

# Towards trustworthy AI: Security risk assessment methodology for Artificial Intelligence systems

Eider Iturbe  
TECNALIA,

Basque Research and Technology Alliance (BRTA)  
Derio, Spain  
eider.iturbe@tecnalia.com  
University of the Basque Country  
Bilbao, Spain

Erkuden Rios  
TECNALIA,

Basque Research and Technology Alliance (BRTA)  
Derio, Spain  
erkuden.rios@tecnalia.com

Nerea Toledo

Faculty of Engineering  
University of the Basque Country  
Bilbao, Spain  
nerea.toledo@ehu.eus

**Abstract**—The digitalization and smartization of modern digital systems include the implementation and integration of emerging innovative technologies, such as Artificial Intelligence. By incorporating new technologies, the surface attack of the system also expands, and specialized cybersecurity mechanisms and tools are required to counter the potential new threats. This paper introduces a holistic security risk assessment methodology that aims to assist Artificial Intelligence system stakeholders guarantee the correct design and implementation of technical robustness in Artificial Intelligence systems. The methodology is designed to facilitate the automation of the security risk assessment of AI components together with the rest of the system components. Supporting the methodology, the solution to the automation of AI risk assessment is also proposed. Both the methodology and the tool will be validated when assessing and treating risks on artificial intelligence-based cybersecurity solutions integrated in industrial systems.

**Index Terms**—Technical robustness, Trustworthy, Artificial Intelligence, Cyber Security, Risk Assessment

## I. INTRODUCTION

The recent threat landscape report published by ENISA includes "Threats against data" as the fourth of the top eight primary threats identified [1]. This category of threat includes both data manipulation and adversarial attacks such as data poisoning against Artificial Intelligence (AI) models. AI models, which have become essential components of modern digital systems, providing intelligence and support for decision making, are increasingly becoming the focus of attacks [1].

Artificial intelligence systems, or 'AI systems' as specified by the European Union regulation, are *designed to operate with varying levels of autonomy and be used on a stand-alone basis or as a component of a product in order to generate outputs such as content, predictions, recommendations, or*

*decisions which influence the environment with which the system interacts* [2].

In April 2021, the European Union proposed the Artificial Intelligence Act, which seeks to establish a unified regulatory and legal structure for artificial intelligence [2]. The main motivation of the European Parliament is to ensure that AI systems used in the European Union are "safe, transparent, traceable, non-discriminatory, and environmentally friendly" [3]. It is important that AI systems are supervised by humans rather than relying exclusively on automation to avoid negative results.

The High-Level Expert Group on AI set up by the European Commission presented their conclusions published in the "Ethics Guidelines for Trustworthy Artificial Intelligence" report in April 2019 [4] which state that trustworthy AI should adhere to the three following principles. First, it must respect all applicable laws and regulations, ensuring its compliance. Secondly, it must adhere to ethical principles and values to prevent harm on humans and respect for their autonomy, as well as to maintain fairness and interpretability. Lastly, it should have technical and social robustness to prevent unintentional damage and be protected from attackers.

Given this context, this paper presents a security risk assessment methodology that aims to support AI system owners and/or providers in evaluating risks of the AI system, while ensuring that technical robustness of the AI components in the system is correctly designed and implemented. The goal is to provide a resource that supports the stakeholders who operate the AI system towards the assurance of a trustworthy AI system development and operation.

The rest of the paper is organized as follows: Section II describes previous works related to the cybersecurity research areas for AI systems. Section III includes an overview of the methodology specialized for the risk assessment of AI systems. Section IV provides insights about the implementation and

validation planned. Finally, section V concludes the paper.

## II. RELATED WORK

The methodology proposed in this paper addresses the needs of AI system owners and providers to systematically assess the security threats and risks of their systems in order to ensure the technical robustness of the AI required by the regulation. In this section, we will explain the most recent developments in cybersecurity risk management solutions for AI systems.

The US National Institute of Standards and Technology (NIST) has created two frameworks that assist organizations in effectively managing cyber risks in their information systems. The first framework, known as the NIST Cybersecurity Framework (CSF), has recently released the public draft of the 2.0 version [?] and has been widely utilized since its original publication in 2014 to mitigate cybersecurity risks. The Risk Management Framework (RMF), the second framework, provides organizations with a seven-step process for guidance on effectively managing both security and privacy risks in their information systems [5]. The RMF is aligned with the CSF.

In January 2023, NIST created the AI RMF framework to enhance the management of risks linked to AI for individuals, organizations, and society [6]. The objective of the AI RMF is to provide a mechanism for organizations involved in the design, development, deployment, or utilization of AI systems. Its purpose is to assist in the management of the risks associated with AI and to encourage the reliable and responsible development and use of AI systems [7]. AI RMF, in contrast to RMF, consists of four activities instead of the seven main activities. These activities are as follows: govern which manages the cross-dependencies throughout the other three activities, map which covers the identification of risks related to context, measure which performs the assessment, analysis and tracking of the detected risks and manage which addresses the prioritization of the assessed risks and implements the appropriate mechanisms. Furthermore, one of the main aspects that is being prioritized in the specification of the AI RMF is ensuring alignment with international standards such as ISO/IEC 5338, ISO/IEC 38507, ISO/IEC 22989 and ISO/IEC 24028 [8].

The NIST AI 100-2 E2023 initiative defined a new taxonomy and terminology for adversarial attacks and their mitigations, which aids in the managements of the security of AI systems [9]. The purpose of the taxonomy is to provide information for standards and future practice guides in order to assess and manage the security of AI systems.

Another prominent organization in the cybersecurity field, MITRE, created the “Systems Engineering for Mission Assurance” methodology to safeguard critical processes and assets that support an organization’s mission [10]. This methodology consists of two main steps: Crown Jewels Analysis and Threat Assessment and Remediation Analysis.

MITRE created a public knowledge base of attack tactics and techniques known as MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) [11]. This knowledge base is built on the analysis of real attacks,

particularly those of advanced persistent threat (APT) nature. MITRE ATT&CK has become a widely used reference tool in the cybersecurity field, as it provides a common taxonomy for Tactics, Techniques, and Procedures (TTPs), which are considered valuable cyber threat information.

Moreover, MITRE created MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) which is a knowledge base that consists of adversary attacks’ TTPs, along with possible countermeasures to mitigate them [12]. ATLAS has been specified following the MITRE ATT&CK framework standard and is complementary to the cyber threat information included in ATT&CK.

ENISA has recently published a report about a multi layer framework for good cybersecurity practices for AI. It introduces a scalable framework that provides guidance to national competent authorities and AI stakeholders on how to ensure the security of their AI systems, operations and processes by following good cybersecurity practices [13]. The framework is structured of three layers, namely: cybersecurity foundations, AI-specific cybersecurity, and sector-specific cybersecurity for AI.

Additionally, ENISA also provided a taxonomy for machine learning algorithms, an in-depth examination of risks and security measures commonly used by standards [14].

There are emerging new commercial solutions that are compliant to MITRE ATLAS knowledge base and offer support for the automation of adversarial analysis, such as the AI Risk Database [15] and the Continuous Validation tool [16] that relies in the former one. However, they do not provide comprehensive methodologies that address the security of the whole perspectives of the security, i.e. the complete AI system.

## III. SECURITY RISK ASSESSMENT METHODOLOGY FOR ARTIFICIAL INTELLIGENCE SYSTEMS

### A. Methodology requirements

When analyzing and evaluating cyber risks in AI-based systems, it is advisable to adopt agile, holistic methodologies that address risks not only at the system development stage but also at the operation stage, and that are oriented towards the continuous assessment of risks. In particular, the methodology shall address the following requirements:

- **Focus on adversarial threats to AI-systems**, which can be of different nature, according to the security aspect of the information handled by the AI-system that is targeted by the adversary: confidentiality, integrity or availability of the data, model or services composing the AI-system.
- **AI-context awareness**, i.e., a holistic methodology is required in terms of supporting assessment of cyber risks to the whole AI-based system, not only the AI components but also the context in which they operate.
- **Comprehensiveness in covering the whole DevOps life-cycle**, i.e., a unified assessment methodology is needed that can be applied to the system in development phase but also to the system in operation.
- **Support to continuous re-assessments**, as the adversaries’ activity and techniques evolve, the cyber risk

status of the system is not static and it needs to be re-evaluated iteratively.

- **Systematic**, i.e. the methodology shall propose a unified method for handling the assessment of multiple types of cyber risks in a well-structured activity flow.
- **Reusable**, the methodology shall be repeatable and applicable to AI-based systems of diverse nature.
- **Agile** in completing the risk assessment process, avoiding tedious tasks addressing the resource shortage suffered by most Small and medium-sized enterprises (SMEs) and allowing to keep pace with fast threat evolution.
- **Facilitate automation**, due to the complexity and modularity of the AI-based systems, the evaluation of the severity of the adversarial threats shall be automated as much as possible in order to facilitate the task to security responsible staff.

The KYKLOS 4.0 risk assessment methodology for security and privacy of industrial solutions, introduced in [17], has been selected as a good candidate methodology to this purpose because it addresses all the requirements above except the two first ones, since its focus are not AI-based systems, but industrial systems, which in general combine both IT and OT assets. The KYKLOS 4.0 methodology allows the systematisation of the risk assessment process in a technology-, system-, and domain-agnostic manner. Therefore, by including some key adaptations required to address the specific needs of AI-systems, the new methodology fulfils all the requirements above. As explained in the following section, these procedural adaptations stress the identification and assessment of adversarial threats against AI components of the AI-systems and seamlessly integrate these activities in the whole process.

### *B. AI4CYBER security risk assessment methodology*

The challenge of security risk assessment of AI-based systems strives in the need of assessing risks to the overall system in an integrated fashion, evaluating the severity of both risks against the AI model(s) supporting some of the system tasks and risks against all the other system components as well. From the AI developer's perspective such surrounding components are seen as the or context of the AI, i.e., the environment in which the AI operates. However, depending on the nature of the AI system they may be even more crucial than the AI itself.

KYKLOS 4.0 continuous risk assessment methodology proposed a process for analysing and evaluating the security risks of an industrial system divided into two phases that are repeated iteratively. The process starts whenever updates are necessary due to certain triggering factors, specifically external and internal triggering factors.

AI4CYBER security risk assessment methodology adopts the same process approach and extends the resources and methods to be used in each of the phases of the methodology. Within the AI4CYBER approach, external triggering factors include the discovery and publication of new threats or vulnerabilities related to the technology and digital assets that are part of the AI system. Internal triggering factors occur when

new specifications require the integration of new components in the AI system, increasing the attack surface and the need for analysis of associated risks.

The initial phase of the risk assessment process follows a procedural approach to identify the necessary security controls that should be incorporated into the AI system. In the second phase, a more dynamic approach is taken, utilizing cybersecurity testing techniques, including adversarial testing techniques, to evaluate potential threats and vulnerabilities of the AI system.

Both phase 1 and phase 2 processes are defined by the original approach from KYKLOS 4.0 methodology but have been specified and extended to address specific AI systems-related threats, adversarial testing techniques, and mitigations. Figure 1 shows the AI4CYBER Procedural risk assessment process (phase 1) and Figure 2 presents the AI4CYBER Security testing-based risk assessment process (phase 2). The yellow elements in the figures are the newly added resources that are included in the AI4CYBER methodology. While the elements in purple are those that depend on the nature of the system to be evaluated, i.e. In this specific case, it is the Cloud Security Alliance's (CSA) Internet of Things (IoT) Security Framework [18] that addresses the specification of the security controls for IoT systems. In case the AI system includes IoT elements such as edge or cloud computing, the CSA IoT Security Framework will certainly be used as a complementary risk remediation analysis method.

The AI4CYBER AI Threat Knowledge Base (KB), one of the main contributions from AI4CYBER to the methodology, inherits from SPARTA AI Threat Knowledge Base [19] and extends it with the result of an extensive systematic literature review of Adversarial Machine Learning (AML) works since 2021 oriented towards identifying practical countermeasures to prevent AML threats. Furthermore, the new AI4CYBER AI Threat KB includes modernised taxonomies to accommodate latest updates in AI standards (including NIST's NIST AI 100-2e2023 on AML [9]), and mappings to TTPs and mitigations in ATLAS.

Additionally, AI4CYBER methodology complements the security testing-based threat assessment function in phase 2 with adversarial testing techniques such as Adversarial Robustness Toolbox (ART) [20].

### *C. AI4CYBER solution supporting the automation of AI system risk assessment*

Considering the large number of works studying adversarial machine learning threats and countermeasures, it is important to automate the search for all possible attack techniques and protections relevant to the AI components of the system. In this section we describe the solution proposed in AI4CYBER project [20] to support the automation of risk analysis for such components. This automation is based on the use of a dedicated tool that searches for AML threats and can execute the tests to verify the vulnerabilities against such kind of threats. The tool supports both phases of the analysis process in the methodology proposed, as shown in Figure 3. First, in

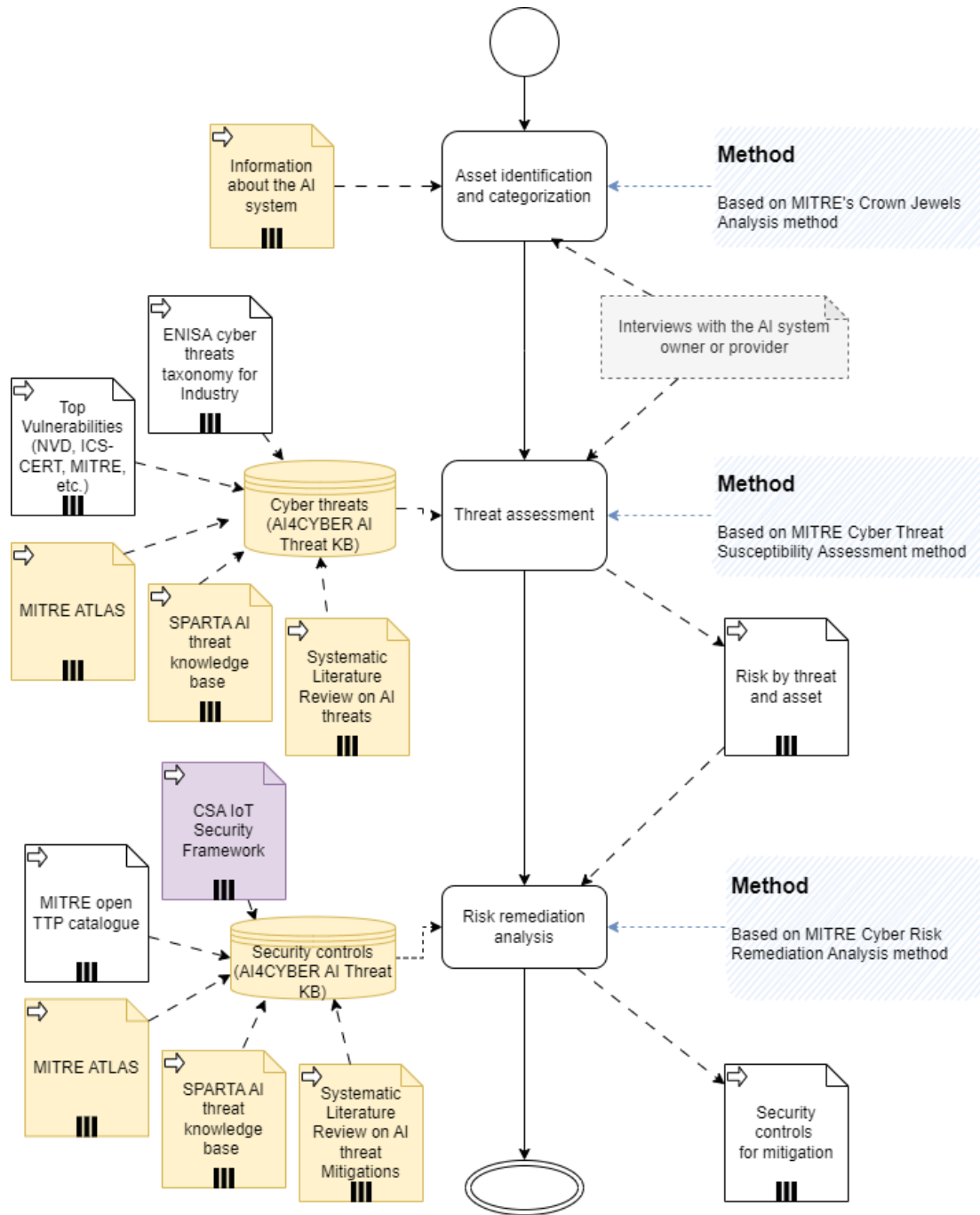


Fig. 1. AI4CYBER Procedural risk assessment process

the procedural assessment phase, the AML Threat Assessment module supports the finding of the known threats against the AI under study (the model used in the AI component under study). To this aim, the search is done on the AI4CYBER Threat KB, which structures AI threats identified in the literature and other public works against multiple classes of AI models. In addition, the AML Mitigation selection module is able to find the corresponding AML mitigations, since the threats in the KB are mapped to documented mitigations. Mit-

igations found are offered as possible controls to implement in the AI model. In this way, the AI system owner can decide on which protections will be implemented. Second, in phase two, the AML Testing service has the possibility of linking with other tools that execute the adversarial tests corresponding to the identified threats. The failing tests will inform the AI system owner of the poor efficacy of the implemented mitigations or the need of further mitigating the risks with additional controls.

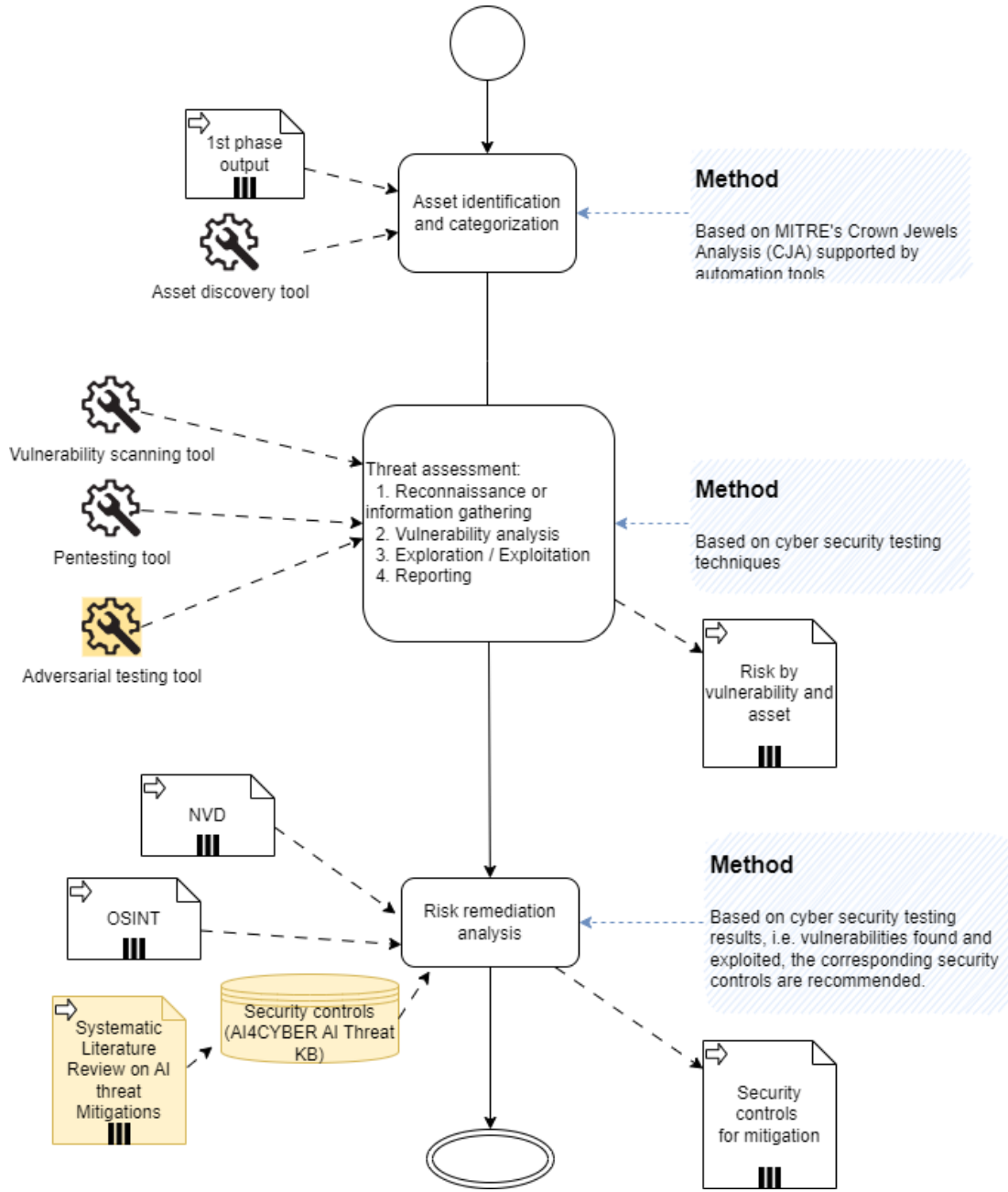


Fig. 2. AI4CYBER Security testing-based risk assessment process

#### IV. IMPLEMENTATION CONSIDERATIONS AND FUTURE DIRECTIONS

The present work runs in parallel with the latest advances in AML research which are extensively increasing in the last three years. Today, a search on google Scholar for “adversarial machine learning” publications since 2021 yields more than 7,730 references. The landscape and functionalities of AML testing tooling as well are rapidly growing, for example, the latest development of ATLAS includes an initial set of 19 TTPs, and the 18 mitigations therein are dated in April 2023. Both are likely to be refined in future months and additional

ones endorsed.

AI4CYBER plans include the validation of the methodology presented in AI-based cybersecurity solutions developed by the project. Such AI-based systems will be integrated in realistic industrial use cases in the following three domains: energy, banking and healthcare systems. The methodology will be used to continuously assess the risks of systems using different AI flavours, such as anomaly-based Intrusion Detection Systems that leverage federated deep learning architectures, user behaviour anomaly detection systems based on deep learning (DL), incident response systems that exploit reinforcement learning, software vulnerability analysis and testing systems

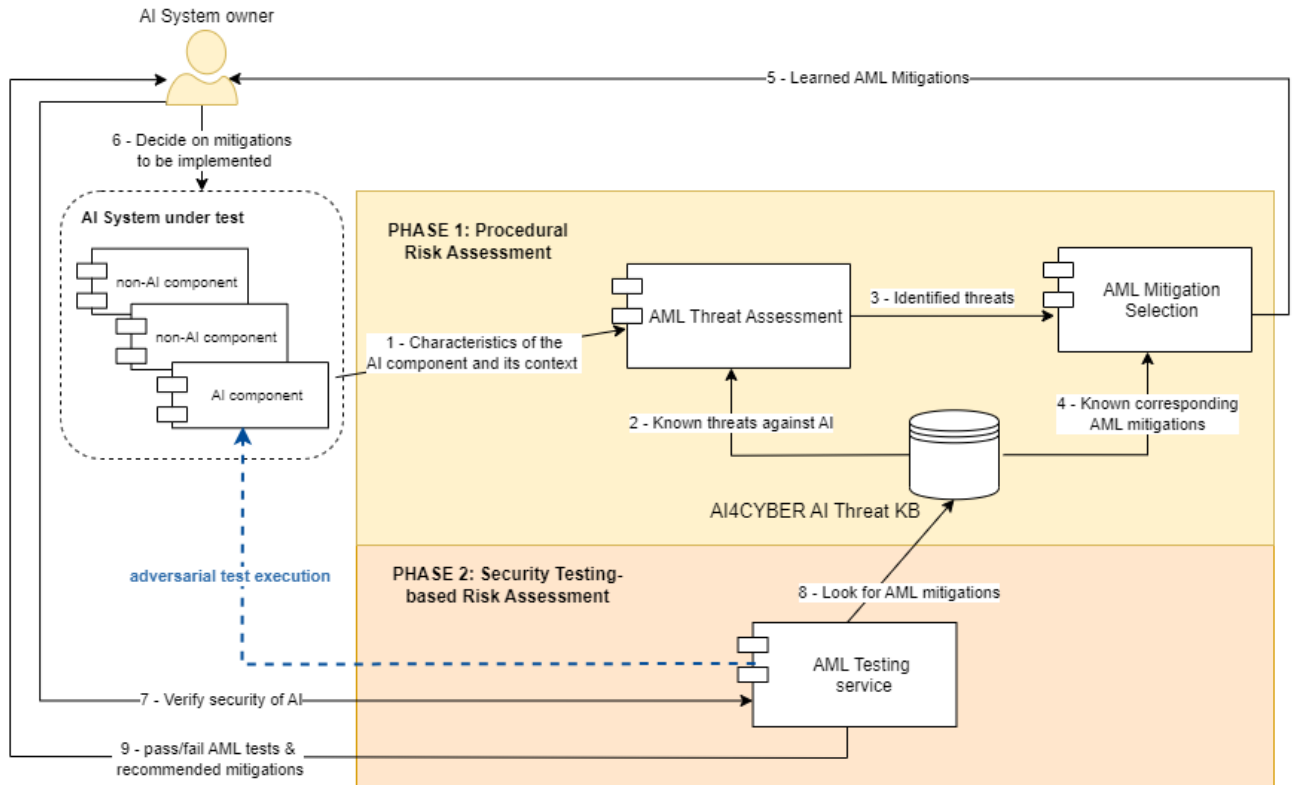


Fig. 3. AI4CYBER solution overview for the automation of AI system risk assessment

that include DL and natural language processing (NLP).

The research findings together with the final version of the AI4CYBER solution supporting the AI4CYBER security risk assessment methodology introduced herein, together with the AI4CYBER AI Threat Knowledge Base, will be released to the cybersecurity community to promote continuous risks assessment of AI systems and support the raising of AI robustness, and therefore, increasing trust in AI when integrated in systems and services.

## V. CONCLUSION

This paper presents a novel holistic security risk assessment methodology for AI-based systems, i.e. a methodology that supports the analysis and assessment of cyber risks in those systems that include some AI component fulfilling part of the system services.

The methodology addresses a set of eight requirements identified as relevant to pragmatically support the risk assessment task in such systems, namely, the need of addressing AI-relevant threats, allow assessing threats against all AI system (i.e., against both AI and AI-context), cover the whole DevOps engineering, enable continuous risk evaluation, systematic, reusable, agile, and facilitate automation.

The methodology proposed is an AI-system specific specialisation of the KYKLOS 4.0 methodology which was originally designed for security and privacy risk assessment of industrial systems.

The refinements of the methodology reside mainly in the fact that the cyber risks of all AI components in the AI system need to be identified and assessed from the adversarial learning perspective. This analysis would consider the poisoning, evasion and oracle threats an adversary could pose on the artificial intelligence methods used, which may lead to confidentiality, integrity and availability of data and services handled by the AI components.

The paper also described the tool support proposed by AI4CYBER project to facilitate the automation of the security risk assessment of AI components. The tool aids in both phases of the AI components' security risks assessment: the procedural assessment where automatically the threats and mitigations relevant for the AI under study are discovered, and the testing-based assessment, where the tool also supports the AML testing of the AI under study and the verification of the effectiveness of the implemented mitigations selected from first phase.

Both the methodology and tool presented herein are planned to be validated in a dual iteration cycle in the industrial use cases of the AI4CYBER project, where the risks of AI components of cyber resilience solutions developed, such as federated deep learning-based intrusion detection systems, will be evaluated.

## ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon Europe research and innovation programme under

grant agreement No 101070450 (AI4CYBER).

## REFERENCES

- [1] I. Lella, E. Tsekmezoglou, R. S. Naydenov, C. Ciobanu, A. Malatras, and M. Theocharidou, “ENISA threat landscape 2022,” Tech. Rep., Oct. 2022.
- [2] European Commission, “EUR-Lex - 52021PC0206 - EN - EUR-Lex,” <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>, Apr. 2021, accessed: 2023-10-06.
- [3] European Parliament, “EU AI act: first regulation on artificial intelligence,” <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>, Aug. 2023, accessed: 2023-10-06.
- [4] High-Level Expert Group on Artificial Intelligence, “ETHICS GUIDELINES FOR TRUSTWORTHY AI,” Tech. Rep., Apr. 2019.
- [5] J. T. Force, “Risk management framework for information systems and organizations,” *NIST Special Publication*, vol. 800, p. 37, 2018.
- [6] NIST, “AI RMF development — NIST,” <https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development>, 2021, accessed: 2023-10-06.
- [7] E. Tabassi, “Artificial intelligence risk management framework (ai rmf 1.0),” 2023.
- [8] NIST, “Roadmap for the NIST artificial intelligence risk management framework (AI RMF 1.0) — NIST,” <https://www.nist.gov/itl/ai-risk-management-framework/roadmap-nist-artificial-intelligence-risk-management-framework-ai>, 2023, accessed: 2023-10-06.
- [9] A. Oprea and A. Vassilev, “Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (draft),” National Institute of Standards and Technology, Tech. Rep., 2023.
- [10] L. S. Metzger, G. Rebovich Jr, R. A. Cormier, S. J. T. Norman, D. L. Schuh, P. A. Smyton, R. S. Swarz, and F. C. Wendt, “Systems engineering guide: Collected wisdom from mitres systems engineering experts,” MITRE CORP BEDFORD MA BEDFORD United States, Tech. Rep., 2014.
- [11] MITRE, “MITRE ATT&CK®,” <https://attack.mitre.org/>, accessed: 2023-10-06.
- [12] —, “MITRE ATLAS,” <https://atlas.mitre.org/>, 2023, accessed: 2023-10-06.
- [13] N. Polemi and I. Praça, “A multilayer framework for good cybersecurity practices for AI,” Tech. Rep., 2023.
- [14] A. Malatras, I. Agraftiotis, M. Adamczyk, and ENISA, “Securing machine learning algorithms,” Tech. Rep., 2021.
- [15] “AI risk database,” <https://airisk.io/>, accessed: 2023-10-06.
- [16] “Continuous validation — robust intelligence,” <https://www.robustintelligence.com/platform/continuous-validation>, accessed: 2023-10-06.
- [17] E. Iturbe, E. Rios, J. Mansell, and N. Toledo, “Information security risk assessment methodology for industrial systems supporting ISA/IEC 62443 compliance,” in *III. International Conference on Electrical, Computer and Energy Technologies (ICECET 2023)*, forthcoming 2023.
- [18] Cloud Security Alliance, “CSA IoT security controls framework v2,” <https://cloudsecurityalliance.org/artifacts/csa-iot-security-controls-framework-v2/>, Jan. 2021, accessed: 2023-10-06.
- [19] E. Rios, M. Knockaert, S. E. de Velp, M. R. Norouzian, C. Palacios, C. Martínez, R. Orduña, X. Etxeberria, A. Gil, M. Pawlicki, and M. Choras, “Strategic programs for advanced research and technology in europe (SPARTA): D7. 1: AI systems threat analysis mechanisms and tools,” Tech. Rep., 2021.
- [20] “Adversarial robustness toolbox (ART),” <https://adversarial-robustness-toolbox.org/>, accessed: 2023-10-6.