., Liatifis, T., Rompolos, K., & Siniosoglou, I. (2019, September). A novel and interactive industrial control system honeypot for critical smart grid infrastructure. In *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)* (pp. 1-6). IEEE.

[46] Scott, C., & Carbone, R. (2014). Designing and Implementing a Honeypot for a SCADA Network. *SANS Institute Reading Room, 39.*

[47] Hyun, D. (2018). Collecting cyberattack data for industrial control systems using honeypots (Doctoral dissertation, Monterey, California: Naval Postgraduate School).

[48] Cao, J., Li, W., Li, J., & Li, B. (2018). Dipot: A distributed industrial honeypot system. In *Smart Computing and Communication: Second International Conference, SmartCom 2017, Shenzhen, China, December 10-12, 2017, Proceedings 2* (pp. 300-309). Springer International Publishing.

[49] López-Morales, E., Rubio-Medrano, C., Doupé, A., Shoshitaishvili, Y., Wang, R., Bao, T., & Ahn, G. J. (2020, October). Honeyplc: A next-generation honeypot for industrial control systems. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (pp. 279-291).

[50] You, J., Lv, S., Sun, Y., Wen, H., & Sun, L. (2021, June). Honeyvp: A cost-effective hybrid honeypot architecture for industrial control systems. In *ICC 2021-IEEE International Conference on Communications* (pp. 1-6). IEEE.

[51] Dutta, N., Jadav, N., Dutiya, N., & Joshi, D. (2020). Using honeypots for ICS threats evaluation. *Recent developments on industrial control systems resilience*, 175-196.

[52] Kendrick, M. M., & Rucker, Z. A. (2019). Energy-grid threat analysis using honeypots (Doctoral dissertation, Monterey, CA; Naval Postgraduate School).

[53] Buza, D. I., Juhász, F., Miru, G., Félegyházi, M., & Holczer, T. (2014). CryPLH: Protecting smart energy systems from targeted attacks with a PLC honeypot. In *Smart Grid Security: Second International Workshop, SmartGridSec 2014, Munich, Germany,*

*February 26, 2014, Revised Selected Papers 2* (pp. 181-192). Springer International Publishing.

[54] Litchfield, S., Formby, D., Rogers, J., Meliopoulos, S., & Beyah, R. (2016). Rethinking the honeypot for cyber-physical systems. *IEEE Internet Computing*, 20(5), 9-17.

[55] Navarro, O., Balbastre, S. A. J., & Beyer, S. (2019). Gathering Intelligence Through Realistic Industrial Control System Honeypots: A Real-World Industrial Experience Report. In *Critical Information Infrastructures Security: 13th International Conference, CRITIS 2018, Kaunas, Lithuania, September 24-26, 2018, Revised Selected Papers 13* (pp. 143-153). Springer International Publishing.

[56] Kato, S., Tanabe, R., Yoshioka, K., & Matsumoto, T. (2021, May). Adaptive observation of emerging cyber attacks targeting various IoT devices. In *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)* (pp. 143-151). IEEE.

[57] Hilt, S., Maggi, F., Perine, C., Remorin, L., Rösler, M., & Vosseler, R. (2020). Caught in the act: Running a realistic factory honeypot to capture real threats. *Trend Micro Research*.

[58] Pahl, M. O., Kabil, A., Bourget, E., Gay, M., & Brun, P. E. (2020). A mixed-interaction critical infrastructure honeypot. *European Cyber Week CAESAR*.

[59] *GasPot Released at Blackhat 2015*. (n.d.). Github. Retrieved March 27, 2024, from https://github.com/sjhilt/GasPot

[60] Kołtyś, K., & Gajewski, R. (2015). Shape: A honeypot for electric power substation. *Journal of telecommunications and information technology*, (4), 37-43.

[61] Wilhoit, K. (2013). Who's really attacking your ICS equipment?. *Trend Micro*, *10*.

[62] Murillo, A. F., Cómbita, L. F., Gonzalez, A. C., Rueda, S., Cardenas, A. A., & Quijano, N. (2018, December). A virtual environment for industrial control systems: A nonlinear use-case in attack detection, identification, and response. In *Proceedings of the 4th Annual Industrial Control System Security Workshop* (pp. 25-32).

[63] Bernieri, G., Conti, M., & Pascucci, F. (2019, October). Mimepot: a model-based honeypot for industrial control networks. In *2019 IEEE International conference on systems, man and cybernetics (SMC)* (pp. 433-438). IEEE.

[64] Siniosoglou, I., Efstathopoulos, G., Pliatsios, D., Moscholios, I. D., Sarigiannidis, A., Sakellari, G., ... & Sarigiannidis, P. (2020, July). NeuralPot: An industrial honeypot implementation based on deep neural networks. In *2020 IEEE Symposium on Computers and Communications (ISCC)* (pp. 1-7). IEEE.

[65] Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 101804.

[66] Petrovic, N., & Jovanovic, A. (2023). Towards Resilient Cyber Infrastructure: Optimizing Protection Strategies with AI and Machine Learning in Cybersecurity Paradigms. *International Journal of Information and Cybersecurity*, 7(12), 44-60.

[67] Schmitt, M. (2023). Securing the Digital World: Protecting smart infrastructures and digital industries with Artificial Intelligence (AI)-enabled malware and intrusion detection. *Journal of Industrial Information Integration*, 36, 100520.

[68] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access*, 10, 112392-112415.

[69] Iliyasu, A. S., Abdurrahman, U. A., & Zheng, L. (2022). Few-shot network intrusion detection using discriminative representation learning with supervised autoencoder. *Applied Sciences*, 12(5), 2351.

[70] Al-Janabi, M., Ismail, M. A., & Ali, A. H. (2021). Intrusion Detection Systems, Issues, Challenges, and Needs. *Int. J. Comput. Intell. Syst.*, 14(1), 560-571.

[71] Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence*, 1(12), 557-560.

[72] Radoglou-Grammatikis, P. I., & Sarigiannidis, P. G. (2018, October). An anomaly-based intrusion detection system for the smart grid based on cart decision tree. In *2018 global information infrastructure and networking symposium (GIIS)* (pp. 1-5). IEEE.

[73] Ham, H. S., Kim, H. H., Kim, M. S., & Choi, M. J. (2014). Linear SVM-based android malware detection for reliable IoT services. *Journal of Applied Mathematics*, 2014.

[74] Bapat, R., Mandya, A., Liu, X., Abraham, B., Brown, D. E., Kang, H., & Veeraraghavan, M. (2018, April). Identifying malicious botnet traffic using logistic regression. In *2018 systems and information engineering design symposium (SIEDS)* (pp. 266-271). IEEE.

[75] Dakheel, A. H., Dakheel, A. H., & Abbas, H. H. (2019). Intrusion detection system in gas-pipeline industry using machine learning. *Periodicals of Engineering and Natural Sciences*, 7(3), 1030-1040.

[76] Markevych, M., & Dawson, M. (2023, July). A review of enhancing intrusion detection systems for cybersecurity using artificial intelligence (ai). *In International conference Knowledge-based Organization (Vol. 29, No. 3, pp. 30-37)*.

[77] Chang, Y., Li, W., & Yang, Z. (2017, July). Network intrusion detection based on random forest and support vector machine. In *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)* (Vol. 1, pp. 635-638). IEEE.

[78] Mughaid, A., AlJamal, M., Issa, A. A., AlJamal, M., Alquran, R., AlZu'bi, S., & Abutabanjeh, A. A. (2023, October). Enhancing cybersecurity in scada iot systems: A novel machine learning-based approach for man-in-the-middle attack detection. In *2023 3rd Intelligent Cybersecurity Conference (ICSC)* (pp. 74-79). IEEE.

[79] Mohamed, T., Otsuka, T., & Ito, T. (2018). Towards machine learning based IoT intrusion detection service. In *Recent Trends and Future Technology in Applied Intelligence: 31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2018, Montreal, QC, Canada, June 25-28, 2018, Proceedings 31* (pp. 580-585). Springer International Publishing.

[80] Sellappan, D., & Srinivasan, R. (2020). Association rule-mining-based intrusion detection system with entropy-based feature selection: Intrusion detection system. In *Handbook of Research on Intelligent Data Processing and Information Security Systems* (pp. 1-24). IGI Global.

[81] Ul Islam, R., Hossain, M. S., & Andersson, K. (2018). A novel anomaly detection algorithm for sensor data under uncertainty. *Soft Computing*, 22(5), 1623-1639.

[82] Diaba, S. Y., Anafo, T., Tetteh, L. A., Oyibo, M. A., Alola, A. A., Shafie-Khah, M., & Elmusrati, M. (2023). SCADA securing system using deep learning to prevent cyber infiltration. *Neural Networks*, 165, 321-332.

[83] Alimi, O. A., Ouahada, K., Abu-Mahfouz, A. M., Rimer, S., & Alimi, K. O. A. (2021). A review of research works on supervised learning algorithms for SCADA intrusion detection and classification. *Sustainability*, 13(17), 9597.

[84] Abdullahi, M., Alhussian, H., Aziz, N., Abdulkadir, S. J., & Baashar, Y. (2022, August). Deep Learning Model for Cybersecurity Attack Detection in Cyber-Physical Systems. In *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA)* (pp. 1-5). IEEE.

[85] Siniosoglou, I., Radoglou-Grammatikis, P., Efstathopoulos, G., Fouliras, P., & Sarigiannidis, P. (2021). A unified deep learning anomaly detection and classification approach for smart grid environments. *IEEE Transactions on Network and Service Management*, 18(2), 1137-1151.

[86] Rajesh, M., Vincent, R., Kathuria, S., Jamalpur, B., Durgam, T., & Jaiswal, T. (2023, December). Design of Deep Learning Models for the Identification of Harmful Attack Activities in the Industrial Internet of Things (IIOT). In *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (Vol. 10, pp. 1765-1771). IEEE.

[87] Santoso, F., & Finn, A. (2023). Trusted Operations of a Military Ground Robot in the Face of Man-in-the-Middle Cyber-Attacks Using Deep Learning Convolutional Neural Networks: Real-Time

Experimental Outcomes. *IEEE Transactions on Dependable and Secure Computing*.

[88] Wang, W., Zhao, M., & Wang, J. (2019). Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing*, 10, 3035-3043.

[89] Yan, B., & Han, G. (2018). Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system. *IEEE Access*, 6, 41238-41248.

[90] Wei, P., Li, Y., Zhang, Z., Hu, T., Li, Z., & Liu, D. (2019). An optimization method for intrusion detection classification model based on deep belief network. *IEEE Access*, 7, 87593-87605.

[91] Rathore, S., & Park, J. H. (2018). Semi-supervised learning based distributed attack detection framework for IoT. *Applied Soft Computing*, 72, 79-89.

[92] Zhang, Y., Wang, J., & Chen, B. (2020). Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach. *IEEE Transactions on Smart Grid*, 12(1), 623-634.

[93] Kurt, M. N., Ogundijo, O., Li, C., & Wang, X. (2018). Online cyber-attack detection in smart grid: A reinforcement learning approach. *IEEE Transactions on Smart Grid*, 10(5), 5174-5185.

[94] Oh, M. H., & Iyengar, G. (2019, July). Sequential anomaly detection using inverse reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & data mining* (pp. 1480-1490).

[95] Constantinides, C., Shiaeles, S., Ghita, B., & Kolokotronis, N. (2019, June). A novel online incremental learning intrusion prevention system. In *2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)* (pp. 1-6). IEEE.

[96] Panfili, M., Giuseppi, A., Fiaschetti, A., Al-Jibreen, H. B., Pietrabissa, A., & Priscoli, F. D. (2018, June). A game-theoretical approach to cyber-security of critical infrastructures based on multi-agent reinforcement learning. In *2018 26th Mediterranean Conference on Control and Automation (MED)* (pp. 460-465). IEEE.

[97] DeCastro-García, N., Muñoz Castañeda, Á. L., & Fernández-Rodríguez, M. (2020). Machine learning for automatic assignment of the severity of cybersecurity events. *Computational and Mathematical Methods*, 2(1), e1072.

[98] Manganiello, F., Marchetti, M., & Colajanni, M. (2011). Multistep attack detection and alert correlation in intrusion detection systems. In *Information Security and Assurance: International Conference, ISA 2011, Brno, Czech Republic, August 15-17, 2011. Proceedings* (pp. 101-110). Springer Berlin Heidelberg.

[99] Dey, A., Totel, E., & Navers, S. (2021). Heterogeneous security events prioritization using auto-encoders. In *Risks and Security of Internet and Systems: 15th International Conference, CRiSIS 2020, Paris, France, November 4–6, 2020, Revised Selected Papers 15* (pp. 164-180). Springer International Publishing.

[100] Obaidat, M., Brown, J., & Alnusair, A. (2021, May). Blind attack flaws in adaptive honeypot strategies. In *2021 IEEE World AI IoT Congress (AIIoT)* (pp. 0491-0496). IEEE.

[101] Lopez–Yepez, J. S., & Fagette, A. (2022, December). Increasing attacker engagement on SSH honeypots using semantic embeddings of cyber-attack patterns and deep reinforcement learning. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 389-395). IEEE.

[102] Pauna, A., Iacob, A. C., & Bica, I. (2018, June). Qrassh-a self-adaptive ssh honeypot driven by q-learning. In *2018 international conference on communications (COMM)* (pp. 441-446). IEEE.

[103] Dowling, S., Schukat, M., & Barrett, E. (2019). Using reinforcement learning to conceal honeypot functionality. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18* (pp. 341-355). Springer International Publishing.

[104] Vishwakarma, R., & Jain, A. K. (2019, April). A honeypot with machine learning based detection framework for defending IoT based botnet DDoS attacks. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1019-1024). IEEE.

[105] Wang, B. X., Chen, J. L., & Yu, C. L. (2022). An ai-powered network threat detection system. IEEE Access, 10, 54029-54037.

[106] Memos, V. A., & Psannis, K. E. (2020, October). AI-powered honeypots for enhanced IoT botnet detection. In *2020 3rd World Symposium on Communication Engineering (WSCE)* (pp. 64-68). IEEE.

[107] Araujo, F., Ayoade, G., Al-Naami, K., Gao, Y., Hamlen, K. W., & Khan, L. (2019, December). Improving intrusion detectors by crook-sourcing. In *Proceedings of the 35th Annual Computer Security Applications Conference* (pp. 245-256).

[108] Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., ... & Schmidhuber, J. (2010). PyBrain. *Journal of Machine Learning Research*, 11, 743-746.

[109] Pauna, A., & Bica, I. (2014, May). RASSH-Reinforced adaptive SSH honeypot. In *2014 10th International Conference on Communications (COMM)* (pp. 1-6). IEEE.

[110] Karuna, P., Purohit, H., Jajodia, S., Ganesan, R., & Uzuner, O. (2020). Fake document generation for cyber deception by manipulating text comprehensibility. *IEEE Systems Journal*, 15(1), 835-845.

[111] Ansari, M. S., Bartoš, V., & Lee, B. (2022). GRU-based deep learning approach for network intrusion alert prediction. *Future Generation Computer Systems*, 128, 235-247.

[112] Al Najada, H., Mahgoub, I., & Mohammed, I. (2018, November). Cyber intrusion prediction and taxonomy system using deep learning and distributed big data processing. In *2018 IEEE symposium series on computational intelligence (SSCI)* (pp. 631-638). IEEE.

[113] Rhode, M., Burnap, P., & Jones, K. (2018). Early-stage malware prediction using recurrent neural networks. *Computers & security*, 77, 578-594.

[114] Radoglou-Grammatikis, P., Sarigiannidis, P., Iturbe, E., Rios, E., Martinez, S., Sarigiannidis, A., ... & Ramos, F. (2021). Spear siem: A security information and event management system for the smart grid. *Computer Networks*, 193, 108008.

[115] Zhang, F., Kodituwakku, H. A. D. E., Hines, J. W., & Coble, J. (2019). Multilayer data-driven cyber-attack detection system for industrial control systems based on network, system, and process data. *IEEE Transactions on Industrial Informatics*, 15(7), 4362-4369.

[116] Fatani, A., Abd Elaziz, M., Dahou, A., Al-Qaness, M. A., & Lu, S. (2021). IoT intrusion detection system using deep learning and enhanced transient search optimization. *IEEE Access*, 9, 123448-123464.

[117] Medjek, F., Tandjaoui, D., Djedjig, N., & Romdhani, I. (2021). Fault-tolerant AI-driven intrusion detection system for the internet of things. *International Journal of Critical Infrastructure Protection*, 34, 100436.

[118] Shukla, K. A., Ahamad, S., Rao, G. N., Al-Asadi, A. J., Gupta, A., & Kumbhkar, M. (2021, December). Artificial intelligence assisted IoT data intrusion detection. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)* (pp. 330-335). IEEE.

[119] European Union Agency for Cybersecurity, Malatras, A., Dede, G. (2020). *AI cybersecurity challenges : threat landscape for artificial intelligence*, European Network and Information Security Agency. https://data.europa.eu/doi/10.2824/238222.

[120] Vitorino, J., Oliveira, N., & Praça, I. (2022). Adaptative perturbation patterns: Realistic adversarial learning for robust intrusion detection. *Future Internet*, 14(4), 108.

[121] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018, May). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE symposium on security and privacy (SP)* (pp. 19-35). IEEE.

[122] Zhong, H., Liao, C., Squicciarini, A. C., Zhu, S., & Miller, D. (2020, March). Backdoor embedding in convolutional neural network models via invisible perturbation. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy* (pp. 97-108).

[123] Vos, D., & Verwer, S. (2021, July). Efficient training of robust decision trees against adversarial examples. In *International Conference on Machine Learning* (pp. 10586-10595). PMLR.

[124] Martins, N., Cruz, J. M., Cruz, T., & Abreu, P. H. (2020). Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. *IEEE Access*, 8, 35403-35419.

[125] ISO/IEC TR 24029-1:202, Artificial Intelligence (AI) — Assessment of the robustness of neural networks (2021, March). In *ISO* (SO/IEC TR 24029-1:2021) Retrieved March 31, 2024, from https://www.iso.org/standard/77609.html.

[126] Artificial intelligence Functional safety and AI systems. (2024, January). In *ISO* (ISO/IEC TR 5469:2024). Retrieved March 31, 2024, from https://www.iso.org/standard/81283.html

[127] OECD. (2019, May 22). Recommendation of the Council on Artificial Intelligence. In *OECD Legal Instruments* (OECD/LEGAL/0449). Retrieved March 31, 2024, from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[128] UNESCO. (2022). Recommendation on the Ethics of Artificial Intelligence. In *UNESCO* (SHS/BIO/PI/2021/1). Retrieved March 31, 2024, from https://unesdoc.unesco.org/ark:/48223/pf0000381137

[129] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act). (2019, June 7). In EUR-Lex (PE/86/2018/REV/1). Retrieved March 30, 2024, from https://eur-lex.europa.eu/eli/reg/2019/881/oj.

[130] *AI Act*. (n.d.). European Commission. Retrieved March 31, 2024, from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#:~:text=The%20AI%20Act%20is%20the%20first%2Dever%20comprehensive%20legal%20framework,powerful%20and%20impactful%20AI%20models

[131] Erokhin, S. D., & Zhuravlev, A. P. (2020, July). A Comparative Analysis of Public Cyber Security Datasets. In 2020 Systems of Signal Synchronization, Generating and Processing in *Telecommunications (SYNCHROINFO)* (pp. 1-7). IEEE.

[132] Kastelic. (2021, April 6). International Cooperation to Mitigate Cyber Operations against Critical Infrastructure: Normative Expectations and Emerging Good Practices. In *United Nations Institute for Disarmament Research (UNIDIR)*. Retrieved April 12, 2024, from https://unidir.org/wp-content/uploads/2023/05/International-Cooperation-to-Mitigate-Cyber-Operations-against-Critical-Infrastructure-April-2021.pdf.

[133] Reeder, F., Pomales, C., Kotras, D., & Lockett, J. (2023). Enabling the Department of Defense's Future to Test and Evaluate Artificial Intelligence Enabled Systems. *IEEE Instrumentation & Measurement Magazine*, 26(5), 31-38.

[134] Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems: A Primer. (2023). In *United Nations Institute for Disarmament Research (UNIDIR)* . Retrieved March 31, 2024, from https://unidir.org/wp-content/uploads/2023/11/UNIDIR_Exploring_Synthetic_Data_for_Artificial_Intelligence_and_Autonomous_Systems_A_Primer.pdf.

[135] Yan, J., Lee, E. J., Conover, D., & Kwon, H. (2020). Synthetic dataset generation and adaptation for human detection (p. 0030). Tech. Rep. ARL-TR-9112, US Army Research Laboratory.

[136] Michel, A. H. (2021). Known unknowns: Data Issues and military autonomous systems. In *UNIDIR* (SecTec/21/AI1). United Nations Institute for Disarmament Research (UNIDIR). Retrieved April 17, 2024, from https://unidir.org/wp-content/uploads/2023/05/Holland_KnownUnknowns_20210517_0.pdf.

[137] *1998 DARPA Intrusion Detection Evaluation Dataset*. (n.d.). LINCOLN LABORATORY, MASSACHUSETTS INSTITUTE OF TECHNOLOGY. Retrieved April 8, 2024, from https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset.

[138] *KDD Cup 1999 Data*. (1999, October 28). LINCOLN LABORATORY, MASSACHUSETTS INSTITUTE OF TECHNOLOGY. Retrieved April 8, 2024, from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[139] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE symposium on computational intelligence for security and defense applications* (pp. 1-6). IEEE.

[140] *Intrusion detection evaluation dataset (CIC-IDS2017)*. (n.d.). University of New Brunswick. Retrieved April 8, 2024, from https://www.unb.ca/cic/datasets/ids-2017.

[141] *The CAIDA "DDoS Attack 2007" Dataset*. (n.d.). CAIDA. Retrieved April 8, 2024, from http://www.caida.org/data/passive/ddos-20070804-dataset.xml/https://www.caida.org/catalog/datasets/ddos-20070804_dataset/.

[142] *ISCX datasets, 2009-2016*. (n.d.). Univeristy of New Brunswick. Retrieved April 8, 2024, from http://www.unb.ca/cic/datasets/index.html.

[143] *SPECIAL DATASET CTU-13*. (n.d.). Stratosphere Lab. Retrieved April 8, 2024, from https://stratosphereips.org/category/datasets-ctu13.

[144] Moustafa, N., & Slay, J. (2015, November). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference (MilCIS)* (pp. 1-6). IEEE.

[145] *DDoS evaluation dataset (CIC-DDoS2019)*. (n.d.). University of New Brunswick. Retrieved April 11, 2024, from https://www.unb.ca/cic/datasets/ddos-2019.html

[146] *The UNSW-NB15 Dataset*. (n.d.). UNSW Sidney. Retrieved April 8, 2024, from https://research.unsw.edu.au/projects/unsw-nb15-dataset.

[147] *CSE-CIC-IDS2018 on AWS*. (n.d.). University of New Brunswick. Retrieved April 8, 2024, from https://www.unb.ca/cic/datasets/ids-2018.html.

[148] *WUSTL-IIOT-2018 Dataset for ICS (SCADA) Cybersecurity Research. (n.d.).* Washington University in St. Louis. Retrieved April 8, 2024, from https://www.cse.wustl.edu/~jain/iiot/index.html.

[149] *ADFA IDS Datasets*. (n.d.). UNSW Sidney. Retrieved April 8, 2024, from https://research.unsw.edu.au/projects/adfa-ids-datasets.

[150] *The Bot-IoT Dataset*. (n.d.). UNSW Sidney. Retrieved April 8, 2024, from https://research.unsw.edu.au/projects/bot-iot-dataset.

[151] Garcia, Parmisano, & Jose Erquiaga. (2020, January 20). *IoT-23: A labeled dataset with malicious and benign IoT network traffic*. Zenodo. Retrieved April 8, 2024, from https://zenodo.org/records/4743746.

[152] Zhang, K., Xu, S., & Shin, B. (2023, October). Towards Adaptive Zero Trust Model for Secure AI. In *2023 IEEE Conference on Communications and Network Security (CNS)* (pp. 1-2). IEEE.

[153] Bak, M., Madai, V. I., Fritzsche, M. C., Mayrhofer, M. T., & McLennan, S. (2022). You can't have ai both ways: Balancing health data privacy and access fairly. Frontiers in Genetics, 13, 929453.

[154] Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (No. NIST Artificial Intelligence (AI) 100-2 E2023). National Institute of Standards and Technology.

[155] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2020). Robust physical-world attacks on deep learning visual classification.

[156] Jing, P., Tang, Q., Du, Y., Xue, L., Luo, X., Wang, T., ... & Wu, S. (2021). Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 3237-3254).

[157] Calvo, A., Ortiz, N., Espinosa, A., Dimitrievikj, A., Oliva, I., Guijarro, J., & Sidiqqi, S. (2023, June). Safe AI: Ensuring Safe and Responsible Artificial Intelligence. In *2023 JNIC Cybersecurity Conference (JNIC)* (pp. 1-4). IEEE.

[158] *ATLAS Matrix*. (n.d.). MITRE ATLAS. Retrieved April 12, 2024, from https://atlas.mitre.org/matrices/ATLAS.

[159] Joint Cybersecurity Information: Deploying AI Systems Securely. (2024, April). In *U.S. Department of Defense* (U/OO/143395-24 | PP-24-1538 | April 2024 Ver. 1.0). Retrieved April 21, 2024, from https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF.

[160] National Security Agency & Cybersecurity & Infrastructure Security Agency. (2020, July). Cybersecurity Advisory: NSA and CISA Recommend Immediate Actions to Reduce Exposure Across all Operational Technologies and Control systems. In *U.S. Department of Defense* (U/OO/154383-20 | PP-20-0622). Retrieved April 22, 2024, from https://media.defense.gov/2020/Jul/23/2002462846/-1/-1/0/OT_ADVISORY-DUAL-OFFICIAL-20200722.PDF

[161] Engaging with Artificial Intelligence (AI). (2023). In *U.S. Department of Defense*. Retrieved April 22, 2024, from https://media.defense.gov/2024/Jan/23/2003380135/-1/-1/0/CSI-ENGAGING-WITH-ARTIFICIAL-INTELLIGENCE.PDF.

**Mojca Volk** received her Ph.D. degree in telecommunications from the University of Ljubljana, Slovenia, in 2010. She is with the Laboratory for telecommunications at the Faculty of Electrical Engineering, University of Ljubljana, holding the role of Scientific Associate and habilitated Assistant Professor. Her main areas of work entail R&D and innovative infrastructure and business model development and prototyping in communication networks, services and applications, with a specific focus on wireless communications, 5G/6G technologies and cybersecurity in different verticals, including critical infrastructure, defence and military, energy systems, digital health and public protection and disaster relief. She has been involved in numerous national and international R&D projects in her fields of interest. Her recent work includes scientific research and managerial roles in research projects from the EDA, EDF and ARIS research programmes on the topic of cybersecurity, green energy and public safety. Her pedagogical activities currently comprise courses on innovation and entrepreneurship on the undergraduate university level and applied study programmes at the Faculty of Mechanical Engineering, University of Ljubljana, and mentorships at the MSc and PhD study programmes at the Faculty of Electrical Engineering, University of Ljubljana. She is a member of the Scientific and research council for interdisciplinary sciences at the Slovenian Research and Innovation Agency (ARIS), the technical council on Artificial Intelligence at the Slovenian Standardisation Institute (SIST/TC), and the IEEE society.