CoEvolution: A comprehensive trustworthy framework for connected machine learning and secure interconnected AI solutions

Antonios Makris[1], Apostolos Fournaris[2], Anita Aghaie[3], Ioannis Arakas[4], Anna Maria Anaxagorou[15], Ioannis Arapakis[5], Davide Bacciu[6], Battista Biggio[7], Georgios Bouloukakis[8], Stavros Bouras[1], Arne Bröring[3], Antonio Carta[6], Marco Caselli[3], Olympia Giannakopoulou[15], Nikolaos Gkatzios[9], Alexandros Gkillas[10], Evangelos Haleplidis[2], Sotiris Ioannidis[11], Eleni-Maria Kalogeraki[12], Panagiotis Karantzias[12], Emmanouil Kritharakis[1], Aris Lalos[10], David Lenk[13], Stella Markopoulou[9], Entrit Metai[11], Andreas Miaoudakis[15], Haralambos Mouratidis[12], Jihane Najar[9], Theodor Panagiotakopoulos[10], Bernhard Peischl[13], Maura Pintor[7], Nikos Piperigkos[10], Vassilis Prevelakis[9], Carlos Segura[5], Georgios Spanoudakis[4], Orestis Tsirakis[15], Omar Veledar[14], and Konstantinos Tserpes[1]

[1] *School of Electrical and Computer Engineering, National Technical University of Athens, Greece*
[2] *Industrial Systems Institute, Research Center ATHENA, Patras, Greece*
[3] *Siemens AG, Munich, Germany*
[4] *Sphynx Technology Solutions, Zug, Switzerland*
[5] *Telefónica Scientific Research, Barcelona, Spain*
[6] *Computer Science Department, University of Pisa, Italy*
[7] *Department of Electrical and Electronic Engineering, University of Cagliari, Italy*
[8] *Télécom SudParis, Institut Polytechnique de Paris, France*
[9] *AEGIS IT RESEARCH GMBH, Braunschweig, Germany*
[10] *AViSense, Patras, Greece*
[11] *DIE, Crete, Greece*
[12] *Security Labs Consulting, Cork, Ireland*
[13] *AVL List GmbH, Graz, Austria*
[14] *Beevadoo e.U., Graz, Austria*
[15] *CBRL, Nicosia, Cyprus*

***Abstract*—The contemporary AI landscape demands a holistic framework to ensure security across the entire AI supply chain and lifecycle. Despite the availability of existing adversarial attack techniques, an end-to-end solution for identifying threats, vulnerabilities, and risks is still lacking. Despite EU initiatives like the AI Act promoting safety and trustworthiness in AI, it lacks a system for managing weaknesses within a networked AI supply chain. This paper introduces CoEvolution, which aspires to address this gap by implementing a complete Security, Trust, and Robustness (STR) assessment solution, capable of addressing evolving AI cybersecurity threats. CoEvolution proposes a universal hub for STR risk assessment and security assurance, aligned with MLDevOps practices and EU AI regulatory frameworks. It introduces innovative AI model descriptions, including an AI Model Bill of Materials, coupled with security monitoring and context awareness. CoEvolution seeks to ensure compliance with EU directives on trust, fairness, data governance, and GDPR guidelines.**

***Index Terms*—adversarial attacks, robustness, security, ai model bills of material, threat models, risk assessment.**

I. Introduction

As the artificial intelligence (AI) market accelerates to over $2575.16 billion by 2032 [1], the landscape is evolving from siloed AI solutions to more collaborative systems (such as federated learning or peer-to-peer learning) and to off-the-shelf, pre-trained third-party AI models that are used as is, with no security guarantees. These interconnected and black-box-based AI solutions promise seamless data flows, models, and services across different sectors, leading to optimized decision making and tailored responses to dynamic situations. However, the seamless integration and collaborative capabilities of AI systems introduce significant vulnerabilities. Adversarial attacks, data poisoning, and man-in-the-middle attacks can compromise the integrity and functionality of these systems. In addition, the AI lifecycle, from data collection to design to deployment, becomes vulnerable to these threats, making the entire AI ecosystem vulnerable.

As the domain of AI use evolves and embraces a more connected and collaborative paradigm, several unprecedented threats and challenges are emerging. At the heart of these threats is the increased risk posed by the integration of multiple AI systems. Such integration, while promising in terms of collaborative intelligence and enhanced functionality, introduces potential points of compromise. For instance, evasion attacks become more insidious in this context [2]. If one system in the network can be fooled by an adversarial input, it could

inadvertently cascade misinformation to other systems in the network/collaboration. Similarly, the shared data ecosystem within these networked systems increases the risk of poisoning attacks, where malicious data introduced into one module can percolate and affect the integrity of others.

In addition to the above, additional vulnerabilities of AI models to adversarial attacks can be introduced when models are unaware of their environment. Without an understanding of the broader context in which they operate, AI systems rely solely on individual data points to make decisions, making them vulnerable to manipulation that exploits these isolated views. Without a comprehensive view of the operating environment, AI models may miss patterns or anomalies that only become apparent when the larger context is considered [3], [4]. Adversarial attacks, particularly those involving subtle input biases, can easily go undetected in such situations. Contextual awareness is essential to provide a holistic perspective that enables models to effectively distinguish genuine variations from adversarial intrusions.

Given the complex landscape of adversarial attacks described above, AI security experts seem to have too many options to consider. On the one hand, the widespread use of AI has created a boom in AI security where a constant stream of adversarial attack variations is provided, while, on the other hand, defences are overly specialised to specific attacks, not formally verified, and lacking in accountability. Although there are some frameworks with assessment of adversarial and poisoning attacks as well as a set of defences, there is still no coherent security assessment flow on how to holistically identify AI security vulnerabilities and risks, as well as provide coherent defences applicable to the entire AI design and development lifecycle regardless of the AI model paradigm in a supply chain. Google has just recently published the Secure AI Framework (SAIF)[1], a conceptual framework for secure AI systems that effectively identifies AI security issues as highly critical, especially in the AI supply chain.

Finally, from a regulatory perspective, to harness the potential of AI without compromising security, there is a significant need for international standards that can guide the development, deployment, and continuous improvement of these AI systems. Such a framework would pave the way for more responsible, secure, robust, ethical and universally accepted AI practice, ensuring that the power of AI in a supply chain is harnessed without compromising security or ethics. To this end, after a multi-year process and several guidance documents, the EU and its member states are making intensive efforts to regulate AI through frameworks like the European Commission's AI Act[2], aimed at ensuring both safety and trustworthiness in AI applications across all industries. This regulation highlights the need for an AI paradigm shift, where each AI model will be accompanied by specific characteristics, risk values, and formal descriptions that will enable transparency, security evaluation, and trust.

In this context, the EU-funded RIA project CoEvolution[3] aims to address key challenges in AI security by developing an integrated framework for Security, Trust, and Robustness (STR) assessment. The core innovation of the project lies in the creation of an end-to-end STR hub, utilizing AI Model Bills of Materials (AI MBOM) to enhance transparency and security assurance throughout the AI supply chain. By aligning with MLDevOps and EU regulatory frameworks—including directives on trust, fairness, data governance, and GDPR guidelines—CoEvolution establishes a structured approach to AI risk assessment. Its architecture encompasses novel AI model descriptions, AI MBOM management, security monitoring, and context-awareness. In addition, CoEvolution promotes the development of open-source trusted datasets and AI models, fostering a robust, adaptable framework for risk analysis and security assessment in AI-driven ecosystems, ensuring alignment with evolving AI cybersecurity threats.

II. Ambition

Ensuring the security, trust, and robustness of AI systems requires a multifaceted approach that addresses emerging threats, enhances resilience, and fosters adaptive intelligence. This section presents CoEvolution's key ambitions, focusing on strengthening AI security through novel methodologies for adversarial defense, decentralized learning, and context-aware decision-making.

A. Detection, Prevention and Mitigation of Adversarial Attacks on AI models

Data-driven AI and ML techniques have shown to be successful in many different applications, including computer vision, speech recognition, and cybersecurity-related domains like malware detection. Despite the deployment of many commercial solutions relying on AI/ML algorithms, these algorithms remain vulnerable to well-crafted attacks. Potential threats against AI / ML are well understood at the research level [5], [6], and can be categorized depending on whether the attacker has the ability to influence the training data used to learn the AI/ML model, or only the test data that will be provided as input to the algorithm during operation. Attacks staged at training time include, but are not limited to, poisoning attacks. Poisoning attacks compromise the learning algorithm by injecting carefully crafted poisoning samples in its training data to cause either a denial of service (i.e., a substantial increase in the test error that makes the system unusable for legitimate users) [7], or specific intrusions/misclassifications at test time (e.g., via backdoors that can be activated at test time by specific input triggers) [8]. Additionally, other cyberattacks can occur during training, such as model extraction, which aim to steal or manipulate the model without necessarily altering the training data.

During operation, attacks can include evasion and privacy attacks, as well as data poisoning in dynamic systems. Evasion attacks (a.k.a. adversarial examples) target a trained

[1]https://saif.google/

[2]https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[3]https://cordis.europa.eu/project/id/101168560

model by perturbing the input data at test time to induce misclassification (e.g., modifying digital input images to be misclassified by an AI/ML system) [9]. Finally, privacy attacks can extract confidential information either about the system itself or about its users by iteratively querying a machine-learning system, e.g., offered as an online service [10]. Even though the research community has largely demonstrated that AI/ML algorithms are vulnerable to different attacks, a more thorough analysis of their practical impact and feasibility, together with proper testing tools to help AI/ML developers quantify the associated risks, is still lacking, especially when such models need to be continuously deployed and updated. The problem of mitigating such attacks from the AI/ML algorithmic viewpoint remains also very challenging, as they undermine the fundamental stationarity assumption of AI/ML algorithms, i.e., that training and testing data are drawn from the same distribution. Although the research community has produced many tools to optimize attacks against AI/ML models (e.g., FoolBox, Cleverhans, IBM ART, SecML), they require strong knowledge of the AI/ML security problems to be used correctly and do not provide accessible user interfaces. In addition, the implemented attacks and defenses are typically very computationally demanding, hindering their applicability in efficient industrial processes.

CoEvolution aims to overcome these issues and facilitate the deployment of more reliable, secure, and trustworthy AI/ML models in industrial/automotive software development processes. It will provide solutions to easily integrate existing ML testing libraries and defensive strategies to mitigate the impact of the aforementioned attacks into current software development lifecycles while making them more efficient and easier to configure and automate.

B. Adversarial attacks on P2P and FL collaborative learning

Traditional machine learning systems are based on a centralized design where all data is transferred, stored and processed in the cloud, and the user endpoints could later fetch the results via API calls. While this approach provides good performance from an ML perspective, it is extremely costly from a system point of view (i.e., storage, processing, bandwidth), especially when the cloud provider also becomes the single point of failure. Further, it does not address successfully the significant privacy concerns, especially considering the sheer volume of the multidimensional and privacy-sensitive user data the modern mobile and IoT devices collect nowadays.

To mitigate these concerns, the distributed ML paradigm of Federated Learning (FL) [11] came to enable mobile devices to produce an ML model without sharing any user data. Instead, in FL, users selectively share subsets of their models' key parameters during training. However, FL requires the coordination of a central server that orchestrates the training, the client's selection, the updates collection, and weights aggregation. This server constitutes a centralized authority that (1) users are forced to blindly trust to protect them from various information leakage attacks , while at the same time (2) such centralized service is a single point of failure, which can become a bottleneck impacting system scalability and performance . To address these emerging issues, existing approaches either use differential privacy [12], thus decreasing the utility of the shared gradients, or specialized hardware (e.g., Trusted Execution Environments). Furthermore, previous peer-to-peer learning systems that do not require a centralized server heavily rely on blockchain ledgers [13] and the Interplanetary File System (IPFS) [14], or alternatively, use differential privacy.

CoEvolution proposes P4L, a novel peer-to-peer learning system that operates without relying on differential privacy or external systems such as blockchain or IPFS. In particular, P4L is a privacy preserving, fully distributed and infrastructure-less peer-to-peer learning system that enables users to participate in an asynchronous decentralized collaborative learning scheme. Specifically, in P4L, users will be able to train their models locally by using their (privacy-sensitive) data and create asynchronous, small, collaborative learning synergies with nearby users, via which they will share their model's gradients. In P4L, we will utilize partial Homomorphic Encryption (HE) during gradient averaging to preserve both the confidentiality and utility of the shared gradients, while at the same time eliminate information leakage attacks. In addition, P4L will utilize the proximity and cross-device communication capabilities of mobile devices, thus operating on infrastructure-independent environments, without requiring centralized federations, Public Key Infrastructure (PKI), or an internet connection. A truly open architecture for federated learning, and even more ambitiously, fully distributed peer-to-peer learning, would enable devices and applications to collaboratively build better and more resilient models by incorporating larger, more diverse datasets and leveraging contextual awareness, while safeguarding security and privacy.

C. Context-aware decision support for threat mitigation using multi-objective criteria

Reinforcement learning (RL) has proven successful across various domains, including robotics, smart systems, and chip design [15]–[17]. In the traditional RL setting, a policy interacts with the environment to maximize a single cumulative reward (in model-based RL algorithms, a simulator is employed to model the environment, allowing the policy to learn through interactions with the simulated environment [18]). However, numerous real-world problems involve multiple, potentially conflicting objectives. For instance, in robotics tasks, there is a compromise between maximizing speed and energy consumption. Unlike single-objective scenarios, these environments are characterized by the measurement of performance using multiple objectives, resulting in multiple Pareto-optimal solutions based on objective preferences [19]. In these scenarios, multi-objective reinforcement learning (MORL) approaches [20] aim to maximize a vector of rewards contingent on specified preferences. Existing MORL algorithms can be broadly categorized into two main groups [21]: single-policy methods and multiple-policy methods. Single-policy methods focus on identifying the optimal policy based on specified preferences

among the objectives [22]. However, single-policy methods face limitations when preferences are unknown. Multi-policy approaches target the approximate Pareto frontier of optimal solutions. However, these methods explicitly maintain sets of policies, making it difficult to scale up to high-dimensional preference spaces. Also, these methods cannot be easily adapted to new preferences at test time.

CoEvolution casts the context-aware decision support task as a recommendation problem. The role of the context-aware decision support system will be to choose the optimal mitigation action considering candidate actions selected at the threat intelligence stage and context vector. Concretely, CoEvolution plans to implement a multi-objective decision support framework using Reinforcement Learning or Decision Transformer (DT) algorithms. In said framework, each action affects multiple key performance indicators (KPIs) of the system, and the changes can be mapped to positive and negative rewards, required to train a decision-making agent. The multi-objective dimension aims to integrate the utility functions of the individual objectives in place into a single reward function (e.g., weighted sum of the utility functions of throughput, latency, resource utilization, energy efficiency). As such, a central idea of this approach is to couple these rewards to target KPIs while accounting for context information.

III. Overall concept

The CoEvolution is envisioned as a multilayer hub of building security and trust into AI solutions by offering robustification tools and services, security assessment and assurance mechanisms, trusted AI models and dataset pools, as well as capacity building mechanisms that instil security, robustness and trustworthiness to AI models. This approach is considered from three different viewpoints, security of the whole AI lifecycle, security of the full AI ecosystem including individual AI models, collaborative learning (Federated Learning, decentralized Learning etc.) paradigm models, as well as interconnected AI model agents in an AI supply chain and the provision of security assurance of all AI model types.

From the AI lifecycle viewpoint, CoEvolution provides security enhancement and robustification in all steps of the AI model lifecycle. This includes security measures to be included for the model design phase, the model training/testing/validation phase (including mechanisms to evaluate the associated data), as well as the deployment and operation phase of an AI model (when performing inference on a given setting). All defences included in the AI model will be part of a Defence Flow, a continuous approach from the design to the deployment/operation phase, complementary to each other based on the security target goals. Eventually, CoEvolution envisions the whole AI model lifecycle to be approached through a security-by-design, privacy-by-design and explainability-by-design perspective that will result into AI models with enough associated security assurance guarantees that render them trusted. Through the CoEvolution approach, an AI model is infused with context awareness throughout its entire lifecycle, enabling it to respond dynamically based on the environment in which it is deployed. Additionally, techniques are introduced that allow the AI model to monitor its own security status, detect potential attacks, and report these incidents to the CoEvolution security runtime monitoring system.

From the AI supply chain viewpoint, CoEvolution acknowledges the AI ecosystem diversity, where AI models are designed and used in various ways. Thus, CoEvolution considers AI adversarial attacks (and providing resistance against them) for individual, single, AI models that operate autonomously in a given system but also for collaborative AI model approaches that operate under a Federated Learning or a distributed, decentralized, Peer-to-Peer Learning. Finally, CoEvolution also provides the mean to secure the whole AI supply chain which may include interconnection of AI models, (interconnected AI agents) that cooperate collectively by exchanging inference results towards a given cause.

From the AI Security and Trust Assurance viewpoint, CoEvolution aims to provide AI models through its security enhancements and robustification operation flows (the CoEvolution security assurance flow) that will be considered secure/trusted. To achieve that, CoEvolution provides a Security Assurance mechanism that records all the assessment and robustification activities that an AI model and associated data undergoe in order to become secure/trusted and generates a security credential that accompanies the AI model throughout its lifecycle. The CoEvolution security assurance flow also provides a management mechanism for AI model credentials, which includes not only credential generation but also credential updating and revocation.

The above three viewpoints in CoEvolution are supported by a universally used structure that is associated to an AI model, the AI Model Bill of Materials (AI MBOM). Adopting the logic of the Software Bills of Materials (SBOM) accompanying software solutions, the AI MBOM provides detailed information about an AI model including the dataset used to train it, the owner and users, its dependencies to ML/DL libraries etc., but most importantly, it incorporates all security related information associated with the use of the CoEvolution solution. The AI MBOM acts as the vessel of the CoEvolution security assurance generated credentials and follows the lifecycle of AI models. It is managed by CoEvolution in an effort to create the necessary momentum for its broad adoption by the AI community as a mechanism of reporting robustness, security, trust, explainability and fairness of an AI model. Existing approaches, such as the Machine Learning Bill of Materials (MLBOM) by CycloneDX[4] and SPDX 3.0 specification[5], an international open standard under the Linux Foundation, will be considered.

To achieve context awareness and continuous adaptation, AI models in CoEvolution are equipped with a unique class of awareness – awareness of the environment and its actors via the provision of an embedding space of cross-modal

[4]https://cyclonedx.org/capabilities/mlbom/

[5]https://spdx.dev/implementing-an-ai-bom/

descriptors - and the capacity to react and interact with its environment, via intelligent adaptations that optimise relevant reward functions.

Both the security-by-design and context awareness features of the solution contribute to the long-term goal of CoEvolution acting as an enabler to a concrete ecosystem around security and trust in AI solutions that captivates the different aspects of ML/DL research and development through their whole lifecycle. CoEvolution enables the exchange of AI security information using the novel AI MBOM approach, provides the support for exchange of trusted models and datasets (with security guarantees) and delivers the tools and methods to assess, harden, assure and monitor deployed AI models from a security perspective (under a series of adversarial threat models), via the Defence Flows and context awareness, in popular use cases.

IV. System Architecture

The concept of the CoEvolution hub, along with its capabilities and viewpoints, is manifested in the conceptual architecture illustrated in Figure 1.

The AI *Risk Assessment and Guidelines/Recommendations Engine*, drawing from a cybersecurity threat intelligence backbone, provides AI-associated risks for a given model design, including risks related to the model's chosen design structure, employed type, training/testing/validation, and its deployment within an AI supply chain (Federated Learning, Decentralized/Distributed/Peer2Peer, or interconnected AI agent approaches). Based on the discovered risks, the engine provides actionable insights on risk mitigation associating each risk with recommended mitigation techniques and tools from the CoEvolution Security Trust and Robustness Defense Framework.

In association with the above risk engine, the *Security Testing/Assessment engine* discovers existing vulnerabilities on a given AI model design (to be trained or pre-trained), as well as data bias that can be used for data poisoning attacks on datasets that are used for training/testing/validation. Similarly, to the AI *Risk Assessment and Guidelines/Recommendations Engine*, the *Security Testing/Assessment Engine* offers recommended AI model robustification/hardening techniques and tools to mitigate the discovered threats. These two engines draw information from existing open source knowledge bases on AI supply chain related risk and/or adversary attacks/vulnerabilities, such as the AI Risk Database[6] and MITRE ATLAS[7]. Eventually, the engines upon discovering new vulnerabilities on existing, publicly available AI models they are capable of reporting such vulnerabilities to such knowledge bases. It should be mentioned that the above two engines produce reports that are included in the AI MBOM of an assessed AI model.

The Core of CoEvolution is the *Security Trust and Robustness (STR) Defence Framework* that consists of methods, libraries, trained models, tools and services capable of increasing the resilience / robustness of an AI model, as well as removing data noise or bias on AI training/testing/validation data in all phases of an AI model/data lifecycle. The Framework solutions are split into three categories, the design phase, the training phase and the deployment phase libraries/tools. For each phase we are providing solutions that consider the various AI design paradigms including individual AI model robustification and adversarial attack prevention, detection and mitigation, Collaborative AI robustification (FL and P2P Learning) and also Interconnected AI model agents that exchange inference results between them.

Eventually, when an AI model is designed, trained and deployed in an actual system following the CoEvolution Security assurance flow (supported by the STR Defence Framework), it integrates one or many CoEvolution defence flows into its structure. A defence flow is envisioned as a series of complementary defence techniques on each AI lifecycle phase and across phases in the AI model robustification process that adopt a common AI core technology (e.g. Ensemble Learning). At deployment phase the defence flow is realized as several "gadgets" that are added to the AI model structure and they activate the AI model's robustness self-awareness and context self-awareness. These CoEvolution gadgets allow the robust CoEvolution enabled AI model to detect inference related adversarial attacks (e.g. poisoning attacks) and report it to the *Security Runtime Monitoring Engine* but also to dynamically adapt itself to the context in which it operates. Hence, the STR Defence Framework will include proper context-awareness mechanisms that will eventually be infused (at the deployment phase tools and services) into the final robust AI model.

The CoEvolution context-awareness approach is diversified based on the architecture paradigm (single, collaborative, interconnected) followed for an AI model. It can follows a static ruleset of context diversification (for single AI models) or a dynamic, evolution-based context adaptation (for the collaborative and interconnected paradigm) that involves dedicated "gadgets" based on reinforcement learning in the robust AI model, as well as a *Central Context-Awareness Engine* that collects context input vectors and using evolutionary algorithms generates new, more relevant to the condition at hand, context vectors. In CoEvolution, context is acknowledged through information related to adversarial attacks at a given time and space, as well as data (and metadata) collected from any in-field use case (e.g., sensors, network, etc.) where an AI model is deployed.

As CoEvolution robust AI model driven applications are deployed, the deployment phase "gadgets" for the models' security-awareness, infused by the deployment phase solutions of some defence flow, take effect and provide feedback to the AI *Security Runtime Monitoring Engine (SRME)*, offering adversarial cyberattack surveillance of AI operations. The goal of this engine is to provide to an operator the means to respond to a possible adversarial attack focused on the runtime operation of an AI model. Such inference focused attacks can be characterized as anomalies by the AI models and trigger an event to the SRME that will then inform its users and also provide updates to the AI MBOM associated with the attacked

[6]https://airisk.io/

[7]https://atlas.mitre.org/

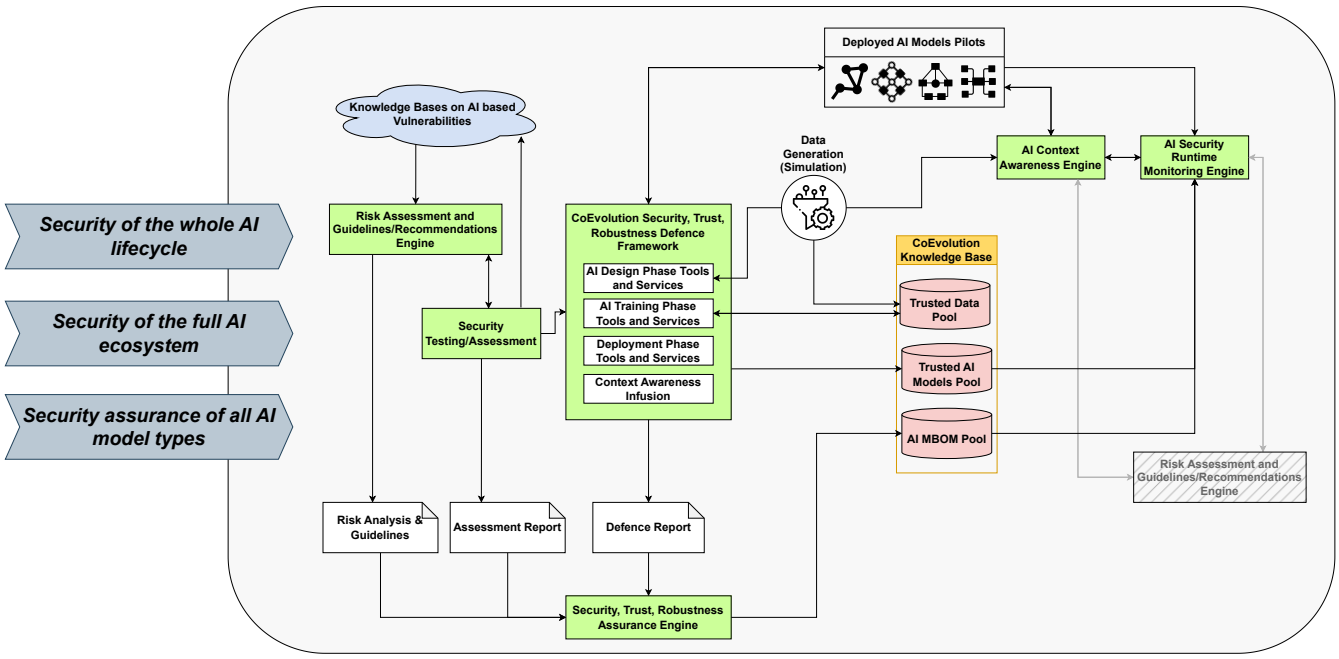

Fig. 1. The CoEvolution conceptual architecture

AI model (e.g. model structure, dataset metadata, discovered security/adversarial attack vulnerabilities, etc.). This engine is also providing a trust evaluation of any AI model associated with the CoEvolution hub. Through its association with the AI MBOM pool where all the generated/managed AI MBOMs are stored, the SRME is acting as a AI root of trust able to answer questions related to the validity, integrity and authentication of existing AI MBOMs associated with an AI model and with the update or revocation of an AI MBOM when an AI model is marked as compromised (based on information collected from the deployed AI model gadgets).

As expected, the CoEvolution hub supports a knowledge base with three data pools. The first data pool includes trusted AI models that have passed through the security scrutiny process of CoEvolution and include all the robustness, security, trust and context awareness features. These models are offered under an open access license (they may also rely on public AI models) to the AI community and aim at enhancing the CoEvolution ecosystem. Similarly, CoEvolution hosts a trusted Data pool that include open access trusted data that have been security assessed using the CoEvolution solution or they have synthetically been generated in a trusted manner using the CoEvolution Simulator component (synthetic data generator). Finally, the third data pool of the CoEvolution knowledge base is related to the management of the AI MBOMs that have been generated by the CoEvolution framework. This pool is accessed by the CoEvolution security assurance engine that generates new AI MBOMs for any CoEvolution secure/trusted, robust, context aware AI model in order to store such MBOM but it is also accessed by the CoEvolution SRME that updates, verifies or revokes an AI MBOM based on the operational status of the associated AI model.

The CoEvolution Security Assurance Process acts as an orchestration component of the overall CoEvolution Risk and Security assessment, and Security robustness/context awareness process of the CoEvolution hub. This process coordinates the above activities, aligns them with the guidelines and standards on AI trustworthiness and fairness, collects the various other engines reports and generates, based on those reports, the AI MBOM for a given assessed/secured/robust AI model.

V. Pilots, Use cases and main challenges

Below, the use case applications, upon which the CoEvolution hub will be tested, are described and their main challenges are identified.

A. Pilot 1: Securing the AI Supply Chain of Manufacturing Infrastructure

In industrial automation, AI components are increasingly important and are used to automate various tasks, such as process management, machine control, and logistics, as well as to enable predictive maintenance of deployed hardware. Often these AI components are brought into the manufacturing environment as part of applications that are executable in an edge computing infrastructure. These applications can be downloaded from marketplaces of edge ecosystems. An application can thereby be provided to the marketplace by third-party vendors. The AI component of the application may have been even developed by another subcontracted party that is supporting the third-party vendor. Hence, for the end-customer—the operator of the manufacturing plant—it is crucial to have security mechanisms in place to ensure

trust in the application and its embedded AI component. Especially, when applications are part of essential workflow chains (consisting of multiple applications), it is of utmost importance to ensure their security and expected behaviour.

CoEvolution focuses on connected AI components that can be used to manage the logistics of the manufacturing environment as well as the management of the network and edge infrastructure itself. An AI model is assumed to arrive (e.g., as part of an edge application) from a third-party vendor or another facility. It may either be a pretrained AI requiring a transfer learning step for adaptation to a new environment or be used as is.

The implementation of such a use case involves the following components: i) Network/Edge Management AI: This AI model supports the management of the network / edge infrastructure, and can, for example, be utilized to efficiently manage edge resources and installed applications, ii) Network switch devices: These devices allow the creation of connected edge networks, iii) Edge device: This is the device on which an AI model will be deployed, iv) Localization AI: This AI model supports the logistics of the manufacturing environment, and can, for example, be utilized to improve the localization accuracy of tracked assets, v) Localization beacon devices: These devices enable wireless localization of tracked assets and vi) Localization tag devices: These devices are being localized by the beacons and carried by the tracked assets.

Ensuring trust in the various AI components is crucial. Adversarial attacks targeting localization accuracy could lead to misplacement or misrouting of critical assets, directly impacting production timelines and operational efficiency. Similarly, attacks compromising edge resource management can disrupt the allocation and functionality of network resources, leading to system downtimes and reduced throughput. Such vulnerabilities not only threaten productivity but also pose significant financial risks, underscoring the need for robust security measures within industrial AI deployments.

B. Pilot 2 Securing the AI Supply Chain in the Autonomous Vehicles Domain

Connected and Automated Vehicles (CAVs) utilize a multitude of onboard sensors, including camera, radar, Light Detection and Ranging (LiDAR), ultrasonics, and GPS to establish a comprehensive understanding of their surroundings. However, the employed sensors have considerable inherent limitations (e.g. weathering and/or lighting conditions degrade camera performance, while active sensors are adversely affected by high humidity, occlusions, and interference), which can severely degrade CAVs performance. In addition, in real-world scenarios, the amount of local data is insufficient to derive resilient and robust learning models, due to the extremely complex and dynamic nature of the CAV's environment. Cooperative perception aims to overcome these challenges by enabling a group of interacting agents to share knowledge, thereby enhancing CAVs' environmental modeling capabilities, extending their local sensing range, and improving the system's resilience to sensor failures and attacks.

Despite advances, CAV perception systems remain vulnerable to adversarial attacks, including LiDAR jamming and GPS spoofing, which can disrupt navigation and control. This use case will test CoEvolution's cybersecurity engine in multi-agent scenarios involving CAVs that cooperate to establish autonomous navigation and co-operative multi-modal four-dimensional scene awareness in the vicinity of the vehicle. To this end, the use case will adopt early (sensory inputs) and late (AI-leveraged knowledge) fusion approaches to introduce a holistic unified framework for co-operative awareness and Multi-Agent-Path Planning (MAPP), jointly interconnected in an iterative loop. For the co-operative awareness phase, CoEvolution will integrate data derived by both mobile and static Interacting Traffic Agents (ITAs), where static ITAs can consist of both sensors lying on the infrastructure as well as landmarks and active sensors, acting as landmarks. In alignment with CoEvolution's principles and objectives, a trustworthy, robust and context-aware cyber resilience approach will be developed, where the concept understanding will be derived by cooperative scene understanding and path planning. Two different sub use cases will be developed, one targeting the vehicle's perception and the other vehicle's path planning and control. In the perception sub-use case, CoEvolution will assess the impact of data poisoning attacks on 3D LiDAR point clouds, focusing on their effects on key perception tasks such as 3D object detection, tracking, segmentation, and SLAM [23]. To counter these attacks, computationally efficient defense mechanisms will be developed using novel and explainable frameworks based on model-based deep learning theory [24]. These frameworks will introduce preprocessing models, such as denoising or super-resolution, designed to enhance the quality of input data and mitigate adversarial noise before it is fed into the perception networks. In addition to that, an effective mechanism for robustifying GPS's reliable operation against spoofing and jamming outliers [25] will be developed. For a unified defense mechanism of perception engine, focusing on securing both LiDAR and GPS against adversarial attacks and outliers, model-based deep learning tools and approaches will be exploited through a dynamic, online and continual collaborative decision-making and learning framework, leveraging the benefits of multi-agent coooperation.

Both sensing and planning problems are inherently variants of accurate geo-localization, which is vulnerable to cyberattacks on multiple surfaces (GPS-spoofing, camera, vehicles ethernet). Therefore, the joint sensing-planning solution will be considered to eliminate the effects of added noise in the traditional two step approach and a sophisticated joint optimization framework resilient to cyberattacks will be deployed. The aim is to compute the complete continuous-time trajectory from start to goal in a synchronous manner, such that, at each time-step, the solutions to the estimation problem (history of the scene) leverage context awareness, and the solutions to the planning problem (prediction of scene evolution and trajectory) automatically emerge. Additionally, performing this joint optimization allows information to flow

between perception and planning resulting in mutual benefits. This approach is highly innovative as the CAV must contend with a potentially high-degree-of-freedom (DOF) trajectory space, as well as to model the uncertainty due to potential cyberattacks, and the stochastic effect of executing actions by the controller.

VI. Conclusions

This papers presents the overall approach and vision of the CoEvolution project toward addressing critical gaps in securing AI systems across their lifecycle and supply chain by introducing a holistic, end-to-end Security, Trust, and Robustness (STR) assessment solution. By integrating AI Model Bills of Materials (AI MBOM) for transparency, context-aware adaptation mechanisms, and a unified STR assessment hub, CoEvolution ensures compliance with evolving regulatory standards like the EU AI Act while mitigating potential security risks. The framework's modular design supports diverse AI paradigms—individual models, collaborative learning, and interconnected agents—enhancing resilience and robustness against adversarial threats. Validation through real-world pilots in autonomous vehicles and industrial automation aims to demonstrates its efficacy in dynamic, adversarial environments.

Acknowledgment

This paper has received funding from the European Union's Horizon Europe research and innovation actions under grant agreement No 101168560 (CoEvolution). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] H. Mann, *Artificial Integrity: The Paths to Leading AI Toward a Human-Centered Future*. John Wiley & Sons, 2024.

[2] Y. Xu, D. Li, Q. Li, and S. Xu, "Malware evasion attacks against iot and other devices: An empirical study," *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 127–142, 2024.

[3] Q. Zhou, M. Zuley, Y. Guo, L. Yang, B. Nair, A. Vargo, S. Ghannam, D. Arefan, and S. Wu, "A machine and human reader study on ai diagnosis model safety under attacks of adversarial images," *Nature communications*, vol. 12, no. 1, p. 7281, 2021.

[4] T. Bai, J. Zhao, J. Zhu, S. Han, J. Chen, B. Li, and A. Kot, "Ai-gan: Attack-inspired generation of adversarial examples," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2543–2547.

[5] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 2154–2156.

[6] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, and F. Roli, "Wild patterns reloaded: A survey of machine learning security against training data poisoning," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.

[7] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *arXiv preprint arXiv:1206.6389*, 2012.

[8] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[9] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. Springer, 2013, pp. 387–402.

[10] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[12] A. Bellet, R. Guerraoui, M. Taziki, and M. Tommasi, "Personalized and private peer-to-peer machine learning," in *International conference on artificial intelligence and statistics*. PMLR, 2018, pp. 473–481.

[13] M. Shayan, C. Fung, C. J. Yoon, and I. Beschastnikh, "Biscotti: A ledger for private and secure peer-to-peer machine learning," *arXiv preprint arXiv:1811.09904*, 2018.

[14] C. Pappas, D. Chatzopoulos, S. Lalis, and M. Vavalis, "Ipls: A framework for decentralized federated learning," in *2021 IFIP Networking Conference (IFIP Networking)*. IEEE, 2021, pp. 1–6.

[15] H. Nguyen and H. La, "Review of deep reinforcement learning for robot manipulation," in *2019 Third IEEE international conference on robotic computing (IRC)*. IEEE, 2019, pp. 590–595.

[16] S. Gupta, S. Bhambri, K. Dhingra, A. B. Buduru, and P. Kumaraguru, "Multi-objective reinforcement learning based approach for user-centric power optimization in smart home environments," in *2020 IEEE International Conference on Smart Data Services (SMDS)*. IEEE, 2020, pp. 89–96.

[17] H. Zheng and A. Louri, "An energy-efficient network-on-chip design using reinforcement learning," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[18] X. Chen, S. Li, H. Li, S. Jiang, and L. Song, "Neural model-based reinforcement learning for recommendation," 2018.

[19] A. Navon, A. Shamsian, G. Chechik, and E. Fetaya, "Learning the pareto front with hypernetworks," *arXiv preprint arXiv:2010.04104*, 2020.

[20] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz *et al.*, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, p. 26, 2022.

[21] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.

[22] G. Tesauro, R. Das, H. Chan, J. Kephart, D. Levine, F. Rawson, and C. Lefurgy, "Managing power consumption and performance of computing systems using reinforcement learning," *Advances in neural information processing systems*, vol. 20, 2007.

[23] Y. Zhang, J. Hou, and Y. Yuan, "A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks," *International Journal of Computer Vision*, vol. 132, no. 5, pp. 1592–1624, 2024.

[24] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465–499, 2023.

[25] N. Ahmed, A. Ameli, and H. Naser, "Detection, identification, and mitigation of false data injection attacks in vehicle platooning," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 1, pp. 1296–1309, 2025.