Artificial Intelligence for next generation cybersecurity: The AI4CYBER framework

Eider Iturbe
TECNALIA Research Innovation, Basque Research and Technology Alliance (BRTA)
Derio, Bizkaia, Spain
Faculty of Engineering, University of the Basque Country
Bilbao, Spain
eider.iturbe@tecnalia.com

Erkuden Rios
TECNALIA Research Innovation, Basque Research and Technology Alliance (BRTA)
Derio, Bizkaia, Spain
erkuden.rios@tecnalia.com

Angel Rego
TECNALIA Research Innovation, Basque Research and Technology Alliance (BRTA)
Derio, Bizkaia, Spain
angel.rego@tecnalia.com

Nerea Toledo
Faculty of Engineering, University of the Basque Country
Bilbao, Spain
nerea.toledo@ehu.es

ABSTRACT

Artificial Intelligence (AI) is playing a crucial role both in the technological advances and emerging advanced threats in cybersecurity. Despite efforts by competent authorities in Europe to regulate the use of AI in a way that aligns with the ethics and individuals' fundamental rights, there are still challenges to be tackled, not to mention the malicious use of AI by cybercriminals. In this paper we present a novel framework that is composed of innovative cybersecurity services that leverage AI to provide support in the management of the incident response and recovery lifecycle of the critical entities' systems against advanced attacks. The paper describes the main components and architecture of the AI4CYBER framework and provides a clear understanding of the application of the autonomous intelligent cybersecurity services and their role in enforcing defensive actions throughout the entire lifecycle of the systems.

CCS CONCEPTS

• **Security and privacy → Systems security**.

KEYWORDS

cybersecurity, artificial intelligence, adversarial attack, adversarial machine learning, critical infrastructure, intrusion detection, incident response, cyber threat intelligence

ACM Reference Format:
Eider Iturbe, Erkuden Rios, Angel Rego, and Nerea Toledo. 2023. Artificial Intelligence for next generation cybersecurity: The AI4CYBER framework. In *The 18th International Conference on Availability, Reliability and Security (ARES 2023), August 29-September 1, 2023, Benevento, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3600160.3605051

ARES 2023, August 29-September 1, 2023, Benevento, Italy

ACM ISBN 979-8-4007-0772-8/23/08…$15.00
https://doi.org/10.1145/3600160.3605051

1 INTRODUCTION

The evolution of cybersecurity has experienced a dynamic progress influenced by technological advances and new emerging risks. The progression of cyber-attacks and their corresponding defences have been interdependent, shaped by the development of new technologies and paradigms in the digital realm.

One of the technologies that has revolutionized the digital world is Artificial Intelligence (AI). Its application has brought countless improvements in multiple domains, such as medical diagnosis in healthcare, autonomous vehicles in transportation or predictive maintenance in manufacturing. AI can be defined as *"the development of computer systems that possess the ability to perform tasks that typically require human intelligence, such as visual perception, natural language understanding, decision-making, problem-solving, and learning."* [35]. All these capabilities of AI are a double-edged sword that can either be used to implement more intelligent defence mechanisms and strategies or serve as an offensive cyber weapon in the era of Advanced Persistent Threats (APTs) and large-scale data breaches.

There is a plan by the European Commission for the regularization of the use and application of AI, known as the Artificial Intelligence Act [11], which has already taken the first steps towards approval in the European Parliament [13]. However, it can certainly be said that cybercriminals will not follow the rules that will be put in place to determine the level of risk that AI systems could present for safety or fundamental rights. As a matter of fact, a recent report published by Europol outlines important conclusions about the malicious use of AI techniques, such as Large Language Models (LLM) for cybercrime, social engineering, fraud or impersonation [14].

Not only is AI being utilized by cybercriminals to damage systems and services, but AI systems used in diverse industrial sectors for the benefit of the society are also targets of adversarial attacks. Adversarial Machine Learning (AML) involves the exploration of techniques and methods to comprehend and protect against adversarial attacks on machine learning systems [6]. These attacks aim to intentionally generate malfunctions in the performance of machine learning models, resulting in incorrect results [6].

The aforementioned cyber-attacks are not only aimed at digital infrastructures across the world, but the attackers are also attempting to cause damage to critical infrastructure systems, which is particularly alarming. According to the Critical Entities Resilience (CER) directive, a critical entity provides one or more essential services (Article 6), which are crucial for the maintenance of vital societal functions, economic activities, public health and safety, or the environment (Article 2) [12].

In view of the presented context, in which AI is being weaponized against digital infrastructures globally or being attacked using AML techniques that may also employ AI maliciously, there is an evident requirement for the implementation of innovative cybersecurity methods and tools that are more intelligent than their offensive counterparts. These novel approaches must be capable of defending digital systems, particularly the critical infrastructures, throughout their entire lifespan.

In this paper we present the AI4CYBER solution which aims to implement an ecosystem framework of next generation AI-based services to support developers and operators of critical systems. These services provide efficient management of system robustness, resilience, and appropriate response in the event of advanced and AI-powered cyber-attacks.

The main contributions of the work presented in this paper are summarized below.

- A framework of AI-driven cybersecurity tools to ensure a continuum of system protection against advanced cyber-attacks, including AI-powered attacks.
- As integral part of such framework, a trustworthy AI framework that provides explainability, fairness and security capabilities of the AI-driven cybersecurity tools.

The rest of the paper is organized as follows: Section 2 describes previous works related to the cybersecurity research areas covered by AI4CYBER. Section 3 includes an overview of the AI4CYBER framework by explaining the main stakeholders and the high-level architecture. Section 4 provides insights about the implementation and demonstration planning of the AI4CYBER cybersecurity solution in critical entities. Finally, section 5 concludes the paper.

2 RELATED WORK

This section presents previous research works in the cybersecurity areas that the AI4CYBER framework covers, namely: (i) preparedness of critical entities to counter advanced attacks, (ii) advanced threat and anomaly detection, (iii) incident response, and (iv) trustworthiness of AI systems.

2.1 AI-based preparedness

By combining AI-powered atomic attack methods with well-known attack techniques, the result produces more damage and faster harmful impact on digital infrastructures [23]. The new techniques that can be produced using AI are diverse, ranging from self-learning intelligent malware capable of adapting its behaviour to different circumstances and situations found in the victim system by modifying the adversarial tactic, technique or the target device, to intelligent evasion techniques aimed at avoiding detection by security mechanisms in place [9, 19, 41].

The modern digital infrastructures must be prepared to counter these types of sophisticated attacks. Breach and attack simulation (BAS) technology aims to emulate well-known individual attacks as well as attack campaigns in order to assess the security posture of a system. The main characteristic of BAS is the automated and controlled simulation of attack techniques using established frameworks such as MITRE ATT&CK to represent the attack flow and provide a more consistent assessment report of the attack techniques that worked successfully in the system constituting a potential security gap [17]. Cymulate, Picus, and Attack IQ Platform are currently the best-rated BAS tools in the market [18], whereas Caldera platform developed by MITRE is well known as the open-source alternative to commercial BAS solutions [28]. Additionally, defensive AML can be used for attack simulation among other security tasks such as countermeasure designs, noise detection and evasion [36].

Furthermore, Cyber Threat Intelligence (CTI) platforms enable improved identification and comprehension of emerging threat vectors which facilitates the continuous update and improvements of the security mechanisms in place to protect the system. There are numerous CTI platforms and tools, both commercial and open-source, available for sharing and gathering CTI indicators and feeds; two of the most popular open-source tools are OpenCTI and MISP. However, the challenge in utilizing multiple sources of CTI data lies in collecting, processing, analysing and evaluating diverse data types [7]. It is also important to determine what information can be shared across the cybersecurity community to improve AI-based threat detection [31] as well as cyber threat information about emerging AI-powered attacks. In addition to that, protecting user privacy and sensitive data is a critical concern in information sharing.

When it comes to code testing, the integration of AI techniques in code analysis leads to significant advances in vulnerability analysis and automated error detection and correction, although it entails important challenges such as the little availability of standard datasets or the need of ensure reproducibility and replicability of the AI models [37]. Today, some commercial solutions have integrated AI models to provide smarter outcomes; e.g., Codeguru uses Machine Learning (ML) techniques to give intelligent recommendations on code quality improvement and bug fixing [39], and Copilot included Deep Leaning (DL) techniques for vulnerability prevention in code [10]. However, [33] revealed that Copilot's performance recommending insecure code was not entirely efficient and they recommend to pair it with appropriate security-aware tools. Similarly, in academia, relevant solutions that integrate AI models can be found such as vulDeePecker [26] and SyseVR [25] that utilize deep learning techniques, specifically Bi-LSTM (Bidirectional Long Short-Term Memory), for fine-grained program representation and vulnerability localization. Another novel approach is μVulDeePecker [43] that improves the accuracy of the previous solutions [40].

2.2 AI-powered intrusion detection

In the last years, Intrusion Detection Systems (IDSs) have leveraged AI to enhance the accuracy of advanced threat detection. Lansky et al. [24] provide a comprehensive analysis of intrusion detection

systems based on DL models for identifying malicious activities. As future work, there is a need to create new datasets that include new kinds of attack signatures and behaviour. Additionally, most of the analised IDS implementations have utilized supervised learning techniques aligned with the use of pre-existing datasets (such as NSL-KDD and KDDCup99); however, there is a need to explore the use of unsupervised and semi-supervised learning methods [24].

Agrawal et al. [1] propose to implement a Federated Learning (FL) architecture to tackle either the problem of insufficient data or the enormous heterogeneity of data in complex digital infrastructures. Another advantage of the use of FL models is the possibility to build privacy-preserving architectures. However, there are also limitations to FL that must be overcome, such as the communication overhead required for effective training of the distributed agents.

Furthermore, adversarial attacks can significantly impact the effectiveness of AI-powered IDSs. In [3], an analysis of adversarial attacks against IDSs is presents as well as the recommended specific defence strategies for them. The authors explain that several defences may be needed by an AI model-based IDS and highlight the GAN(Generative adversarial network)-based defence strategies such as APE-GAN++.

2.3 AI-based incident response

Security Orchestration and Automation Response (SOAR) solutions in the market utilize machine learning capabilities to enhance incident triage, extend and improve the incident-related information, or assist the SOC analyst in the decision making of the best response actions [30]. In addition, new AI methodologies are emerging that can improve SOAR systems performance, including but not limited to reinforcement learning, selected defensive adversarial learning methods, and Bayesian networks [36]. It is important to note that each AI methodology may function differently depending on the type of system it is deployed in, such as enterprise IT or an industrial IoT environment, which is an area of future research [36]. Reinforcement learning has proven effective in controlling Software Defined Network (SDN)-based systems and can be utilized as a security measure [27]. Other AI methodologies have been used to enhance cyber incident responses in various applications, like: defending against DDoS attacks with hidden Markov models and cooperative reinforcement learning [42], protecting IDS for botnet detection from adversarial attacks through deep reinforcement learning mechanisms [4], and designing optimal defence for Cyber-Physical Systems (CPSs) using deep reinforcement learning and game theory mechanisms [15].

2.4 Trustworthiness of AI

The "Ethics Guidelines for Trustworthy AI" report published by the European Commission defines the main factors of trustworthy AI, including technical robustness and safety, transparency, and diversity, non-discrimination and fairness, which need to be included into AI system's life cycle [20]. Technical robustness implies ensuring the reliability of AI models, as well as accuracy, security and resilience against adversarial attacks. On the other hand, providing transparency of AI systems refers to enabling explanations about their decision and outcomes so the users can clearly understand how the AI system functions. Finally, addressing fairness of AI systems ensures there will be no bias for any sensitive attribute and the fundamental rights of individuals are respected by providing equal treatment.

In recent years, established entities in the field of cybersecurity have made an effort to define concepts and taxonomies of adversarial attacks in order to provide reference structured knowledge around AML. It is worth noting the NIST AI 100-2e2023 initiative that defines a new taxonomy and terminology of adversarial attacks and mitigations to support the management of AI systems' security [32]. On the other hand, MITRE ATLAS specifies a knowledge base that includes adversarial threat landscape for AI systems that contains adversary tactics and techniques as well as potential mitigations for them [29]. Both mentioned initiatives have recently updated their AML-related classification to include mitigation recommendations, which is crucial when dealing with adversarial attacks.

Regarding interpretability of AI systems, [5] provides a taxonomy of explainability techniques related to different ML models and distinguishes between interpretable or transparent models, and model interpretability through external techniques known as post-hoc explainability. Ali et al. [2] propose four-axes explainability methodology based on the following explainability methods: scoop-based explainers, model-based explainers, complexity-based explainers and methodology-based explainers. Those models produce three different types of explainability: data, model and post-hoc explainability. One aspect to be considered is the potential malicious use of explainability methods to generate adversarial attacks [5].

3 AI4CYBER FRAMEWORK

With the continuous increase of advanced threats, it is essential that the cybersecurity mechanisms part of the defence system be capable of countering those threats while ensuring systems robustness, resilience, and early response of systems. The AI4CYBER framework tackles the growing need for innovative and intelligent methods, tools and services in the cybersecurity field, covering all phases of the incident response workflow.

This section describes the AI4CYBER framework from different viewpoints: (i) first, the main stakeholders interacting with the framework are presented; (ii) second, the high-level architecture that includes the core logical components of the framework is reported; (iii) third, the information model of AI4CYBER is defined; (iv) and finally, the alignment of the AI4CYBER framework with the security incident workflow defined by NIST [8] is explained.

3.1 AI4CYBER stakeholders

As shown in figure 1, there are three main roles that directly interact with the AI4CYBER solution. During the development phase of the system, the *system developer* is the responsible actor for releasing non-vulnerable versions of the code to be installed in the system. Following NIST SP 800-37 v2, an established risk management framework, the system developer is mapped to the role *"responsible for conducting systems security or privacy engineering activities as part of the SDLC"*, which is called a *systems security engineer* [16].

During the operational phase of the system, the *security operator* is in charge of ensuring the proper functioning of the system from a cybersecurity perspective. According to NIST SP 800-37 v2, the

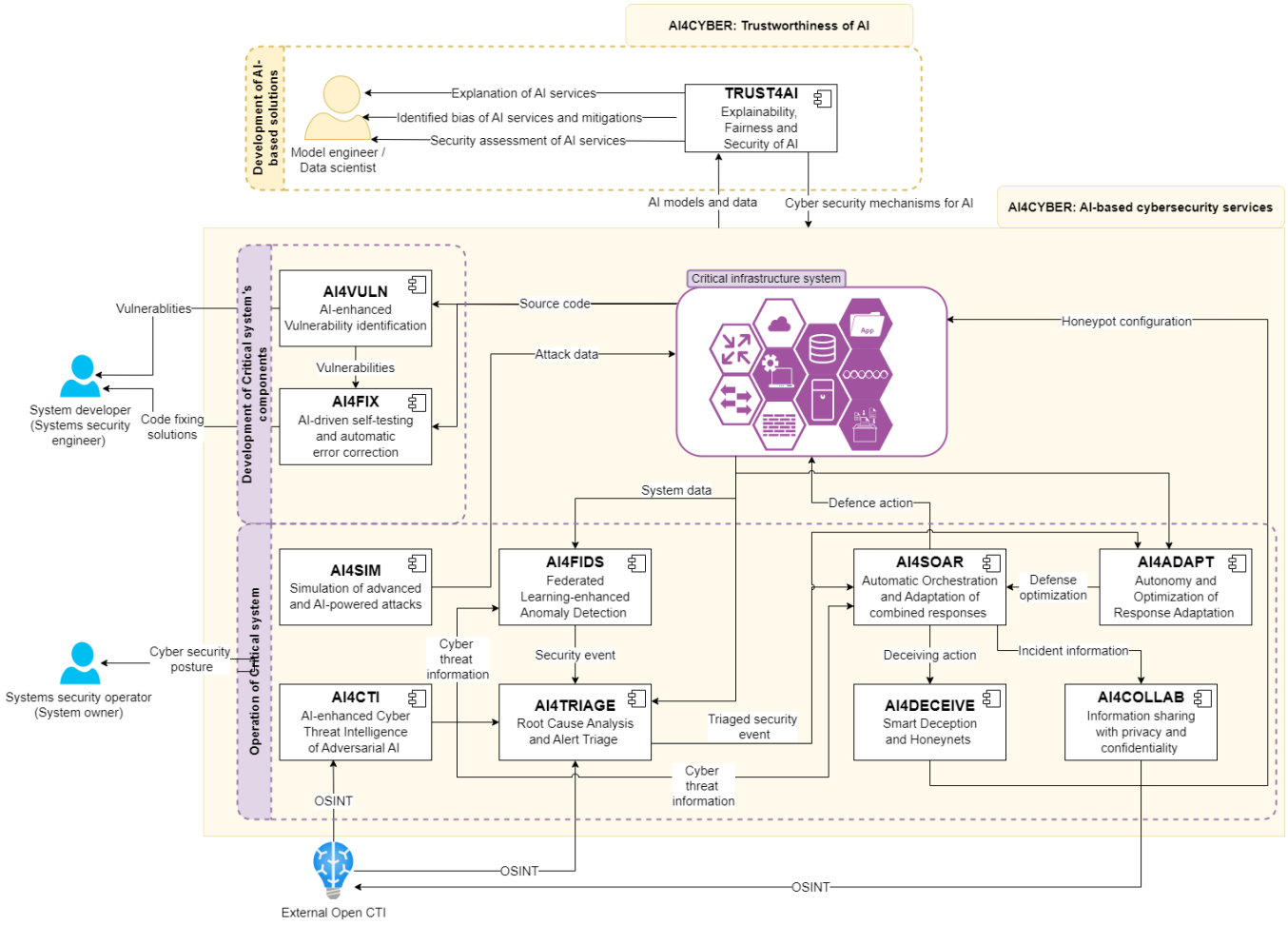

Figure 1: AI4CYBER Components diagram

security operator is mapped to the *system owner* which is responsible for *"... integration, modification, operation, maintenance, and disposal of a system"* [16].

Finally, an essential role in the development, verification and validation of AI models is the *model engineer* or *data scientist* who is in charge of identifying and correcting problems [38].

3.2 AI4CYBER architecture

Figure 1 shows the high-level architecture of the AI4CYBER framework. It represents the primary logical components of the solution, their interdependence, and the data required by these interactions. Two main parts can clearly be differentiated in the architecture: (i) AI-based cybersecurity services, and (ii) services for trustworthiness of AI.

3.2.1 *AI-based cybersecurity services.*

(1) **AI4VUN AI-enhanced vulnerability identification.** AI4-VULN is an innovative solution that utilizes symbolic execution integrated with AI for automatic identification and verification of potential vulnerabilities and weaknesses in the code, achieving a higher accuracy rate and improved scalability.

(2) **AI4FIX AI-driven self-testing and automatic error correction.** AI-driven self-testing and automatic error correction. AI4FIX is an automated code testing and error correction service that leverage AI technology for identifying and rectifying most dangerous "warning" level issues that could potentially lead to security vulnerabilities. By addressing the main security gaps of a code, AI4FIX can reduce the number of exploitable vulnerabilities and enhance the overall robustness of the system.

(3) **AI4SIM Simulation of advanced and AI-powered attacks.** AI4SIM is an attack simulation solution designed to generate advanced and AI-powered attacks. It implements AI-powered attacks, adversarial AI techniques, and advanced attacks to emulate emerging attack techniques used by malicious groups, such the ones identified in the MITRE ATT&CK knowledge base. The goal of AI4SIM is to test the security mechanisms of an organization's cyber defence strategy and prepare them for the operational phase.

(4) **AI4CTI AI-enhanced cyber threat intelligence of adversarial AI.** AI4CTI collects cyber threat information from diverse open CTI data sources such as CTI platforms, attack vectors databases, and reports from intelligence analysts. This information is used to generate structured and comprehensive cyber threat intelligence that can be utilized by the security mechanisms responsible for the system's defence, such as Indicators of Compromise (IoCs) of emerging attacks, possible mitigations, etc.
(5) **AI4FIDS federated learning-enhanced detection.** AI4FIDS is FL-based IDS that identifies a broad range of threats and anomalies while simultaneously protecting the privacy and confidentiality of the relevant entities.
(6) **AI4TRIAGE Root cause analysis and alert triage.** AI4TRIAGE examines the contextual data and details of identified cyber security incidents and offers a comprehensive analysis of the underlying causes. Additionally, it categorizes and ranks the security events using AI-enhanced triage methods. It provides situational awareness of the cybersecurity status of the system and guides both security operators and tools in making decisions on the subsequent responses to counter the attack.
(7) **AI4SOAR Automatic orchestration and adaptation of combined responses.** AI4SOAR provides the intelligent orchestration of multiple security measures across different levels of the system to effectively respond to cyber incidents and attacks.
(8) **AI4ADAPT Autonomy and optimization of response adaptation.** AI4ADAPT employs AI-based methods for optimizing incident response against sophisticated cyber-attacks. It uses reinforcement learning techniques that aim to learn the optimal defence strategies of the system in a given environment by obtaining the highest possible reward. Within the framework of AI4CYBER, AI4ADAPT's outcome is used by AI4SOAR to enhance the defensive strategy of a system.
(9) **AI4DECEIVE Smart deception and honeynets.** AI4DECEIVE provides intelligent deception methods to respond to cyber-attacks. It incorporates intelligence to enhance the deception strategy, which decides on the setup and configuration of honeypot networks that prolong the time attackers spend in those honeypots, while keeping them away from the real system.
(10) **AI4COLLAB Information sharing with privacy and confidentiality.** AI4COLLAB enables third parties, such as CERTS and industry organizations, to prepare their systems to counter reported cyber-attacks by automatically sharing incident information while preserving both privacy and confidentiality.

3.2.2 Trustworthiness of AI. **TRUST4AI** component has the responsibility of offering advanced techniques and models to guarantee the trustworthiness of AI systems. Three complementary services are provided, namely:

- Interpretability (i.e. explainability also known as XAI) of AI models used in AI4CYBER services, for better understanding of AI model's outcomes to the data scientist or model engineer, or even to a potential end-user.
- Fairness of the AI models used in AI4CYBER services to rectify potential bias of sensitive attributes.
- Technology robustness or security of AI models used in AI4CYBER services, based on AML methods to protect against adversarial attacks.

3.3 Information model

Table 1 collects the list of data items handled by the AI4CYBER framework components. It describes the inwards-outwards data specified in the high-level architecture flows shown in Figure 1.

3.4 Alignment with NIST 800-61

Figure 2 shows how the services provided by the AI4CYBER components are aligned with the security incident process defined in NIST SP 800-61 Computer Security Incident Handling Guide [8]. NIST incident repose life cycle is composed of four main phases: (i) preparation, in which the system is protected before going into operation by installing security mechanisms; (ii) detection and analysis, which is essential to identify potential threats during the operational phase; (iii) containment, eradication and recovery, which involves an early reaction against detected attacks with mitigation and recovery actions and may also include a loop to detection to improve the entire system analysis; and (iv) finally, post-incident activity, which includes the forensic work of the incident to be leveraged in preparation for future incidents.

Below it is explained how AI4CYBER services are providing capabilities for each of the phases in the security incident workflow defined in NIST 800-61.

- **Preparation.** During the development phase of the system, AI4VULN and AI4FIX verify that no vulnerabilities exist in the code that can be exploited by adversaries. Additionally, TRUST4AI provides a means to identify AML-related weaknesses in AI systems and correct them. During the pre-operational phase of the system, AI4SIM runs advanced attack simulations in an automated and controlled manner to identify potential misconfigurations and security gaps in the system. AI4CTI continuously provides cyber threat information for the correct configuration at operation of the security mechanisms in place.
- **Detection and analysis.** During the operational phase, AI4FIDS detects sophisticated attacks and AI4TRIAGE determines their level of importance and underlying causes.
- **Containment, eradication and recovery.** During the operational phase, AI4SOAR coordinates AI4ADAPT and AI4DECEIVE together with the set of security mechanisms deployed in the system, to provide automated, intelligent, self-adapting orchestration of incident response mechanisms, including deception methods.
- **Post-incident activity.** After an incident has been addressed, AI4COLLAB enables the system owner to share incident-related information, such as attack technique used and most appropriate response actions, while preserving privacy and confidentiality.

Data item	Description
AI model and data	They represent trained AI models and the datasets utilized at training phase, e.g., AI4FIDS component.
Attack data	It represents the data generated during the execution of a cyber-attack in a digital infrastructure. E.g., network traffic or device's system activity generated during the attack.
Code fixing solution	It represents the correction of the source code installed in the system that fixes previously identified vulnerabilities or misconfigurations.
Cyber threat information	NIST defines cyber threat information as *"any information that can help an organization identify, assess, monitor, and respond to cyber threats. Cyber threat information includes indicators of compromise; tactics, techniques, and procedures used by threat actors; suggested actions to detect, contain, or prevent attacks; and the findings from the analyses of incidents. Organizations that share cyber threat information can improve their own security postures as well as those of other organizations"* [22].
Cyber Threat Intelligence (CTI)	NIST defines cyber threat intelligence as *"cyber threat information that has been aggregated, transformed, analyzed, interpreted, or enriched to provide the necessary context for decision-making processes"* [22].
Cybersecurity event	NIST defines cybersecurity event as *"a cybersecurity change that may have an impact on organizational operations (including mission, capabilities, or reputation)"* [34].
Cybersecurity posture	NIST defines cybersecurity posture as *"the security status of an enterprise's networks, information, and systems based on information security resources (e.g., people, hardware, software, policies) and capabilities in place to manage the defense of the enterprise and to react as the situation changes"* [21].
Deceiving action	It represents a specific type of defence action that lures the attacker to interact with devices called honeypots that emulate the behaviour of real devices in the system. The more time the attacker spends interacting with honeypots, the better defensive action is.
Defence action	It involves the execution of a cybersecurity mechanism as a defence method against potential threats or real attacks. The cybersecurity mechanism can be enabled either as a direct reaction to an identified security incident or as preventive action as part of the cyber defence strategy.
Explanation of AI service	It represents the information elements provided by TRUST4AI-XAI to interpret a trained AI model.
Honeypot configuration	It represents the data required to execute any configuration update of the honeynet as part of the defence mechanism of the system.
Identified bias of AI services and mitigations	They represent the information element provided to the data scientist by TRUST4AI-Fairness about the bias identified in a trained AI model as well as the possible mitigations to rectify it.
Open source intelligence (OSINT)	It is cyber threat intelligence that has been publicly shared within the cyber security community to assist security experts in enhancing the security posture of their systems.
Security assessment of AI services	It represents a list of potential security gaps linked to a trained AO model as well as the possible mitigations to rectify them.
Source code	It is a set of instructions coded in a particular programming language that can be executed as a software solution in a system.
System data	It involves heterogeneous types of data collected from the system to be protected. The data sources can be diverse, ranging from network traffic to device's log data and operational data and even outcomes from other security mechanisms such as IDSs.
Triaged cybersecurity event	It is a cybersecurity event that has been further analysed to define its underlying causes and explore its relation to other cyber threat information in order to determine its level of priority.
Vulnerability	NIST defines vulnerability as *"a weakness in an information system, system security procedures, internal controls, or implementation that could be exploited or triggered by a threat source"* [16].

Table 1: AI4CYBER Information model

4 IMPLEMENTATION AND DEMONSTRATION CONSIDERATIONS

The operating premise of the AI4CYBER solution is based on the fact that the provided cybersecurity services can function individually and be integrated into the global defence infrastructure of the system, complementing existing cybersecurity mechanisms. Within the AI4CYBER project, the demonstrations have been organized to ensure that the components with capabilities for each phase of the NIST 800-61 workflow are integrated.

As mentioned in previous sections, one of the most significant challenges in developing AI-based cybersecurity solutions, including both ML and DL, is obtaining or generating datasets that include emerging advanced attacks. This is necessary to ensure that the models are continuously updated and capable of learning to respond intelligently to novel real attacks. Having access to these datasets is essential when training supervised models, which currently dominate the anomaly-based intrusion detection field [24].

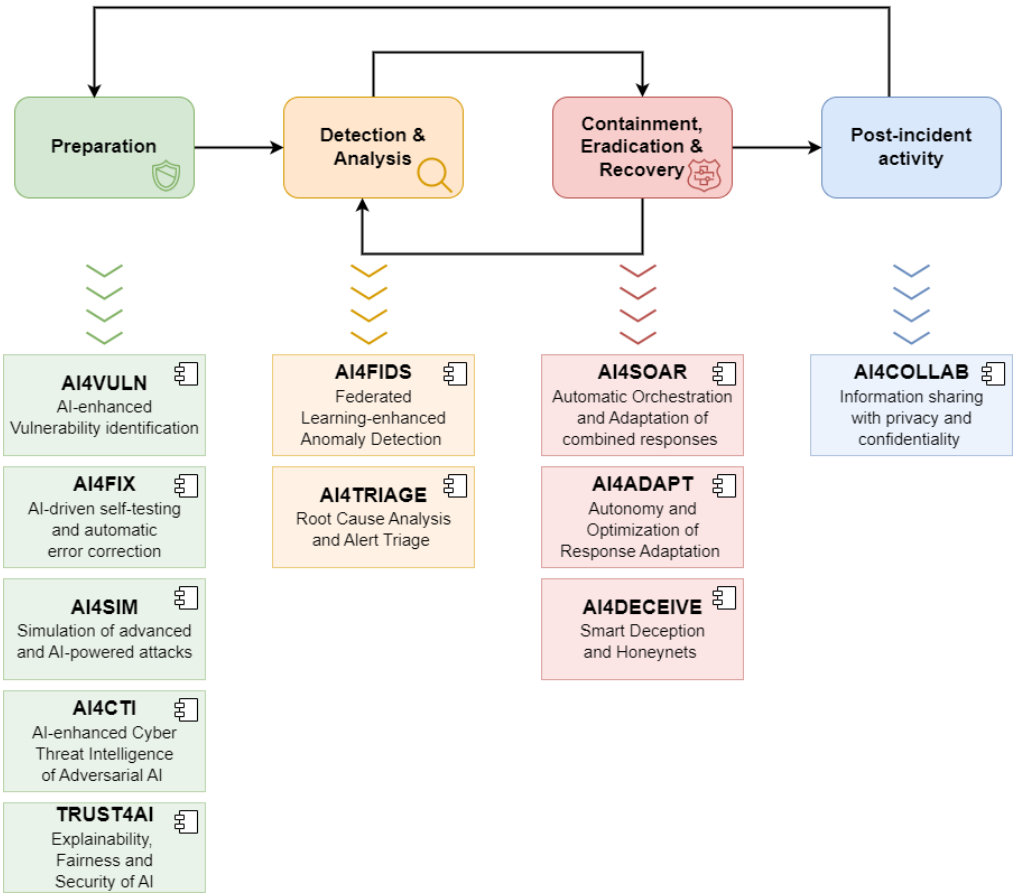

Figure 2: AI4CYBER Alignment with NIST 800-61

AI4CYBER plans to work on realistic industrial use cases; in fact, testing labs with representative digital infrastructure of the following three domains will be used: energy, banking and healthcare. During the project, datasets containing signatures and behaviours of advanced and AI-based attacks will be generated in order to efficiently train the components' models for detection. Additionally, specific datasets will be created with the aim of learning how incident response mechanisms and actions work and estimate their impact against the attacks.

The final version releases of AI4CYBER components, together with the previously curated datasets, will be released to the cybersecurity community to promote continuous enhancement of defensive systems. This represents a major contribution to research in the field of cybersecurity.

5 CONCLUSIONS

Technological advancements will continuously guide as well as condition the evolution of cybersecurity. The use of AI technology entails growing concerns about compliance with ethics and fundamental rights. Therefore, achieving trustworthy AI systems has become one of the priorities at the European level.

AI is an instrument that can be utilized for both offensive and defensive purposes. Moreover, a novel field called AML within cybersecurity is gaining importance to better understand how the adversarial attacks can cause malfunctions in AI systems, resulting in incorrect decisions and outcomes.

The paper describes the main contributions provided by the AI4CYBER framework, an ecosystem of cybersecurity services that use the potentiality of AI to support critical infrastructure owners on the management of the entire lifecycle of the response incident process (i.e. preparedness, intrusion detection, incident response and post-incident processing). It also implements services to ensure trustworthy AI (i.e. explainability, fairness and security of AI) is achieved within the AI systems.

Another major planned contribution from AI4CYBER is the generation of cybersecurity-related datasets to promote the improvement of cybersecurity knowledge in the community and build more intelligent security mechanisms. It has already been considered as major requirement during the implementation of the AI4CYBER components.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450 (AI4CYBER).

REFERENCES

[1] Shaashwat Agrawal, Sagnik Sarkar, Ons Aouedi, Gokul Yenduri, Kandaraj Piamrat, Mamoun Alazab, Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta, and Thippa Reddy Gadekallu. 2022. Federated Learning for intrusion detection system: Concepts, challenges and future directions. *Comput. Commun.* 195 (2022), 346–361.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* (2023), 101805.

[3] Afnan Alotaibi and Murad A Rassam. 2023. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet* 15, 2 (2023), 62.

[4] Giovanni Apruzzese, Mauro Andreolini, Mirco Marchetti, Andrea Venturi, and Michele Colajanni. 2020. Deep reinforcement adversarial learning against botnet evasion attacks. *IEEE Trans. Netw. Serv. Manag.* 17, 4 (2020), 1975–1987.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[6] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognit.* 84 (2018), 317–331.

[7] Rebekah Brown and Robert M Lee. 2021. 2021 sans cyber threat intelligence (cti) survey. In *Tech. Rep.* SANS Institute.

[8] Paul Cichonski, Tom Millar, Tim Grance, Karen Scarfone, et al. 2012. Computer security incident handling guide. *NIST Special Publication* 800, 61 (2012), 1–147.

[9] Darktrace. 2018. *The next paradigm shift AI-Driven Cyber-Attacks.* Technical Report.

[10] Sergio De Simone. 2023. GitHub enhanced Copilot with new AI model and security-oriented capabilities. https://www.infoq.com/news/2023/02/github-enhanced-copilot-codex/. Accessed: 2023-5-20.

[11] European Commission. 2021. EUR-Lex - 52021PC0206 - EN - EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Accessed: 2023-5-19.

[12] European Commission. 2022. EUR-Lex - 32022L2557 - EN - EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022L2557. Accessed: 2023-5-19.

[13] European Parliament. 2023. AI Act: a step closer to the first rules on Artificial Intelligence. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence. Accessed: 2023-5-19.

[14] Europol. 2023. The impact of large language models on law enforcement. https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf. Accessed: 2023-5-19.

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

[16] Joint Task Force. 2018. Risk management framework for information systems and organizations. *NIST Special Publication* 800 (2018), 37.

[17] Gartner. 2021. Quick Answer: What Are the Top Use Cases for Breach and Attack Simulation Technology? https://www.gartner.com/en/documents/3997362. Accessed: 2023-5-19.

[18] Gartner, Inc. 2023. Breach and Attack Simulation (BAS) Tools reviews 2023. https://www.gartner.com/reviews/market/breach-and-attack-simulation-bas-tools. Accessed: 2023-5-20.

[19] Blessing Guembe, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz, and Vera Pospelova. 2022. The emerging threat of Ai-driven cyber attacks: A review. *Appl. Artif. Intell.* 36, 1 (2022), 1–34.

[20] High-Level Expert Group on Artificial Intelligence. 2019. *ETHICS GUIDELINES FOR TRUSTWORTHY AI.* Technical Report.

[21] Arnold Johnson, Kelley Dempsey, Ron Ross, Sarbari Gupta, Dennis Bailey, et al. 2011. Guide for security-focused configuration management of information systems. *NIST special publication* 800, 128 (2011), 16–16.

[22] Chris Johnson, Lee Badger, David Waltermire, Julie Snyder, Clem Skorupka, et al. 2016. Guide to cyber threat information sharing. *NIST special publication* 800, 150 (2016).

[23] Nektaria Kaloudi and Jingyue Li. 2021. The AI-based cyber threat landscape: A survey. *ACM Comput. Surv.* 53, 1 (2021), 1–34.

[24] Jan Lansky, Saqib Ali, Mokhtar Mohammadi, Mohammed Kamal Majeed, Sarkhel H Taher Karim, Shima Rashidi, Mehdi Hosseinzadeh, and Amir Masoud Rahmani. 2021. Deep learning-based intrusion detection systems: A systematic review. *IEEE Access* 9 (2021), 101574–101599.

[25] Zhen Li, Deqing Zou, Shouhuai Xu, Hai Jin, Yawei Zhu, and Zhaoxuan Chen. 2022. SySeVR: A framework for using deep learning to detect software vulnerabilities. *IEEE Trans. Dependable Secure Comput.* 19, 4 (2022), 2244–2258.

[26] Zhen Li, Deqing Zou, Shouhuai Xu, Xinyu Ou, Hai Jin, Sujuan Wang, Zhijun Deng, and Yuyi Zhong. 2018. VulDeePecker: A deep learning-based system for vulnerability detection. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, Reston, VA.

[27] Shih-Chun Lin, Ian F Akyildiz, Pu Wang, and Min Luo. 2016. QoS-aware adaptive routing in multi-layer hierarchical software defined networks: A reinforcement learning approach. In *2016 IEEE International Conference on Services Computing (SCC)*. IEEE.

[28] MITRE. 2023. CALDERA. https://caldera.mitre.org/. Accessed: 2023-5-20.

[29] MITRE. 2023. MITRE ATLAS. https://atlas.mitre.org/. Accessed: 2023-5-22.

[30] Claudio Neiva, Craig Lawson, Toby Bussa, and Gorka Sadowski. 2020. *Market guide for security orchestration, automation and response solutions.* Technical Report. Technical Report. Gartner.

[31] Talha Ongun, Simona Boboila, Alina Oprea, Tina Eliassi-Rad, Alastair Nottingham, Jason Hiser, and Jack Davidson. 2021. Collaborative information sharing for ML-based threat detection. (2021). arXiv:2104.11636

[32] Alina Oprea and Apostol Vassilev. 2023. *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (draft).* Technical Report. National Institute of Standards and Technology.

[33] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2022. Asleep at the keyboard? Assessing the security of GitHub copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE.

[34] Ron Ross, Victoria Pillitteri, Richard Graubart, Deborah Bodeau, and Rosalie McQuaid. 2019. *Developing cyber resilient systems: a systems security engineering approach.* Technical Report. National Institute of Standards and Technology.

[35] S Russell and P Norvig. 2010. *Artificial Intelligence: A Modern Approach.* Prentice Hall.

[36] Sagar Samtani, Murat Kantarcioglu, and Hsinchun Chen. 2020. Trailblazing the artificial Intelligence for cybersecurity discipline: A multi-disciplinary research roadmap. *ACM Trans. Manag. Inf. Syst.* 11, 4 (2020), 1–19.

[37] Tushar Sharma, Maria Kechagia, Stefanos Georgiou, Rohit Tiwari, Indira Vats, Hadi Moazen, and Federica Sarro. 2021. A survey on machine learning techniques for source code analysis. (2021). arXiv:2110.09610

[38] Elham Tabassi. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). (2023).

[39] Darryl K Taft. 2020. Amazon CodeGuru uses machine learning to improve code. https://www.techtarget.com/searchsoftwarequality/news/252485652/AWS-CodeGuru-uses-machine-learning-to-improve-code-quality. Accessed: 2023-5-20.

[40] Jingjing Wang, Minhuan Huang, Yuanping Nie, and Jin Li. 2021. Static analysis of source code vulnerability using machine learning techniques: A survey. In *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE.

[41] Christopher Whyte. 2020. Problems of poison: New paradigms and "agreed" competition in the era of AI-enabled cyber operations. In *2020 12th International Conference on Cyber Conflict (CyCon)*. IEEE.

[42] Xin Xu, Yongqiang Sun, and Zunguo Huang. 2007. Defending DDoS attacks using hidden Markov models and cooperative reinforcement learning. In *Intelligence and Security Informatics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 196–207.

[43] Deqing Zou, Sujuan Wang, Shouhuai Xu, Zhen Li, and Hai Jin. 2019. *M*VulDeePecker: A deep learning-based system for multiclass vulnerability detection. *IEEE Trans. Dependable Secure Comput.* (2019), 1–1.