IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 April 2026, accepted 27 April 2026, date of publication 5 May 2026, date of current version 12 May 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3690620

SURVEY

Adversarial Machine Learning: A 20-Year Survey of Attacks, Defenses, and Standards

BENHUR TEKESTE[1], KHALIL AL-HUSSAENI[1], (Member, IEEE), BENJAMIN C. M. FUNG[2], (Senior Member, IEEE), IBTESAM ALAWADHI[3], AND CLAUDE FACHKHA[4], (Senior Member, IEEE)

[1]Electrical Engineering and Computing Sciences Department, Rochester Institute of Technology, Dubai, United Arab Emirates
[2]School of Information Studies, McGill University, Montreal, QC H3A 1X1, Canada
[3]Dubai Police Academy, Dubai, United Arab Emirates
[4]College of Engineering and IT, University of Dubai, Dubai, United Arab Emirates

Corresponding author: Khalil Al-Hussaeni (kxacad@rit.edu)

This work was supported by Dubai Future Foundation through the Dubai Research, Development, and Innovation (RDI) Program under Grant 2025/DRDI0188.

ABSTRACT Adversarial Machine Learning (AML) presents a significant barrier to the large-scale deployment of Artificial Intelligence (AI) in safety-critical environments. While early research focused on algorithmic robustness, the field has evolved into a complex intersection of security, assurance, and policy. This paper presents a comprehensive, multidisciplinary survey of the AML landscape, covering over 250 peer-reviewed contributions. We implement a lifecycle-oriented taxonomy that maps attack vectors and defense mechanisms to specific stages of the AI pipeline from data collection to deployment, expanding the traditional Confidentiality, Integrity, and Availability (CIA) triad to include Governance and Regulation. We identify critical research gaps, including certified robustness for Natural Language Processing (NLP), and emerging threats in Generative AI. To ground these theoretical insights in practice, we analyze five domain-specific case studies: Autonomous Vehicles, Medical AI, Financial Systems, Natural Language Processing (NLP), and the Internet of Things (IoT). Uniquely, this survey bridges the gap between academic literature and industrial practice by mapping technical AML findings to emerging standards, including the NIST AI Risk Management Framework (RMF), MITRE ATLAS, and ISO/IEC 42001. We conclude by providing a roadmap for researchers, practitioners, and regulators to build verifiable, trustworthy, and compliant AI systems.

INDEX TERMS Adversarial machine learning, AI security, AI assurance, certified robustness, MLSecOps, security governance, LLM safety, AI risk management, privacy.

I. INTRODUCTION

Advances in modern systems, based on a deep understanding and utilizing machine learning and data-driven AI, have enabled new capabilities across a range of applications, including speech recognition and visual perception (e.g., self-driving vehicles and robotics), as well as cybersecurity functions such as spam filtering and malware detection [1]. However, these systems are susceptible to adversarial examples, which are inputs with deliberately introduced perturbations, many of which may be imperceptible, intended to cause the system to produce an erroneous output at test time. This susceptibility was highlighted in 2014 when Szegedy et al. [2] demonstrated that several high-performing classification systems could be deceived by small perturbations and subsequently by Goodfellow et al. [3], Nguyen et al. [4], and Moosavi-Dezfooli et al. [5].

In fact, a substantial body of research on adversarial attacks and defenses has been conducted since 2014, primarily in computer vision, e.g., [6], [7], [8], [9], [10], [11], [12], reviving interest in the broader area of AML, while also creating a widespread misconception that AML originated in 2014. Several recent surveys and position papers appear to accept 2014 as a de facto inception year

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

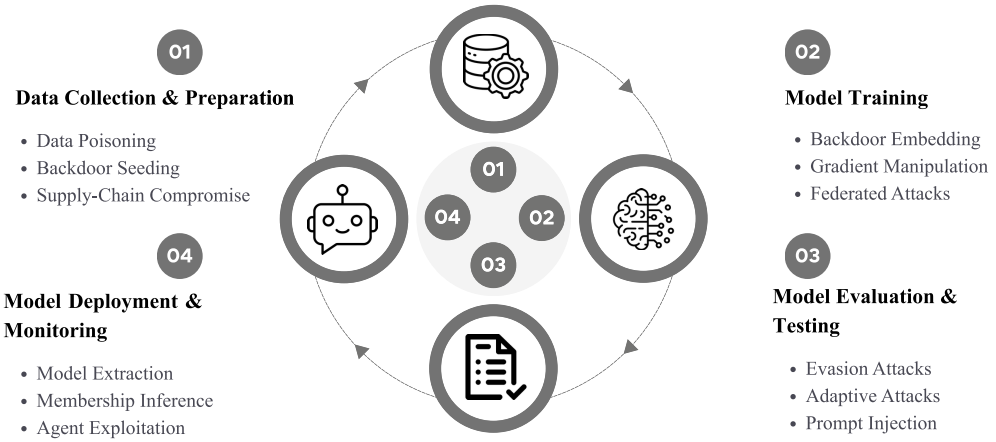

FIGURE 1. High-level consolidation of adversarial vulnerability classes mapped to the machine learning lifecycle.

for AML [13], [14], [15], [16], although AML predates deep learning and has repeatedly discovered phenomena that were well-known outside of deep vision. Research on non-adversarial robustness [17], [18] and domain adaptation [19] address related but distinct challenges in machine learning. In particular, non-adversarial robustness and domain adaptation focus on generalization under natural distributional shifts without assuming an explicit adversary. However, AML is primarily focused on the worst-case or malicious perturbations from adversaries. Actual robustness will require more than just patching gradients. Consequently, some researchers argue that improving robustness may require learning causal or invariant representations that capture stable relationships across environments [20], [21]. However, other approaches emphasize alternative strategies, such as distributionally robust optimization or domain generalization methods that aim to maintain performance under shifting conditions without explicitly enforcing causal invariance [22], [23].

The first works in AML predated the advent of deep learning. Dalvi et al. [24] studied adversarial manipulation in spam filtering and demonstrated that even slight modifications to the content of spam messages could evade linear classification systems. Subsequently, Lowd and Meek [25] developed the concept of ''good word'' attacks, exploiting access to queries made by statistical filters. In biometrics, Matsumoto et al. [26] demonstrated that fake fingerprints could lead to incorrect identification decisions in fingerprint recognition systems. Building on these works, Barreno et al. [27] developed a framework for classifying attacks on learning systems into evasion, poisoning, and other categories, and discussed defense mechanisms from the perspective of a system.

The maturity of this pre-2014 body of research is evident in the early community's activities, including the 2007 NIPS Workshop on Machine Learning in Adversarial Environments, where machine learning researchers joined with researchers from computer security to explore how attackers could interfere with learning systems. The Machine Learning journal special issue [28] introduced the core ideas of threat models, assessing performance under attack, and evaluating a model's robustness, which were then formally defined in papers such as Biggio and Roli's work [13] and implemented using benchmarks such as RobustBench [29]. Additionally, the 2013 Dagstuhl seminar on ML and Security [30], and the AISec workshop series [31] helped shape the AML discipline. Together, these groups establish that AML has been active long before the development of deep learning; however, it was from 2014 onwards that AML became recognized across all areas of computing.

To provide a foundation for our review of the AML space, we establish standard terms and definitions. Threat models vary in the information available to the adversary, ranging from full access to model parameters in white-box scenarios to limited query access in black-box scenarios. Additionally, threat models can differ by their budgetary limitations (number of queries and computational resources). Allowed perturbation types can vary significantly across settings, including small, bounded changes; perceptually aligned modifications; and physical transformations that are feasible to implement in real-world settings. We define the term ''adversarial example'' to represent any input whose characteristics have been altered according to some constraints (budget, type of perturbation) to deceive a model during testing. In this survey, we will consider attacks and defenses throughout the life cycle of development and deployment of a machine learning model.

In the past ten years, there has been major growth in the area of AML. Some of the most well-known areas of advancement are evasion attacks during testing [2], [25], [32], and poisoning attacks during training [33], [34], [35], [36]; evaluation frameworks and benchmarks to evaluate model robustness [12], [29], [37], [38]; and defense against these types of attacks through adversarial training and

robust optimization, detecting and rejecting, and certified robustness [8], [39], [40], [41].

Additionally, we expand the lens of AML beyond vision. For instance, in NLP, attacks exploit discrete constraints (e.g., via paraphrases and edits) to create adversarial examples; in audio and speech, attacks are performed by adding noise over-the-air to challenge Automatic Speech Recognition systems; in reinforcement learning, adversarial observations and reward poisoning can affect the decision-making of agents. More recently, the large transformer models being used for various NLP tasks raise several additional concerns regarding jailbreaking, prompt injecting, circumventing safety filters, and data poisoning at pre-training scale. These areas typically diverge from small-norm-bounded perturbations, focusing on higher-level content constraints, tool usage, and interactions. It is essential to perform good evaluations. To do so, we follow established best practices: utilize strong, adaptive attacks and specify attack configurations; avoid gradient masking and apply transformations [10]; report both standard and robust accuracy under equal budgets; reveal query/computational budget limits in black-box scenarios; and, whenever possible, leverage standardized evaluation suites [12], [29], [38]. Finally, we examine the relationship between robustness and accuracy, as well as the roles of data, model size, and architecture.

Our survey presents a comprehensive and historically contextualized overview of the AML field, spanning the early roots of AML in non-deep models through the current state of the art in deep learning. In addition to providing a historical context, we focus on the period 2010-present and examine the advancements in certified robustness, the theory of adversarial training, attacks and defenses on large transformer models, and robustness in interactive and security-critical systems. Our objective is to connect developments across different eras and domains and address some ongoing misconceptions about the evaluation of security and the reliability of learning systems.

The contributions of this survey are outlined below:

- A comprehensive chronological analysis of adversarial machine learning spanning 2002 to 2025, tracing the evolution from early linear classifier exploits to modern attacks on foundation models and generative AI.
- A lifecycle-centric taxonomy that organizes attacks and defenses by their injection point in the Continuous Integration and Continuous Development (CI/CD) pipeline (Data Collection, Training, Evaluation, and Deployment), providing a practical framework for the secure development of ML models.
- A cross-domain analysis examining vulnerabilities across diverse modalities, including computer vision, NLP, audio, and Internet of Things (IoT), highlighting transferability of threats across domains.
- A strategic mapping of academic findings to industrial standards, specifically bridging the gap between technical research, best practices, and regulatory frameworks such as MITRE ATLAS, OWASP ML Security Top 10, NIST AI RMF, the EU AI Act, and ISO/IEC 42001.

The remainder of the paper is organized as follows. Section II outlines the methodology we used for this survey. Section III presents a brief background on adversarial machine learning and its evolution. Section IV presents a survey of different categories related to adversarial machine learning. Section V presents a taxonomy of AML attacks and defenses across the development lifecycle of ML models. Section VI instantiates AML in high-stakes domains (autonomous driving, healthcare, finance, LLM/NLP, IoT/edge) to show that the same threat patterns recur under different operational constraints. Section VII consolidates metrics, datasets, and benchmarking practice so results from other works can be compared fairly. Section VIII connects these technical results to MITRE ATLAS, NIST AI RMF, ISO/IEC SC 42, the EU AI Act, and related industry efforts, clarifying what ''robustness to attacks'' means in practice. Section IX provides a discussion and recommendations for various stakeholders and open research problems for future work on adversarial machine learning. Section X concludes our work with our findings of the current research contributions

II. SURVEY METHODOLOGY

Our methodology is a structured survey approach that combines manual curation and a protocol-driven filtering process to provide a comprehensive and reproducible overview of AML research over the last two decades. The methodology described below is composed of several components:

A. SEARCH STRATEGY

We developed a search strategy to identify scientific literature related to AML threats, defenses, and governance. Our literature search was conducted across the scientific databases IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar. The search was restricted to peer-reviewed sources and published archival sources. Queries used combinations of Boolean terms: (''adversarial machine learning'' OR ''adversarial attack'' OR ''evasion attack'' OR ''poisoning attack'' OR ''model extraction'' OR ''membership inference'' OR ''backdoor attack'' OR ''certified robustness'') AND (''defense'' OR ''mitigation'' OR ''benchmark'' OR ''survey'' OR ''AI security'' OR ''MLSecOps''). The coverage window spanned publications from 2002 to the present, corresponding to the emergence of modern AML literature. The final search update was conducted in March, 2026.

B. NUMBER OF SOURCES IN THE CORPUS, REMOVING DUPLICATES, AND THE FINAL LIST OF REFERENCES

We had an initial number of 437 possible sources after searching relevant literature. After title-based duplicate removal, we were able to remove 25 duplicate sources from our database (412 potential sources) to screen them for

relevance to our topic. The total number of references in the final bibliography was 253.

C. INCLUSION CRITERIA

Papers were included in this review based on the following criteria:

1) They presented novel theoretical or empirical insights;
2) They proposed or analyzed an attack or defense mechanism;
3) They provided systematic benchmarks, taxonomies, or real-world case studies; or
4) They addressed assurance, explainability, or certification within an adversarial context.
5) They are highly cited articles published in reputable venues.
6) Preprints are included when they correspond to widely used datasets, benchmarks, or technical reports.

D. EXCLUSION CRITERIA

Excluded from this review were:

1) Redundant references which were eventually replaced by peer-reviewed versions;
2) Speculative commentary that was without empirical or formal support; and
3) Papers that addressed general AI ethics or safety that did not address adversarial robustness.

E. DATA EXTRACTION AND CATEGORIZATION

Each paper selected for inclusion was coded with metadata that included information about:

- Threat Model (white-box, black-box, transfer);
- Attack/Defense Type (e.g., PGD, certified defense, DP-based);
- Modality;
- Evaluation Metric (e.g., ASR, accuracy under perturbation); and
- Whether Detection-Aware Evaluation was Considered.

If possible, the authors attempted to normalize the attack budget and the type of evaluation metric utilized to enable comparison of the results across papers.

F. COMPARATIVE TABLES AND TAXONOMIES

All of the comparative tables were manually curated using the coded metadata. Footnotes were added to comparative tables when the evaluation metrics were different (e.g., ASR vs. ASRD). The authors annotated outliers and non-standard settings (e.g., gradient-free attacks with unlimited queries) to prevent misinterpretation.

G. LIMITATIONS

Although the main objective of our survey was to gather literature on AML across tasks and modalities, it should be noted that our taxonomy has been largely shaped by high-AML-research-activity areas, i.e., image classification and NLP. For example, future work may develop a publicly accessible database or live benchmark of additional metadata.

III. BACKGROUND

A. MACHINE LEARNING FUNDAMENTALS

In today's data-driven world, machine learning (ML) is a foundational element of almost all fields, including finance, healthcare, and security. Machine learning enables systems to learn from experience and develop intelligence without direct programming, thereby facilitating intelligent automation and informed decision-making. Figure 1 shows the end-to-end development and deployment of an ML model.

The development of ML-based systems has followed several established principles and practices as defined by the ISO/IEC standardization body [42], [43], [44], [45], [46], [47], which includes ensuring the reliability, safety, and compatibility of ML-based system components.

Recent conversations have also focused on the DevSecMLOps paradigm [48], which integrates security and compliance into an ML developer's overall development cycle; visualization tools and dashboards have helped increase model transparency, particularly in high-risk applications [49].

Future-oriented research efforts will continue to focus on developing robust frameworks for ML-based systems that support current national innovation objectives, data governance, and ethics of AI; these developments reflect a shift from a strictly technical approach to a broader social, institutional, and regulatory approach [31].

B. EVOLUTION OF ADVERSARIAL MACHINE LEARNING

In this section, we will present the evolution of AML over the past 20 years. Figure 2 provides the chronological developments in AML across various periods.

The early foundations were laid in the mid 2000s when the first studies on the security of machine learning emerged. Dalvi et al. [24], and Lowd and Meek [50] proved that spam filters can be evaded with maliciously manipulated inputs. They are the first to introduce the concept of adversarial manipulation to traditional machine learning models based on linear classifiers. Finally, Biggio et al. [32] formalized adversarial machine learning as a scientific discipline. They have defined a structured threat model that distinguishes between evasion and poisoning attacks during the testing and training phases, respectively.

For the first time, the deep learning vulnerability was demonstrated by Szegedy et al. [2]. They proved that imperceptible perturbations generated with L-BFGS can mislead the best neural network architectures currently available. Then, Goodfellow et al. [3] proposed the Fast Gradient Sign Method (FGSM) and pointed out that neural networks have a linear nature that is responsible for their fragility. At the same time, Biggio et al. [37] presented one of the first taxonomies of attacks and countermeasures.

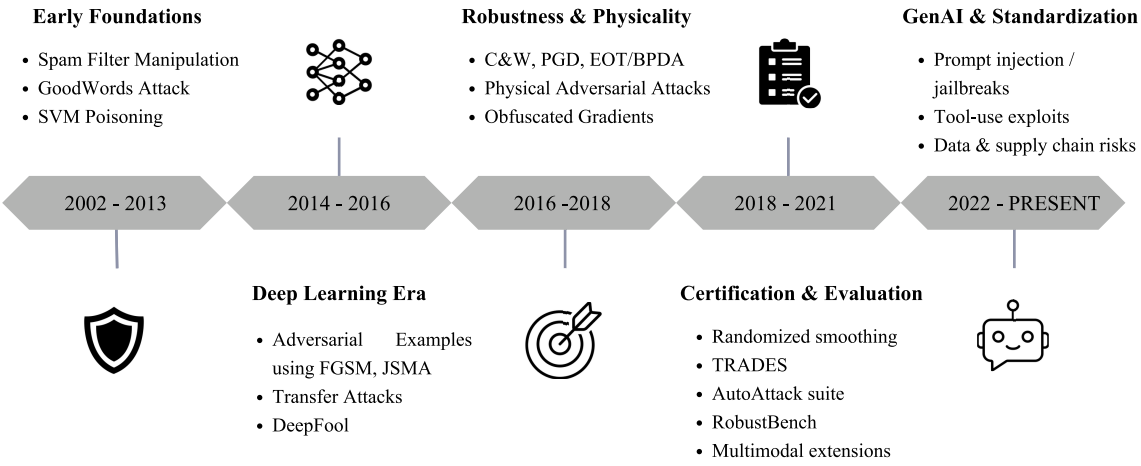

FIGURE 2. AML evolution.

As a consequence of these pioneering works, several countermeasures emerged during this period, such as adversarial training [3] and defensive distillation [51].

In subsequent years, new and stronger attacks emerged. Carlini and Wagner [9] developed powerful attacks based on optimization techniques. Madry et al. [8] proposed Projected Gradient Descent (PGD) and set up a reference point for first-order attackers. Brown et al. [52] extended the concept of adversarial examples to the physical world, introducing adversarial patches. Moreover, several authors have explored the application of AML to modalities beyond visual, such as audio [53], text [54], [55], and 3D/physical-world attacks [56]. In addition, defense mechanisms have evolved significantly, particularly in certified robustness approaches [39], [57] and improved variants of adversarial training. Finally, Athalye et al. [10] investigated the limits of obfuscation techniques to defend against AML attacks.

By the end of the last decade, the research community began developing systematic reviews and taxonomies to organize the growing body of knowledge in AML. For instance, Papernot et al. [58] and Pitropokis et al. [14] produced comprehensive reviews of attacks and countermeasures, while Zhang et al. [59] analyzed the relationship between accuracy and robustness. In addition, the research expanded to areas such as reinforcement learning [60], graph neural networks [61], and federated learning [62]. The RobustBench [29] initiative represented a milestone towards standardizing the evaluations of AML attacks and countermeasures.

Recently, due to the success of foundation models and large language models (LLMs), the attacks against AML have been moved from low-level manipulation of input data (perturbation) to high-level manipulation of input data, such as prompt injection, jailbreaks, and backdoored pre-trained checkpoints [63]. Consequently, the number of potential attack surfaces has increased dramatically, including model extraction, sleeper agents, and prefix leakage in LLMs. National and international institutions have also formalized the taxonomy of AML attacks and countermeasures, e.g., NIST's AI RMF [64] clearly distinguishes between threats to Predictive AI (PredAI) and threats to Generative AI (GenAI). Currently, AML is widely recognized as a fundamental component of the AI safety and assurance landscape, with implications for critical infrastructure, disinformation, and public confidence in AI systems.

C. DISTINCTIONS BETWEEN ROBUSTNESS, SAFETY, AND GOVERNANCE

Robustness, safety, and governance are three distinct but hierarchical levels of protection for AML (see Figure 3). In the context of AML, robustness is a *model-level* attribute that describes the ability to maintain correct outputs even under input perturbations introduced by an adversary with malicious intent, within a defined threat model. Robustness to natural distribution shifts arising from the data-generating process (e.g., domain shift or noise) is considered out of scope. Adversarial robustness can be measured using metrics such as robust accuracy or formal certification guarantees [29], [40]. Safety, on the other hand, is a *system-level* attribute that is concerned with preventing adverse outcomes from occurring during the deployment of the system; utilizing alignment techniques such as RLHF [65] and runtime guardrails, safety provides assurance that the system will act in a manner consistent with being safe even though the underlying model may remain vulnerable to certain types of perturbations. The highest level of protection for AML systems is provided by governance which is an *organizational* attribute comprised of the regulatory frameworks and standards that provide no direct technical assurances of robustness and safety, but require the administrative validation and enforcement of robustness and safety attributes throughout the development lifecycle.

IV. RELATED WORK

AML has been analyzed from multiple perspectives. We will divide previous works into three categories that include

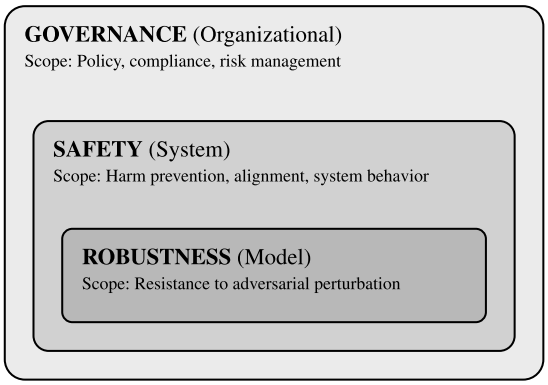

FIGURE 3. Hierarchical relationship between robustness (model-level), safety (system-level), and governance (organizational-level).

fundamental reviews, attack/defense taxonomy reviews, domain-specific reviews, and a final review in comparison to our review.

A. FOUNDATIONAL AML REVIEWS

Barreno et al. [27] established an initial framework for studying the security of ML systems, and proposed a taxonomy of attacks on ML systems, and examined possible defenses and challenges. Laskov and Lippman [28] studied the emerging field of adversarial learning, developing classifications of attacks in adversarial environments and stressing the importance of developing strategies to create classifiers resistant to manipulations of malicious data. Huang et al. [66] were among the first to formally define ''Adversarial Machine Learning'' as a field, defining types of threats, and identifying some of the earliest defensive methods to encourage systematic study of attacks on learning algorithms. Biggio and Roli [13] conducted a comprehensive review of AML over the past decade, examining evolving forms of attacks and threat models from before the advent of deep learning to the advent of deep learning, and highlighting pitfalls of AML evaluation, misunderstandings, and future research areas toward creating more secure ML models. A summary of the fundamental AML reviews is shown in Table 1.

B. AML ATTACK AND DEFENSE TAXONOMY SURVEYS

As Akhtar and Mian [16] provided an early survey focused on the threats caused by adversarial attacks to deep learning-based computer vision and the many ways to develop these attacks and defend against them, Yuan et al. [15] surveyed the generation of adversarial examples and the defenses to counter them in deep neural networks and developed a taxonomy of attack types and defenses and explored challenges to creating robust deep learning solutions, while Pitropakis et al. [14] introduced a taxonomy of adversarial attacks on machine learning systems and organized these attacks based on the primary characteristics of attacks (e.g., when the attack occurs and what the attacker knows about the system), with the intent of organizing the terminology used to describe attacks across multiple disciplines and identifying areas that require additional research (specifically, the authors focus on attacks and leave defenses as future research); A summary of the AML reviews of attacks and defenses is shown in Table 2.

C. DOMAIN-SPECIFIC AML SURVEYS

Ilahi et al. [67] conducted a systematic review of the threats to deep reinforcement learning (DRL) systems in cyber-physical domains, including smart grids, autonomous driving, and traffic control. The authors identified and described various vulnerabilities in DRL systems that can be exploited by adversaries, and they compared and summarized the most effective defenses to prevent these types of attacks, and also identified the open challenges to protect the real-world DRL applications. Deng et al. [68] studied the security of autonomous and agentic AI systems, which operate in highly dynamic and complex environments. The authors identified the main security threats, including unpredictable multi-step inputs, complex internal decision-making processes, and varying operating environments, and proposed directions for future research to secure these types of systems. Liu et al. [69] conducted a systematic review of the emerging ''jailbreak'' attacks on large-scale generative models, including multimodal and conversational AI systems. The authors cataloged how adversaries bypass the built-in safety constraints of these systems and summarized defense mechanisms, identifying the distinct security challenges posed by generative AI systems and the need for novel safeguards in this context. A summary of the domain-specific AML surveys is shown in Table 3.

D. COMPARISON: THIS SURVEY VERSUS PRIOR WORKS

Table 4 provides a comparison of this work to prior research on adversarial machine learning (AML). As described above, prior work has been limited to specific aspects of AML. For example, Biggio and Roli [13], Yuan et al. [15], and Pitropakis et al. [14] examined evasion and poisoning attacks during the training and test phases of the machine learning lifecycle, primarily from a model-centric viewpoint. Although Yuan et al. [15], and Pitropakis et al. [14] have established taxonomies of adversarial examples and attack characteristics for predictive models (i.e., classifiers), these works also did not examine generative AI systems, Agentic workflows, or regulatory and assurance requirements.

Biggio and Roli [13] were among the first to investigate evasion and poisoning attacks at both the training and test phases of the machine learning lifecycle. They offered a model-centric view of machine learning, but did not address upstream data risks, downstream deployment threats, or organizational governance. Similarly, Yuan et al. [15], and Pitropakis et al. [14] have focused on predictive models (i.e., classifiers) and developed taxonomies of adversarial examples and attack characteristics; however, neither Yuan et al. [15] nor Pitropakis et al. [14] have addressed the

TABLE 1. Fundamental AML reviews.

Review	Year	Focus	Contributions
Barreno et al. [27]	2006	Initial investigation of the security of ML systems; described the influences of an adversary on both the training phase and test phase of the system.	Defined one of the first attack taxonomies (indiscriminate/targeted, exploratory/causative) and listed research barriers to secure learning.
Laskov and Lippmann [28]	2010	Definition of the foundations of adversarial learning, and definition of formal threat models that can be used across ML domains.	Defined general principles of designing robust classifiers, and stated the need for interdisciplinary methods to improve the resilience of ML systems.
Huang et al. [66]	2011	Formalization of Adversarial Machine Learning as a new research field.	Differentiated between exploratory and causative attacks; provided a summary of initial defensive methods to protect against adversarial perturbations.
Biggio and Roli [13]	2018	A decade-long look back at the development of AML from shallow to deep models.	Described the growth of attacks (both evasion and poisoning) over time, described pitfalls of evaluating AML systems, and presented future research areas to create more secure ML systems.

TABLE 2. AML review of attacks and defenses.

Review	Year	Focus	Contributions
Akhtar and Mian [16]	2018	Complete survey of the adversarial attacks on deep learning in computer vision.	Summarized methods for generating adversarial images, strategies to defend computer vision models, and demonstrated real-world vulnerabilities in vision models.
Yuan et al. [15]	2019	Wide-ranging review of the attacks and defenses of adversarial examples for deep neural networks.	Developed a taxonomy of adversarial example generation, reviewed methods to defend against adversarial examples, and identified research barriers to enhance the robustness of DNNs.
Pitropakis et al. [14]	2019	Classification-based taxonomy of attacks in the Adversarial Machine Learning domain.	Created a taxonomy of Adversarial ML attacks that is unifying (training vs. inference, attacker knowledge), explained the threat landscape, and identified open research directions for defending Adversarial ML attacks.

TABLE 3. Domain-specific AML surveys.

Survey	Year	Focus	Contributions
Ilahi et al. [67]	2022	Adversarial attacks on deep reinforcement learning (DRL) systems in cyber-physical domains such as smart grids, autonomous driving, and traffic control.	Reviewed DRL vulnerabilities exploited by adversaries, summarized state-of-the-art defenses, and identified open challenges for securing real-world DRL applications.
Deng et al. [68]	2025	Security challenges for autonomous and agentic AI operating in dynamic, complex environments.	Categorized emerging threats such as goal manipulation and unpredictable inputs; outlined research directions for securing multi-agent and autonomous systems.
Liu et al. [69]	2024	Adversarial "jailbreak" attacks against generative AI and foundation models.	Proposed a taxonomy of safety-bypassing attack vectors, surveyed defense mechanisms, and analyzed robustness and safety gaps unique to generative AI deployments.

use of generative AI systems, Agentic workflows, or regulatory and assurance requirements.

Recently, Patel et al. [70] built upon prior efforts by mapping adversarial attacks to the MLOps (DevSecOps) pipeline and developing an engineering-oriented view of secure machine learning operations. However, the authors' analysis is mostly restricted to traditional predictive models and engineering controls. Specifically, Patel et al. [70] do not consider (i) threats to generative and foundation models (e.g., jailbreaks, RAG poisoning, indirect prompt injection); (ii) Agentic AI systems that perform multi-step planning and tool invocation; (iii) the computational and operational overhead of defensive mechanisms under industrial constraints; or (iv) the regulatory compliance and post-market obligations that arise from the failure to comply with regulations. Therefore, Patel et al. [70] definition of "defense" is mainly technical and implementation-centric.

Instead, this survey employs an AI assurance perspective that encompasses predictive, generative, and agentic AI threats throughout the entirety of the ML lifecycle while directly relating technical defensive mechanisms to organizational governance and regulatory standards. The survey maps attacks and mitigation measures to frameworks such as NIST AI RMF, MITRE ATLAS, ISO/IEC 42001, and the EU AI Act. This alignment of perspectives will shift the conversation from how to merely patch models to how to develop and deploy trustworthy and compliant AI systems within real-world settings.

Table 4 provides a summary of the differences among the surveyed works.

TABLE 4. Comparison: this survey vs. prior works.

Survey	Year	Primary Focus	Lifecycle Scope	GenAI / LLM?	Agentic AI?	Standards & Gov.?
Foundational Works						
Biggio & Roli [13]	2018	Attacks & Defenses	Training & Test Only	No	No	No
Yuan et al. [15]	2019	Deep Learning Attacks	Model-Centric	No	No	No
Pitropakis et al. [14]	2019	Attack Taxonomy	Model-Centric	No	No	No
Concurrent Works (2024–2025)						
Patel et al. [70]	2025	Secure MLOps	Yes	Predictive Only	No	Operational (Non-Regulatory)
Liu et al. [69]	2024	Jailbreak Attacks	No	Yes	No	No
Deng et al. [68]	2025	Agentic Security	No	Yes	Yes	No
This Survey	**2025**	**AI Assurance**	**Yes**	**Yes**	**Yes**	**Yes (ISO/EU/NIST)**

V. ADVERSARIAL MACHINE LEARNING ACROSS THE MODEL LIFECYCLE

AML has traditionally been treated as a "loose" collection of techniques, e.g., adversarial examples, data poisoning, and model extraction. As such, the current view of AML is incomplete. Modern ML systems are long-lived pipelines with feedback loops, fine-tuning cycles, telemetry ingestions, deployment governance, and API exposures. Therefore, attacks occur at every point throughout the pipeline lifecycle, and defenses need to counter those attacks at all points in the pipeline lifecycle.

In this section, we will organize ML adversarial threats and their defenses, stage by stage, along the ML lifecycle. Figure 4 presents a high-level classification of attacks and defenses along the ML lifecycle. Organizing attacks and defenses by lifecycle stage reveals security gaps at each phase and highlights the need for end-to-end robustness throughout model development

1) Data Collection & Preparation
2) Model Training
3) Model Testing & Evaluation
4) Model Deployment & Monitoring

At each stage we will describe (i) the primary attack surfaces, (ii) the characteristic techniques used by the adversary, and (iii) the defensive strategy families available for addressing those threats.

A. STAGE 1: DATA COLLECTION & PREPARATION

1) ATTACKS

During data collection and preparation, adversaries get their first chance to carry out a myriad of attacks as seen in Figure 5. Adversaries may poison the data source or corrupt the integrity of training samples to create models that behave in biased, degraded, or covert ways. There are two primary forms of threats at this stage: data poisoning and data injection or corruption.

a: DATA POISONING

Data poisoning involves deliberately corrupting the training dataset or adding malicious samples to the training set so that the learned model will express the adversary's desired behavior during the evaluation or deployment stage.

Indiscriminate poisoning is designed to generally decrease model performance and learning stability. Biggio et al. [33] demonstrated that small changes to the training data significantly decreased classifier accuracy and prevented service by reducing decision boundary reliability. Laskov and Lippmann [28] demonstrated that such perturbations decreased model availability and reliability by inducing instability in model convergence. Bhagoji et al. [71] demonstrated that, using gradient-based optimization, small but structured corruptions of the data can result in catastrophic performance degradation in deep systems.

Targeted or clean-label poisoning involves introducing malicious samples into the training set that appear either visually or semantically plausible to induce the model to behave as the adversary desires. Shafahi et al. [35] proposed clean-label poisoning, where correctly-labeled data induces targeted misclassifications. Huang et al. [36] generalized this concept using MetaPoison to develop attacks against multiple architectures. Aghakhani et al. [72], Kasichainula et al. [73], and Jagielski et al. [74] demonstrated that even high-trustworthiness data sets can be contaminated with significant systematic integrity failures.

Backdoor or Trojan poisoning embeds hidden behaviors within the model that activate upon the detection of a specific trigger pattern. Bagdasaryan et al. [62] and Saha et al. [75] demonstrated that very small, nearly invisible trigger patterns could completely override the prediction made by the model. Xie et al. [76], Nguyen and Tran [77], and Zheng et al. [78] later developed stealthy and distributed backdoors that propagate in federated and compressed settings. Han et al. [79] analyzed how backdoors could be introduced into multimodal learning platforms that integrate vision and language. They proposed a cross-modal consistency attack that included poisoned training data to make sure the backdoor would only trigger under certain feature combinations from both of the different modalities. Although they were able to achieve very high attack success rates using very low amounts of poisoned training data, the method has two limitations. First, the alignment of triggers can become too complex for various

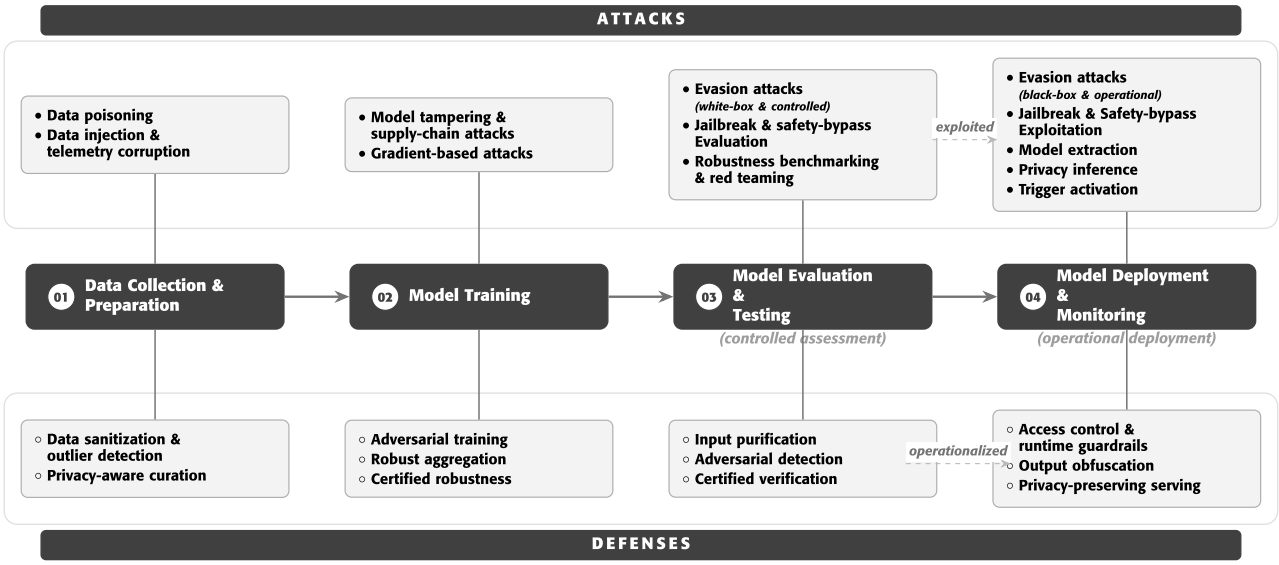

FIGURE 4. Adversarial machine learning taxonomy across the ML lifecycle. Evasion and jailbreak attacks appear in both Stage 3 and Stage 4 with context-specific qualifiers: Stage 3 addresses them under controlled, white-/semi-white-box assessment conditions, while Stage 4 addresses their black-box, rate-limited exploitation in production. Dashed arrows indicate that attacks characterized during evaluation are exploited at deployment, and defenses validated during testing are operationalized in production.

datasets. Second, there exists an open challenge as to how defense mechanisms can be developed for detecting cross-modal subtle adversarial patterns.

b: DATA INJECTION AND CORRUPTION

Modern ML systems continually retrain or fine-tune on evolving data sources (logs, telemetry, feedback, and federated client updates), providing additional opportunities for the continued corruption of the data.

Behzadan and Munir [60] demonstrated that reinforcement learning agents could be steered through the manipulation of feedback loops. Melis et al. [80] identified the inherent vulnerabilities of federated learning due to malicious clients injecting poisoned updates. Fang et al. [81] and Bhagoji et al. [71] further demonstrated how adversarial participants could manipulate the aggregation rounds of federated learning and ultimately shift the overall behavior of the model. These attacks alter the statistical representation of the world the system is trained on, thereby compromising the system's integrity and operational safety.

2) DEFENSES

Defensive methods implemented at the data stage primarily aim to detect and prevent the use of untrusted data and minimize the impact of malicious data contributions on model performance.

Data sanitization and outlier detection methods attempt to identify and filter out poisoned samples before model training. Tran et al. [82] used Spectral Signatures to demonstrate that backdoor triggers create a unique signature in the covariance matrix of feature representations; therefore, by examining the top singular values, defenders can identify and isolate the poisoned samples from the clean data. Gao et al. [84] introduced STRIP (Strong Intentional Perturbation), which detects backdoors by superimposing different clean images on each other over the suspect image. The poisoned images will always have low entropy (constant prediction), regardless of how many images they are superimposed with, while clean images will always have high entropy (highly variable prediction). Despite its effectiveness, STRIP can be less reliable when attackers design triggers that mimic clean data distributions. For the decentralized area, Nguyen et al. [85] applied this concept to Federated Learning, by establishing frameworks such as FLAME that uses clustering and adaptive clipping on client model update (gradients) instead of the raw data, and therefore filter out the malicious contributions made by compromised edge devices before aggregation although it may be vulnerable to sophisticated model poisoning attacks that evade gradient-based detection.

Privacy-aware data curation limits the potential for future information leakage during model training. Papernot et al. [58] established the PATE (Private Aggregation of Teacher Ensembles) framework, where multiple teacher models were trained on different portions of the sensitive data, and then a single student model was trained on the noisy, aggregated vote of the teacher models. This decoupling ensures that the final model provides the strongest possible differential privacy (DP) guarantees, preventing the student model from memorizing individual training examples. Lecuyer et al. [90] provided the first method that relates privacy to robustness, through injecting noise into either the input layer or the latent layer of a neural network to establish a certified robustness bound. This approach guarantees that the model's output remains stable (and thus privacy-preserving) even when the input data undergoes small perturbations.

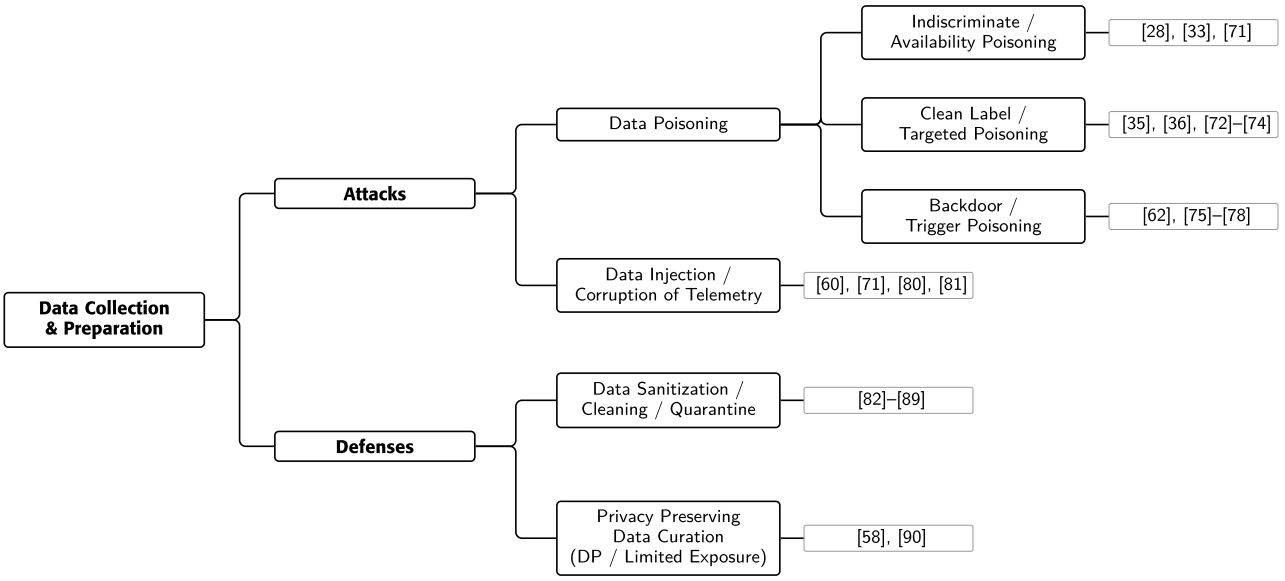

FIGURE 5. Stage 1 - data collection & preparation: attacks and defenses taxonomy.

3) SUMMARY OF FINDINGS FROM V-A

- Poisoning is a fundamental attack that can compromise the model's integrity before training. Indiscriminate corruption can severely impact learning stability, while stealthy and targeted manipulation of model behavior is possible.
- Backdoors continue to be one of the most difficult-to-detect attack forms, since they are embedded in latent malicious functionality that can be activated after deployment with minimal detectability.
- Telemetry and federated data streams extend the attack surface, enabling persistent low-profile manipulation of the model over time in continuous learning or distributed environments.
- Defensive approaches rely more heavily on "trust aware" learning in order to use data cleaning, outlier detection, and robust aggregation to reduce the impact of adversarial examples in large-scale systems.
- Privacy-preserving data collection creates a governance layer, enhancing the confidentiality of the data and decreasing the risk of future model inversion or membership inference attacks.

B. STAGE 2: MODEL TRAINING

The second phase, which is also the most critical, occurs during model training. It includes two types of threats: (i) attacks on the model while it is being trained, by modifying the learned parameters; (ii) attacks on the training process itself, to obtain sensitive information about the model as outlined in Figure 6. This makes the training process a double risk: the attacker can both insert malicious features into the model and steal sensitive data about the model's training process.

1) ATTACKS

a: MODEL TAMPERING AND SUPPLY CHAIN BACKDOORS

An attacker who can interfere with the optimization, fine-tuning, or model's supply chain can insert malicious functions into the model's learned parameters.

Trigger backdoors. Liu et al. [91] first demonstrated that "Trojaning attacks" are possible, i.e., one can retrain a neural network to behave incorrectly when given a specific input (trigger), while maintaining good performance on clean inputs. Standard accuracy metrics will not detect such breaches due to the dual behavior. Furthermore, Hammoud et al. [98] present a new type of backdoor attack for video action recognition (VARS) that uses both visual and auditory channels. A VARS backdoor attack has two components. It creates synchronized triggers in both channels during the model's learning process; it also causes the model to misclassify an action only if both triggers appear during testing. In particular, one of its main limitations is the need for precise timing synchronization between modalities. Moreover, this paper identifies defense against multimodal attacks as an important open problem, since most anomaly detection techniques can detect attacks in only one modality at a time.

Stealthy backdoors. Early backdoor attacks used static, visible, and obvious triggers that could be detected by humans or via outlier analysis. Saha et al. [75] have generalized this idea to "stealthy hidden triggers", which remain invisible after pruning and compression. Aghakhani et al. [72], Zheng et al. [78] demonstrated that backdoors can survive rigorous pruning, quantization, and fine-tuning. Therefore, a model can be compromised during pre-training and remain vulnerable to the same attack even after being adapted by a secure developer for a different task, thereby creating a significant risk of a supply chain breach.

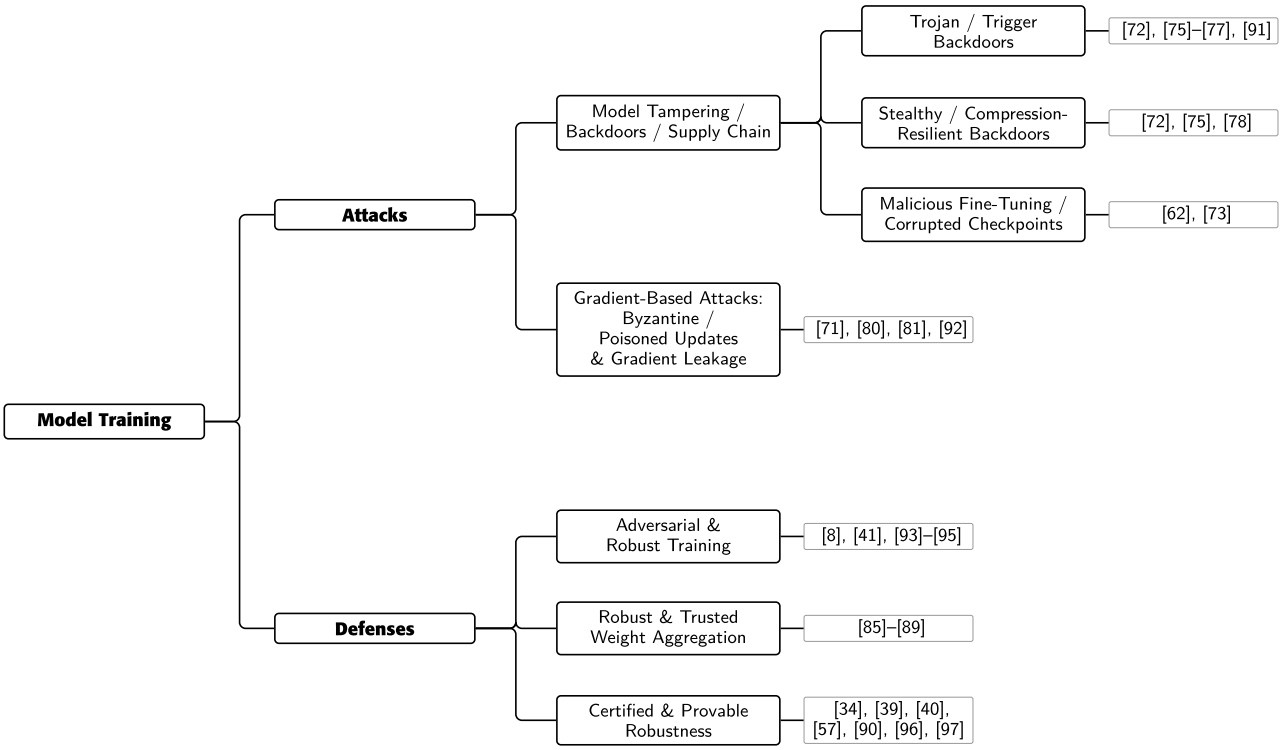

FIGURE 6. Stage 2 - model training: attacks and defenses taxonomy.

Malicious fine-tuning and corrupt checkpoints. The threat surface increases significantly in distributed and generative environments. Nguyen and Tran [77] and Xie et al. [76] have proposed new backdoor architectures and distributed poisoning techniques for federated learning, in which malicious clients send their local models to the server, which aggregates them into a final global model, and the backdoors remain undetected. Additionally, the advent of foundation models provides new avenues for tampering; Bagdasaryan et al. [62] and Kasichainula et al. [73] demonstrated that malicious fine-tuning can remove safety alignment from Large Language Models (LLMs) while preserving their general utility.

b: POISONED OR BYZANTINE GRADIENT UPDATES IN FEDERATED/COLLABORATIVE LEARNING.

In federated and collaborative learning, there are additional problems due to the fact that model parameters are combined among many clients. Melis et al. [80] have shown that a malicious client can submit poisoned gradients to affect the shared model, while Fang et al. [81] and Bhagoji et al. [71] have studied the impact of small manipulations of gradients in order to create a biased or degraded global model. Aghakhani et al. [72] and Xie et al. [76] have described the same type of threat in a federated backdoor context, i.e., the possibility of manipulating the space of model parameters to produce a phenomenon similar to classical data poisoning. In this case, the integrity of the global model is compromised, and the reliability of the decentralized training process is called into question.

Gradient leakage/training data reconstruction. In addition to manipulating the model, the adversary can exfiltrate private data from the training dynamics. Chen et al. [92] have proposed GAN-Leaks, which demonstrates how a malicious client in a federated learning setting can reconstruct a sensitive image from the shared gradients. However, the majority of these types of attacks are based on strong assumptions regarding the attackers' ability to obtain gradients of the model, the model's architecture, and the training configuration. These assumptions may not be valid in most real-world applications. The adversaries' ability to effectively exploit these attacks can be diminished through the use of defenses such as; Gradient Clipping, Noise Injection, and Secure Aggregation. There is an open challenge in understanding the scalability of this type of attack to large models and heterogeneous datasets.

2) DEFENSES

The defense mechanisms applied during the training phase address both integrity and confidentiality concerns using complementary techniques.

Adversarial training and robust optimization expose the model to adversarial perturbations during training in order to increase the model's robustness. Madry et al. [8] formalized this approach with Projected Gradient Descent (PGD)-based inner maximization, and Tramèr et al. [93] generalized it using ensemble adversarial training. Gowal et al. [41], Liu and Lai [94], and Sen [95] have improved the efficiency of training and the generalization of robustness, showing

that robust optimization can provide protection against both test-time evasion and some forms of poisoning. However, Bondok et al. [99] examine the vulnerabilities of *Adversarial Training* (AT) in *Federated Learning* in order to protect *Smart Grids*. The researchers argue that while AT is an acceptable defense against sophisticated, domain-specific evasion attacks in time series prediction, AT does not always generalize well to new evasion attacks; they present an alternative evasion strategy to create perturbations based on the non-IID (Independent and Identically Distributed) of power consumption data to harden models that can be evaded. The attack creates statistically acceptable perturbations that cause the global server to produce a large forecasting error. Their research illustrates the major flaw in the existing defenses: the robust training methods developed for static image data sets are not inherently secure for the temporal, high-risk infrastructure of smart grids.

Robust and trust-weighted aggregation has the objective of protecting the integrity of federated learning updates in Byzantine environments. In order to reduce the impact of malicious gradients on federated learning updates in a Byzantine environment, several aggregation methods have been proposed: e.g., Krum [88], Trimmed Mean [87], FLTrust [86]. These aggregation methods are being improved upon with the use of dynamic weighting (adaptive) mechanisms [85], [89], [100] that can adjust their weights based on the behavior of each participant and dynamically penalize updates from participants who are outliers so that the quality of the global model remains within an acceptable range even when attacked. Gad et al. [101] focus on the ''Privacy-Utility-Efficiency'' dilemma for bandwidth-constrained IoT networks. The authors present a joint framework using *Joint Knowledge Distillation* (JKD) and *Differential Privacy* (DP) to create an efficient learning protocol that also provides secure communication. Unlike sending high-dimensional model weights, which consume bandwidth and leak client information, clients send their distilled knowledge to a central teacher model. In addition to reducing the bandwidth consumed by the communications layer via JKD, using DP noise when creating distilled knowledge prevents an attacker from reverse-engineering sensitive client data. Therefore, the combination of these two techniques significantly reduces the bandwidth used by the communications layer while still meeting the privacy budget constraint (ϵ).

Certified and provable robustness establishes formal security guarantees. Hein and Andriushchenko [57], Steinhardt et al. [34], Wong and Kolter [39], and Lecuyer et al. [90] have developed mathematical frameworks to quantify upper bounds on worst-case perturbations. In contrast, Cohen et al. [40], Wang et al. [96], and Zhang et al. [97] have generalized randomized smoothing and established provable robustness certificates. These methods provide measurable evidence of the integrity and confidentiality of deployed systems, establishing reliable guarantees of their trustworthiness.

3) SUMMARY OF FINDINGS FROM V-B

- Training is a significant point of vulnerability regarding the integrity of the model: the manipulation of the parameters of a model during its training (via backdoor, supply chain contamination, or poisoned gradients) can definitively modify the behavior of the model covertly.
- Federated learning amplifies the range of possible attacks: a malicious actor in a federated learning environment can generate a global model. This is biased toward his/her interests by submitting poisoned gradients that will be processed by all other actors.
- Gradient leakage represents a previously unknown form of confidentiality breach: even if an adversary does not have direct access to the data used to train a model, he/she can nevertheless recover sensitive training samples from the gradients shared by the actors.
- Exposure to adversarial perturbations during training increases resistance: by exposing a model to adversarial perturbations during its training, we increase its resistance to evasion attacks at testing time and to poisoning attacks.
- Robust aggregation is proven and trust-based methods increase confidence: certified robustness and trust-weighted aggregation allow us to establish quantitative measures of integrity that can be used to guarantee the correctness of our learning processes formally.

C. STAGE 3: MODEL EVALUATION / TESTING

After a model is trained, it must undergo systematic evaluation, auditing, and testing prior to deployment. Although the goal of this evaluation phase is to increase confidence and trust in the performance of the model, evaluation itself is yet another adversary interface for launching various attacks. In particular, while defenders use the evaluation phase to assess and strengthen the robustness of their models, adversaries use the evaluation phase as a means to test, optimize, and refine their own attacks. Therefore, evaluation serves two purposes: it acts as a diagnostic tool for identifying and characterizing vulnerabilities and as a rehearsal arena for developing and optimizing attacks. Figure 7 outlines the attacks and defenses pertinent to this stage.

1) ATTACKS

a: EVASION ATTACKS

Evasion is the most common type of adversarial attack. The attacker generates test-time input data that induces incorrect predictions by the model without visibly modifying the input data.

White-box evasion provides the attacker with complete knowledge of the model's architecture and gradients. Goodfellow et al. [3] first showed that deep neural networks are vulnerable to small, imperceptible changes in the input data through the fast gradient sign method (FGSM). Kurakin et al. [7] generalized FGSM to multiple iterations

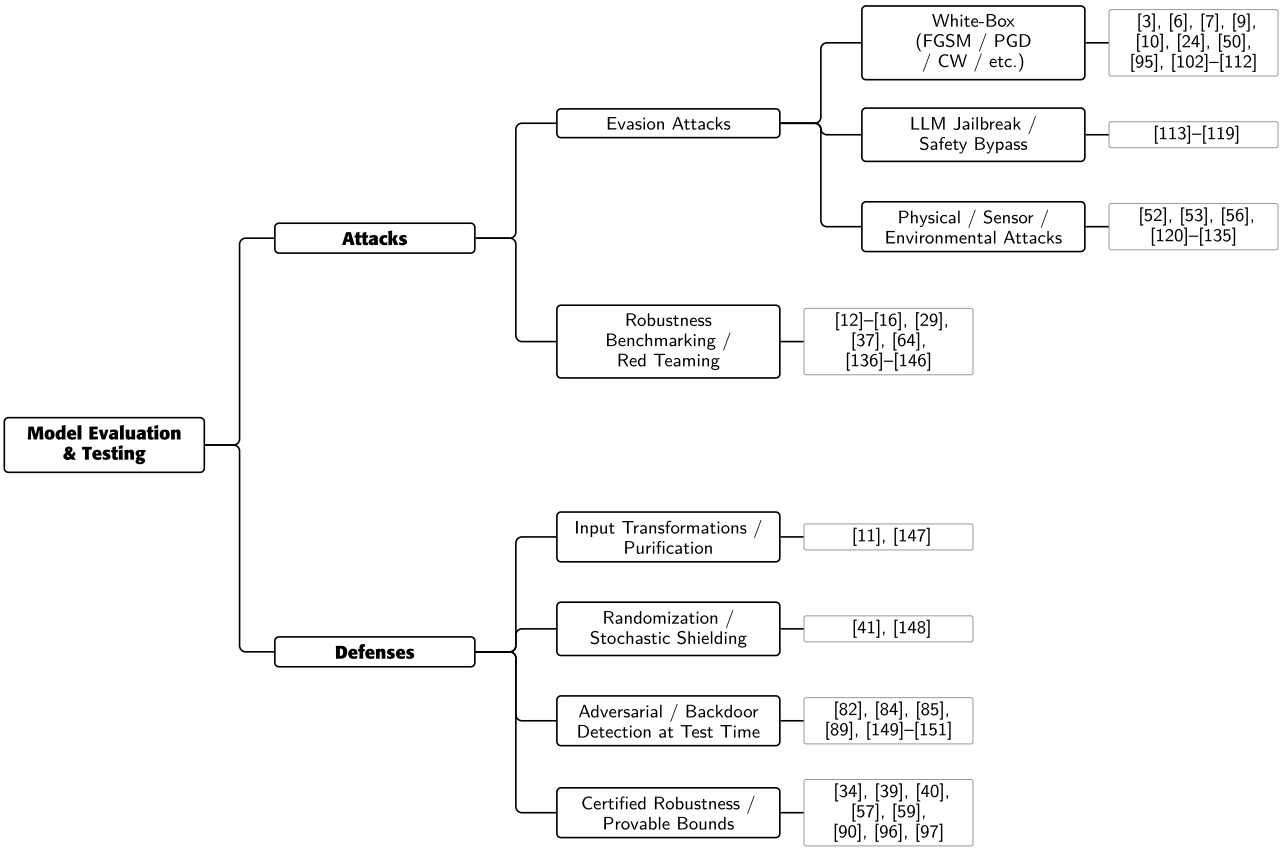

FIGURE 7. Stage 3 – model evaluation & Testing: Attacks and defenses taxonomy. Evasion attacks are presented here under controlled, white-box assessment conditions; black-box evasion under operational constraints is covered in Stage 4 (Figure 8).

of attacks. Carlini and Wagner [9] developed optimization-based methods to generate adversarial examples that were indistinguishable from the original example. Chen et al. [107], Liu et al. [105], and Sen et al. [95] further developed these ideas using elastic-net regularization, transferable attacks, and enhanced gradient methods. Athalye et al. [10] showed that some defenses (which rely on gradient masking) are actually ineffective. Croce and Hein [111] and Chakrabarty [112] then developed and improved minimal-perturbation attacks. Collectively, the results demonstrate that the decision boundary of a model is highly susceptible to adversarial perturbations, thereby posing a significant integrity risk to the model.

Jailbreak and safety-bypass attacks extend evasion attacks to generative and language models. Initial studies showed that through prompt injection (a process of manually inputting a specific phrase) and persona manipulation (manipulating the "personality" of the model), an attacker was able to circumvent the model's internal "safety" filters and generate adversarial inputs [114]. The attacker's method of generating adversarial examples changed as the defensive capabilities of LLMs improved; this transitioned from using manual processes to using automated optimization methods to create adversarial examples [152], [153]. A study by Kumar [115] both indicated that various types of gradient-based and evolutionary algorithms can effectively and systematically optimize for the maximum reward from a model to generate highly complex and sophisticated adversarial suffixes. There is also an emerging risk associated with agentic systems, which have been explored by Zhang et al. [119], and Sheth et al. [154]; they both illustrated that an attacker can use a variety of techniques to alter the behavior of an autonomous agent so that it does not comply with the policy of the system. Yang et al. [155] developed an automated approach called SneakyPrompt, to find and exploit vulnerabilities in generative text-to-image (T2I) models through using reinforcement learning. In their view, jailbreaking the T2I model is a discrete optimization problem where they iteratively modified banned prompts to identify "sneaky" tokens which have the same semantic value as the original prompt but do not trigger false positives by appearing harmless when input into a classifier based on the text. Since they were targeting the latent space of text encoder models like CLIP [156], they showed how successful they could be against both open-source models (Stable Diffusion), and closed-source systems (DALL·E 2). Even though they were able to achieve great success with finding keyword-based filtering methods, however, they identified significant limitations of the method to detect multi-modal, image-based safety mechanisms; and its low query efficiency.

Physical, sensor, and environmental attacks move the adversarial perturbations into the physical world. Sharif et al. [120] first demonstrated that accessories could be used to generate adversarial perturbations in face recognition. Brown et al. [52] demonstrated that printouts of adversarial patches could be created for image classification. Eykholt et al. [56] showed that stop signs could be made physically invisible to an autonomous vehicle's cameras. Since then, additional works have systematically demonstrated that adversaries can exploit a diverse array of hardware ranging from microphones via hidden audio commands to LiDAR and 3D perception systems via adversarial meshes and optical projections, proving that the physical environment itself serves as a highly viable, multi-sensory vector for model exploitation [53], [126], [128], [129], [130], [131], [132].

Robustness Benchmarking and Red Teaming. The evaluation phase has moved beyond simply being a means of passively testing models to actively probing them using adversarial methods. Biggio and Roli [13], [37] and Akhtar and Mian [16] started the trend toward systematic benchmarking of model robustness in adversarial environments. Dong et al. [136], Croce and Hein [12], [29], and Awadid and Robert [137] provided a framework for creating large-scale robustness benchmarks and reproducible pipelines for evaluating robustness. Marulli et al. [138], Ishii et al. [140], and Goyal et al. [141] broadened this effort into formal taxonomies and methodologies for evaluating robustness. Recent meta-surveys [64], [142], [144] integrated the individual parts of the robustness evaluation process into cohesive frameworks and linked technical evaluations to policy-focused evaluations.

As a result, the adversarial red teaming process has transitioned from a solely defensive necessity to an assurance mechanism: it connects technical evaluation to accountability and, in doing so, has positioned evaluation as a core component of AI safety.

2) DEFENSES

Defenses in the evaluation/testing phase focus on reducing the damage caused by adversarial attacks after a model has been deployed for inference.

Transformation and purification techniques pre-process incoming data to remove possible adversarial perturbations. Guo et al. [147] proposed discrete transformations (e.g., bit-depth reduction and JP compression) and randomizations (e.g., random resizing and padding), creating a stochastic target to prevent attackers from fixing pixel coordinates. Xie et al. [11] employed generative models to reconstruct input data, projecting it onto the clean data manifold for input reconstruction. Such methods examined image denoising, compression, and projection as ways to reduce the sensitivity of models to adversarial attacks. However, Athalye et al. [10] demonstrated that all three techniques relied on gradient masking (gradient obfuscation) to prevent attackers from locating vulnerabilities, but allowed adaptive attackers to bypass these defenses using BPDA (Backward Pass Differentiable Approximation). Although modern diffusion-based purifiers such as DiffPure [157] and ADBM [158] have addressed this issue by enforcing strict manifold projection, they have introduced a new trade-off by increasing inference latency by orders of magnitude.

Randomization and Stochastic Shielding add random elements to the inference process to prevent deterministic exploitation. Pinto et al. [148] introduced Robust Adversarial Reinforcement Learning (RARL), which jointly trains a protagonist agent against a destabilizing adversary that applies random forces to the system dynamics, thereby forcing the policy to learn stability under diverse perturbations. Complementing this, Gowal et al. [41] demonstrated that training with generated data, even simple Gaussian-sampled inputs, can significantly improve robustness against norm-bounded attacks by artificially expanding the training distribution and reducing the robust-accuracy gap.

Adversarial and Backdoor Detection Systems focus on the real-time detection of malicious inputs. The early work established the basic principles for detection using statistical anomaly-detection techniques: [149] and [150] used reconstruction error generated by an autoencoder to detect anomalous inputs (i.e., inputs that do not reside within the manifold of clean inputs) while [82] used spectral signatures (i.e., clusters in the covariance matrix of latent representations) to expose hidden backdoor triggers. Building upon these methods, [84] proposed the STRIP method that used entropy signatures to determine if a trigger was malicious. Specifically, it determined whether an input was malicious (or not) by assessing the unpredictability of its predictions when the input was perturbed. As machine learning began to adopt decentralized architectures, [85] and [151] extended the capabilities of previous detection frameworks to Federated Learning environments. In doing so they addressed the security gap that exists when using raw data is not feasible in environments in which the raw data cannot be accessed. These authors filtered malicious model updates based on anomalies in the distribution of model weights. Although there have been several advancements made in regards to the capability to detect malicious activity in models; many open questions exist. For example, *adaptive attackers* can include the detection mechanism within their adversarial loss function(s), thereby allowing them to evade detection. Another major challenge remains, i.e., the overhead imposed by performing real-time spectral analysis or reconstructing inputs results in a significant and persistent latency/security tradeoff that hinders scalability in high-throughput applications.

Certified Robustness serves as a key bridge connecting the training and testing phases. Hein and Andriushchenko [57], and Steinhardt et al. [34] have provided exact verification bounds on smaller networks, but Cohen et al. [40] developed formal robustness certificates through randomized smoothing that provide quantifiable risk estimates at the time of inference, thus converting the evaluation process into an

assurance process. A critical open problem remains the certification of robustness for NLP models. Certification techniques for continuous domains (like computer vision) have improved significantly in recent years. The challenge lies in applying these same guarantees to discrete input spaces in NLP. Adversarial attacks on NLP models can include a variety of discrete transformations, such as synonym substitution, paraphrasing, or inserting new tokens. These discrete transformations are difficult to model using traditional, norm-based methods of bounding changes in the input space. Jia et al. [159] proposed a certification framework that provides provably guaranteed protection against word substitutions or synonym-based perturbations. Their approach leverages Interval Bound Propagation (IBP) over word embeddings. However, compared to computer vision models, certification frameworks for NLP are much less scalable and provide far fewer guarantees when applied to real-world language data.

3) SUMMARY OF FINDINGS IN V-C

- Testing serves as an assurance and a defense mechanism for the defender while it is a refinement mechanism for the adversary in the evaluation phase.
- Testing validates robustness but also provides an opportunity for adversaries to identify new attack surfaces that they will use to improve their evasion techniques and optimize their attacks.
- Jailbreak and safety-bypass attacks expand the threat landscape beyond simple misclassification. As such, the evaluation of generative models now includes governance issues since adversaries can manipulate the alignment constraints, policy filters, and agentic behaviors to produce policy-violating or unsafe output.
- Physical and sensor-based attacks verify that there is an intersection between the vulnerability of AI models to attacks in a controlled digital testbed and in the real world. Thus, evaluation must take into consideration environmental noise, sensor error, and physical perturbation, which can reliably cause failure in a real-world setting.
- Robustness benchmarking has progressed from ad hoc testing to formalized, structured, and systematic evaluation. Current robustness benchmarking standards, taxonomies, and meta-surveys include technical analyses of attacks along with safety and governance considerations, thus solidifying robustness evaluation as an essential component of AI assurance.
- Detection systems, certified robustness, randomization, and purification are examples of defensive mechanisms that can be employed to protect against attacks during the testing process. As such, these mechanisms provide the means to prevent, detect, and quantify the threats that exist at inference time and serve as the means to bridge the gap between the testing and deployment phases of a model.
- Furthermore, collectively, these findings demonstrate that testing is no longer a passive quality control measure. Rather, it is an active security and assurance measure that links adversarial robustness, safety and compliance throughout the entire AI development life cycle.

D. STAGE 4: MODEL DEPLOYMENT & MONITORING

1) ATTACKS

The most risky phase of the machine learning lifecycle, as seen in Figure 8, is once a model has been deployed, and users can submit inputs to the model through an interface. In addition, attacks from Stage 3 can be exploited at this stage while defenses are operationalized. This is because the model's behavior will be visible, allowing users to elicit information about how the model makes decisions, circumvent safety policies, and possibly induce unintended behavior.

Black-box evasion removes the assumption of having white-box information and uses either transferability or query-based estimations to generate adversarial examples. Papernot et al. [160] demonstrated that if a model can generate an adversarial example for one model (the surrogate model), the generated example will likely have similar success for other models. Chen et al. [161] used this idea to develop the Zoo algorithm, which uses zeroth-order optimization to estimate gradients from the continuous confidence values returned by an API. Many real-world systems only return discrete labels (e.g., "Cat" or "Dog"); therefore, Brendel et al. [163] proposed the Boundary Attack. The Boundary Attack is a large perturbation that is minimized at each iteration while tracking the model's decision boundary. Although there is now a substantial body of research focusing on improving the query efficiency of both score-based and hard-label attacks [164], [165], [167], [168].

Model Extraction and Policy/Filter Circumvention. Deployed model interfaces can become sources of information leakage. An adversary can send well-designed inputs to a model, and based on the model's responses, determine its decision boundary (model extraction), replicate its learned policy, or systematically identify and bypass safety filters. Examples of such adversarial querying against large language models (LLMs) include work such as [113] and [114]. Researchers have also demonstrated how policy-steering and jailbreaking techniques used against large language models can evolve into system-wide compromises. Fathima [117] and Liang et al. [182] demonstrate scalable extraction and filter-circumvention strategies against black-box APIs, establishing the dual nature of this threat as a confidentiality breach when proprietary parameters are inferred, and a governance breach when policy filters are systematically neutralized.

Membership Inference and Model Inversion. Although adversaries cannot obtain a model's parameters, they can still infer information about a model's training data from its

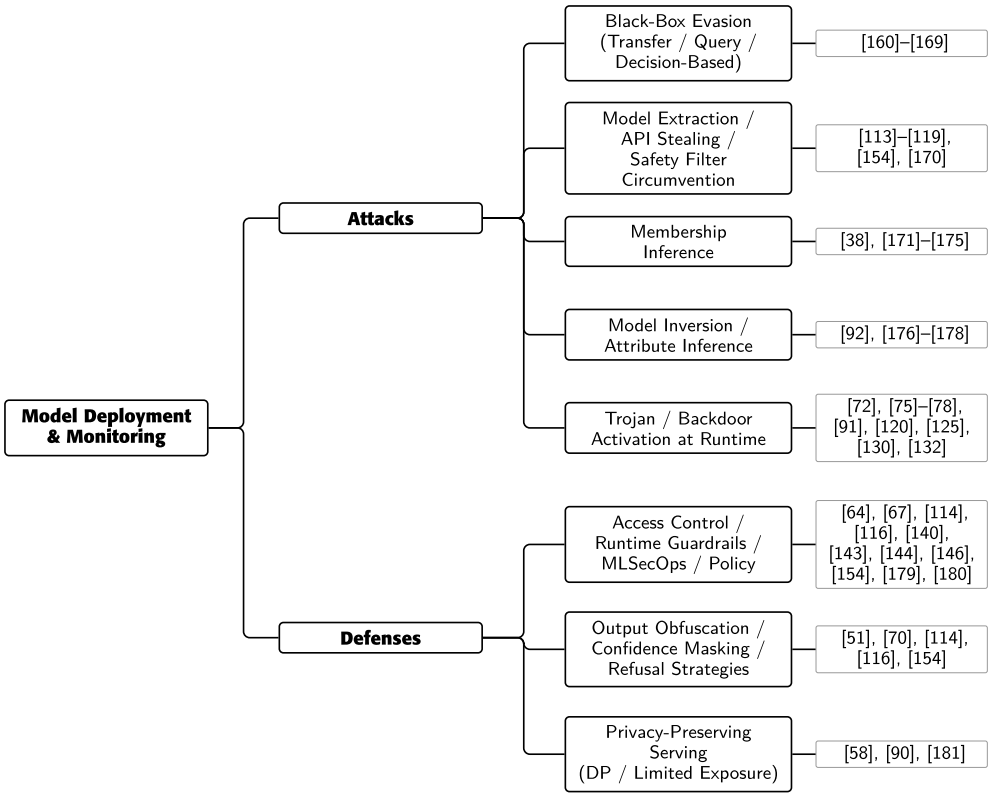

FIGURE 8. Stage 4 – Model deployment & Monitoring: Attacks and defenses taxonomy. Black-box evasion attacks, which require only query access to a deployed model, are placed here under operational constraints (rate limits, decision-based feedback only), complementing the white-box evasion assessment presented in Stage 3 (Figure 7).

outputs. Shokri et al. [171] first formalized membership inference attacks, demonstrating that confident predictions often disclose whether a sample was included in the training set. Salem et al. [173] and Nasr et al. [174] extended these attacks to realistic settings with minimal prior knowledge, and Carlini et al. [38], [175] provided standardized methodologies for evaluating the amount of leakage from a model's outputs. Fredrikson et al. [176] and Song and Shmatikov [183] established the complementary threat of model inversion, reconstructing representative training samples or sensitive attributes from a model's outputs. Chen et al. [162] applied gradient-based inversion and reconstruction in federated learning. These works illustrate how modern inference interfaces can unintentionally compromise the privacy and confidentiality of sensitive data, thereby posing significant challenges to regulatory compliance (e.g., GDPR, HIPAA).

Trojan/Backdoor Activation at Runtime. Deployment is also when latent threats become apparent. As previously mentioned, if a model was previously poisoned or fine-tuned with a hidden trigger, the deployed system provides a mechanism for activating the attacker-controlled backdoor. Trained backdoors can remain dormant until a specific trigger (e.g., visual patches, phrases, or tokens) is presented [75], [91]. Distributed activation mechanisms shown by [11], [77], and [78] also demonstrated that these triggers can persist even after pruning and quantization. Demonstrations of physical-world manifestations such as those of Sharif et al. [120], Thys et al. [125], Lovisotto et al. [130], and Wei et al. [132], which confirm that the risk extends beyond simulations. Therefore, runtime backdoor activation represents a compromise of both the integrity and safety of production systems, which undermines the predictability of the latter.

2) DEFENSES

The socio-technical nature of production ML systems makes defense mechanisms in this phase primarily focused on engineering controls, policy enforcement, and monitoring. It is important to remember that most of the mechanisms presented here function as external "wrappers." They cannot remove the inherent flaws present in the model. The primary goal is to increase the difficulty and time needed for an attacker to exploit the vulnerabilities in the model.

Access control, runtime guardrails, and MLSecOps practices specify who can query a model, how often, under what circumstances, and with what level of logging. The use of operational frameworks that enforce endpoint validation and query throttling to heuristically detect and block malicious traffic was implemented in the work of Patel et al. [70], Narajala and Narayan [181], and others as the basis of the standardized AI assurance pipeline suggested in the institutional guidelines developed by Tabassi et al. [146] and

Vassilev [64]. However, as demonstrated by Sheth et al. [154] and Jin et al. [184] applying semantic safety filters to the external filter has significant structural limitations, and adversaries may employ sophisticated role-playing jailbreaks or multi-turn social engineering to cognitively bypass the input validation layer. These solutions protect operational integrity by ensuring usage aligns with the declared safety boundaries, and protect policy compliance by enforcing usage restrictions.

Output Obfuscation, Refusal Strategies, and Confidence Masking intentionally limit the amount of information a model exposes regarding its outputs. Papernot et al. [51] proposed defensive distillation as an early example of controlled output transformations such as gradient masking; however, such methods are easily defeated by adaptive attacks that can calculate gradients. In addition, although Xu et al. [114] and Kumar [115] analyzed refusal frameworks for large language models that limit semantic interpretability and withhold logits or confidence scores to limit jailbreak susceptibility, both acknowledged that withholding logits or confidence scores simply changes the threat model from a white-box to a decision based black-box, increasing the number of queries allowed per attack but does not eliminate the vulnerability.

Privacy-Preserving Serving uses differential privacy (DP) and least-privileged protocols. Papernot et al. [58] and Lecuyer et al. [90] formalized the application of DP to prevent inversion. However, Narajala and Narayan [181] show that implementing strict DP guarantees in secure AI deployments often require significant compromises in model utility and inference latency, possibly making them unfeasible for many real-time applications.

Finally, continuous monitoring and trigger scanning, like those presented by Gao et al. [84] and Nguyen and Tran [77] serve as real-time integrity checks. These methods, however, are inherently reactive and therefore struggle to distinguish between benign distribution shifts (drift) and stealthy, low-confidence adversarial attacks, leading to a high false-positive rate that can cause security personnel to become fatigued.

3) SUMMARY OF FINDINGS FROM V-D

- Deployment provides an operationalization of the threat environment, and live systems expose a simultaneous threat to confidentiality, integrity, and governance, thereby transforming the AML research problem into an organizational security problem.
- Model extraction and policy bypass present a double jeopardy; attackers are able to either steal confidential intelligence (confidentiality) by extracting models or subvert usage limitations (governance) through policy bypass.
- Inference-Time Privacy Attacks Remain Underregulated: Membership inference and inversion continue to disclose private or sensitive information regardless of model abstraction layers.
- Runtime Trojan Activation Unifies Safety and Integrity Failures: Hidden behaviors introduced during model poisoning can be activated after deployment, thereby degrading user trust in critical systems.
- MLSecOps and Runtime Assurance Bridge Technical and Organizational Defenses: Security controls such as rate limiting, logging, and least privileged interfaces now support algorithmic defenses and signify a convergence between AI assurance and classical cybersecurity.

E. THE "NO FREE LUNCH" OF ROBUSTNESS

It is essential to note the trade-offs in the claims of efficacy for these defenses. As illustrated in Table 5, no single method offers complete protection, and each carries distinct trade-offs as detailed below. The overhead ranges reported in Table 5 should be interpreted as *indicative general values rather than benchmark measurements*, since the actual computational cost can vary substantially depending on the model architecture, dataset, threat model, and evaluation configuration used in each study.

- Adversarial Training remains the state-of-the-art for empirical robustness but introduces a significant computational tax (generating adversarial examples at every training step) and often degrades clean accuracy by 10–20% on difficult tasks (e.g., ImageNet).
- Certified Defenses such as Randomized Smoothing offer provable guarantees but shift the cost to inference time, requiring hundreds of forward passes per prediction to estimate the smoothed classifier, rendering them impractical for real-time applications like autonomous driving.
- Alignment Techniques such as RLHF and Guardrails act as ''soft'' defenses. They do not remove the underlying vulnerability (the model's capacity to cause harm), but rather mask it behind a refusal policy. Consequently, they remain vulnerable to adaptive ''jailbreak'' attacks that leverage competing objectives or role-play scenarios to bypass the safety filter.

F. LIFECYCLE VIEW VERSUS SECURITY PROPERTY VIEW

The intersection of the holistic view of the AML lifecycle and the traditional security triad (*availability, integrity, confidentiality*) reveals several common security concerns across the phases of the ML pipeline.

Integrity failures occur early on in the lifecycle and continue through all phases. Examples include poisoning of data [33], [75] and backdoors inserted during data collection; Byzantine gradient manipulation during distributed training [71]; jailbreaking and other forms of evasion during inference [3]; and Trojans activated during deployment [78]. The common thread across these phases of the AML lifecycle is the presence of an *integrity failure*, indicating that the model does not operate as originally intended. In fact, integrity failures can be described as the model operating outside its original intended *policy* or *task*.

TABLE 5. Approximate computational and operational overhead of representative AML defenses. Values denote indicative ranges synthesized from representative studies and may vary substantially by architecture, dataset, and threat model.

Defense Category	Technique	Training	Inference	Memory	Notes / Trade-offs
Adversarial Training	PGD / TRADES	3×–10×	None	Moderate	Strong empirical robustness; may reduce clean accuracy; expensive retraining.
Certified Robustness	Randomized Smoothing	2×–5×	10×–100×	High	Probabilistic guarantees; limited to small perturbations.
Formal Verification	MILP / SMT-based	Very High	Offline	High	Poor scalability; mainly for small safety-critical models.
Input Preprocessing	Denoising / Compression	None	Low	Low	Often bypassable by adaptive attacks.
Ensemble Defenses	Model Ensembles	$N\times$	$N\times$	High	Improves robustness but increases latency and cost.
Runtime Detection	Adversarial Detectors	None	Low–Moderate	Low	Detects suspicious inputs; may incur false positives.
RLHF (LLMs)	Preference Optimization	High	None	High	Improves alignment; no robustness guarantees.
Guardrails / Filters	Rule-based or Classifier Filters	None	Low	Low	Reduces harmful outputs; bypassable via prompt reformulation.

Confidentiality and privacy risks exist primarily at the training boundary and the deployment boundary. Gradient inversion, Membership Inference Attacks (MIAs), and Model Extraction Attacks [92], [171], [175] demonstrate how maliciously designed attacks can successfully obtain sensitive information about individuals or organizations through indirect means (access to internal model components or outputs). To address this concern, researchers and practitioners use techniques such as differential privacy [58], controlled output exposure, and query governance [114], which closely mirror confidentiality-based protection mechanisms used within cybersecurity.

Availability and safety/governance provide the final elements of the entire picture. Poisoning or Resource Denial-of-Service (DoS) type attacks [71] target degrading functional capabilities or inducing DoS. Violations of safety/governance policies (e.g., jailbreaks, policy circumvention) [113], [114], [154] signify a failure of compliance rather than capability. As a result, a new area of study known as *MLSecOps* has emerged, including runtime policy enforcement, refusal/redaction strategies, tool usage auditing, and human-in-the-loop escalation [64], [144], bridging operational assurance and technical robustness.

Summary: The lifecycle perspective informs us where we need to intervene in the ML pipeline to prevent attacks, such as data ingestion, model training, testing, etc.; the security property perspective informs us what we are trying to protect, the integrity, confidentiality, availability, and governance of our systems. Therefore, developing a mature AI security posture requires understanding both perspectives and applying them together to achieve end-to-end resilience.

VI. DOMAIN-SPECIFIC CASE STUDIES

In this section, we present specific case studies across five key domains to ground theoretical discussion in practical impact: autonomous vehicles, healthcare, financial systems, large language models (LLMs), and Internet of Things (IoT) systems. Each case highlights representative attack vectors, measured impacts, and mitigation strategies supported by peer-reviewed literature.

A. CYBER-PHYSICAL MOBILITY AND AUTONOMOUS VEHICLE

Adversarial attacks against autonomous vehicles (AVs) have exposed critical vulnerabilities in perception systems, particularly in traffic sign recognition and object detection.

Traffic Sign Physical Manipulation. Eykholt et al. [56] introduced Robust Physical Perturbations (RP2) against a traffic sign classifier trained on the LISA dataset in a white-box setting. They successfully achieved complete targeted misclassification in a controlled setting and 84.8% in field test drive-by scenarios. For example, stop signs were misclassified as speed limit signs using perturbations optimized under robust physical constraints. The authors noted that their results focused on a standalone classifier rather than the entire stack of an autonomous vehicle system, including detection, tracking, and sensor fusion. The study [185] found that the system-wide misrecognition rate for tested vehicles sold in the U.S. in 2023 was approximately 33%. This suggests that while full-autonomy stacks may provide additional resiliency, the legacy perception modules remain vulnerable. Similarly, LiDAR spoofing attacks [109] have been shown to project fake objects into point clouds, significantly reducing object detection accuracy.

Object Detection Availability-Based Attacks. In addition to misclassification, attacks against the non-maximum suppression (NMS) stage can significantly decrease system availability. Wang et al. [186] demonstrated that the Daedalus attack can compress bounding-box regressions to bypass NMS and achieve up to 99.9% false positives

and mAP≈0 on YOLO-v3 and RetinaNet (COCO dataset), demonstrating this using physically printed posters. Additionally, Shapira et al. [187] presented Phantom Sponges, universal perturbations designed to overwhelm detectors with false positives in real-time. Like the stop-sign attack, the purpose of these methods is to slow down or overload collision-avoidance pipelines rather than directly mislabeling objects, thereby gradually decreasing system reliability and user confidence.

Defense mechanisms include:

- Multi-sensor fusion: Combining camera, LiDAR, and radar inputs reduces reliance on any single modality, thereby mitigating single-vector attacks [128].
- Temporal consistency checks: These systems detect inconsistencies across video frames to identify adversarial outliers that appear transiently [188].
- Robust model training: This incorporates adversarial examples into training pipelines to enhance generalization against physical perturbations.

Empirical results demonstrate that multi-sensor fusion can reduce adversarial misclassification from over 90% to under 20% for traffic signs, illustraoutputs by routing low-confidence predictions to clinicians, which is ting the necessity of multimodal redundancy [56], [128].

B. HEALTH CARE AND MEDICAL AI SYSTEMS

AI systems designed to aid medical professionals are highly susceptible to imperceptible perturbations, which pose severe risks in high-stakes clinical decision-making. Finlayson et al. [189] showed that white-box PGD attacks could be performed using fine-tuned ResNet-50 models for both chest X-rays and dermatology classifications. The results of this study were that the attackers were able to generate false diagnoses with confidence and create realizable skin patches that could be applied to mimic a melanoma for dermatologists to visually examine.

The sensitivity of decision boundaries to minor changes (catastrophic failures) is also due to the significant statistical differences between medical and natural images. As pointed out in [190], Medical image databases have smaller intra-class variances and larger inter-class similarities than natural images, as they were obtained under much stricter conditions than natural images. This makes them very sensitive to small perturbations, resulting in catastrophic failure modes. A recent study reported that White-Box Attacks reduced accuracy from more than 87% on clean data to almost 0% across several architectures and modalities [191].

Methods for increasing the robustness of medical AI include:

- Adversarial training: Enhances a model's resilience against real-world perturbations in complex modalities like CT and MRI scans [192].
- Quantifying uncertainty: Techniques such as Bayesian inference and Monte Carlo dropout help identify instances where a model has low confidence, triggering human review [193].
- Multi-reader validation: Provides cross-validation of model output by routing low-confidence predictions to clinicians, essential for radiology workflows.

Incorporating these methods enhances resilience against attacks while concurrently building clinician trust in AI-driven diagnostics.

C. FINANCIAL SYSTEMS AND FRAUD DETECTION

Financial institutions rely heavily on ML for credit scoring and fraud detection, making them prime targets for evasion and poisoning attacks. Unlike regulatory anti-money laundering, adversarial attacks in this context target the integrity of the predictive models themselves.

Carminati et al. [194] evaluated evasion attacks against four types of fraud detectors (BankSealer, Random Forest, Neural Network, Hybrid) using proprietary transactional datasets from a single financial institution. Under white-box conditions, evasion rates are almost entirely successful against specific types of fraud detectors. However, in more realistic black-box settings, evasion rates dropped to 60-75%. In particular, the unsupervised fraud detector (BankSealer) was shown to be more resilient than both the supervised random forest and supervised neural network-based fraud detectors. Complementary work has demonstrated that evasion can also be achieved through backdoors created via false data injection [195]. Moreover, Paladini et al. [196] have demonstrated, in their research using two real-world, anonymized banking datasets, that even the most advanced ML-based fraud detection systems are highly vulnerable to manipulation by an adversary. While the threat models in grey-box and black-box conditions may show that the system detects between 55% to 91% of the attacks; the exploitation window is still two months long at which time the adversary can extract funds from the account(s) on a systematic basis over those two months, with the average amount extracted ranging from €41,995 to €388,342 per victim, before when the system identifies the anomaly.

Defenses for adversarial threats to financial systems include:

- Adversarial Training: Training financial models with perturbations applied to transaction data samples to harden decision boundaries against evasion [197].
- Surveillance Analytics: Identifying anomalies in the time series of trades or transactional metadata to detect coordinated poisoning attempts [197].
- Data Provenance: Rigorous validation to identify poisoned data or samples with low confidence scores before they impact financial decision-making.

Given the potential for significant economic loss, implementing robust MLSecOps frameworks is critical for financial organizations.

D. NLP CONTENT SYSTEMS

Large Language Models (LLMs) face unique security risks, including prompt injection, jailbreaking, and training data poisoning.

Zou et al. [63] used white-box access to Vicuna models to generate adversarial suffixes, achieving an average success rate of 88% for Vicuna-7B and 56% for LLaMA-2-7B-Chat (AdvBench) and showing some transferability to GPT-3.5 and GPT-4. The attacks were computationally expensive, requiring many resources (multiple A100 GPUs and hours of processing time per suffix).

In addition, Carlini et al. [198] and Liang et al. [182] have shown that well-designed prompts are capable of extracting memorized training data from large language models. They demonstrated how, by repeatedly asking questions, users can potentially recover sensitive information sequences contained within the training dataset. Adi et al. [199] proposed an embedding watermark that utilized a backdoor trigger in a deep neural network that would cause the model to give a pre-defined output given the appropriate input. They demonstrated how a method typically considered a poisoning attack can be repurposed for intellectual property rights protection and model validation. In systems that use agents or tools, indirect prompt injection [200] enables users to create adversarial instructions as part of the system's external content, thereby allowing them to take unauthorized actions. Figure 10 illustrates an example of an indirect prompt injection attack.

Key mitigation techniques include:

- RLHF (Reinforcement Learning from Human Feedback): Aligns LLM outputs with human safety preferences to reduce toxicity [63].
- Differential Privacy Training: Limits the LLM's ability to memorize and regurgitate sensitive training data [201].
- Automated Red-Teaming: Continuously testing LLM safety against evolving jailbreak heuristics.

While these measures significantly decrease risk, they do not entirely eliminate it. Ongoing research focuses on formalizing threat models for agentic workflows and multimodal interactions [154].

E. EDGE COMPUTING AND INTERNET OF THINGS (IOT)

IoT devices are resource-constrained and physically accessible, making them highly susceptible to physical and over-the-air attacks. Bhagoji et al. [71] demonstrated that Federated Learning (FL) approaches in IoT can be compromised via poisoned gradients from one client, and that the security of the other nine clients can be compromised by model replacement attacks, resulting in an approximate 75% decrease in the targeted class's accuracy.

Yuan et al. [202] demonstrated how to embed inaudible commands into audible speech to hijack voice assistant systems using *CommanderSong* with an almost perfect success rate when utilizing Kaldi-based systems.

The main vulnerabilities include:

- Sensor Spoofing: Disrupting physical sensors via optical, magnetic, or acoustic interference [135].
- Model Tampering: Injecting malicious models into devices via compromised over-the-air updates [131].
- Firmware Attacks: Exploiting weak attestation to modify device firmware.

Defending against these attacks includes using secure enclaves to protect model weights [203], implementing edge-based anomaly detection to identify adversarial payloads [204], and developing robust federated learning techniques such as robust aggregation [205]. Thus, the design of both physical security and software robustness must be coordinated.

VII. EVALUATION METHODOLOGIES AND BENCHMARKING

A. STANDARDIZED EVALUATION FRAMEWORKS

1) ROBUSTNESS EVALUATION METRICS

To evaluate how well a model resists attacks, Cohen et al. [40] defined certified robustness using randomized smoothing. They measured the proportion of test points in which the model's output was provably unchanged within an ℓ_2 radius. Zhang et al. [59] studied the fundamental trade-off between standard accuracy and robust accuracy and concluded that it must be explicitly optimized and cannot be assumed to follow from optimizing accuracy. Likewise, Carlini et al. [38] stated that without a specific definition of the threat model, the perturbation budget, and the attack strength, any measurement of robustness is meaningless.

All three research efforts converged upon a set of three metrics that have become the foundation for evaluating adversarial robustness. The Attack Success Rate (ASR), defined in Eq. (1), measures the effectiveness of attacks (the higher the value, the worse for the defender). Robust Accuracy (RA) and Certified Accuracy (CA), defined in Eqs. (2) and (3) respectively, measure the defender's ability to reliably predict outputs even under perturbations (the higher the value, the more resilient the model).

$$\mathrm{ASR} = \frac{N_{\mathrm{succ}}}{N_{\mathrm{att}}} \tag{1}$$

$$\mathrm{RA} = \frac{N_{\mathrm{corr,adv}}}{N} \tag{2}$$

$$\mathrm{CA} = \frac{N_{\mathrm{cert}}}{N} \tag{3}$$

where N_{succ} is the number of successful attacks, N_{att} is the total number of attack attempts, $N_{\mathrm{corr,adv}}$ is the number of correctly classified adversarial samples, N_{cert} is the number of certifiably correct predictions, and N is the total number of evaluated samples.

The limitations of all three metrics (ASR, RA, CA) are substantial. The reliance on ASR on the selected attack and its parameters means that ASR may be either underestimated due to weak or non-adaptive attacks, as well as being dependent on the attack used, and the attack's

parameters [10]. The use of RA limits it to a particular perturbation type and budget (for example, l_∞ at a particular value of ϵ). This results in benchmark overfitting and poor generalization to other types of perturbations or semantic distortions. However, CA provides formal guarantees only when using narrow threat models, and the certified radius of distortion is typically far below what would be considered practically relevant, and is primarily limited to classification tasks. In addition, these metrics have been developed for predictive models and therefore do not lend themselves easily to generative, multimodal, or agent-based systems: small input perturbations to LLMs do not always result in meaningful semantic changes; multimodal attacks operate across multiple and/or disparate input spaces; and failure in agent-based systems often occurs due to multi-step reasoning errors, tool misuse, or goal manipulation, rather than input noise alone. Furthermore, there is no unique ''ground truth'' for open-ended generation; discrete token perturbation budgets are undefined; and established benchmarks for comparing jailbreak, or agent compromise rates do not currently exist [206], [207]. These studies represent the first steps towards benchmarking; however, the development of standard metrics for multi-step, tool-augmented workflow robustness, safety, and system-level reliability remains an open research question, motivating new paradigms for evaluating these characteristics.

Metrics Evaluating Perturbation Quality. Zhang et al. [208] showed that LPIPS (Learned Perceptual Image Patch Similarity) has a much better correlation with human perception than ℓ_p-norm distances; and Wang et al. [209] proposed SSIM (Structural Similarity Index) to compare structural and luminance coherence. Heusel et al. [210] developed FID (Fréchet Inception Distance) to assess how much an image has changed with respect to its distribution. The following three measures of distance represent how effectively a perturbation can remain imperceptible (or produce visually natural distortion):

In this way, each of these metrics captures a distinct aspect of visual integrity: LPIPS emphasizes the deep feature alignment of two images, SSIM emphasizes the local texture consistency of two images, and FID emphasizes the statistical realism of two images.

2) EVALUATION PROTOCOL STANDARDS

Carlini et al. [38] support consistent and transparent adversarial attacks to evaluate robustness, which includes the specification of adversarial assumptions for attacks, configuration of an attack, and splitting the data into training and test sets. In NLP, similar concerns are found by McCoy et al. [211]. They have demonstrated that NLP models can rely heavily on heuristics even when trained on standard datasets, without diagnostic testing. A benchmark for evaluating robustness is proposed by Hendrycks and Dietterich [212] using the corruption benchmarks (CIFAR-10-C and ImageNet-C). Hendrycks et al. [213] extend this using ImageNet-A, which contains natural adversarial examples. The principles outlined above are operationalized by Croce et al. [29], who created *RobustBench*, a single, unified leaderboard for evaluating the robustness of deep learning models. The leaderboard enforces the same ε-budget and attack protocol for all submitted models.

Key standardization principles include:

1) Threat Model Specification: Clearly define adversarial capabilities (white-box or black-box) and constraints (e.g., ℓ_p-norm and ϵ budget) [38].
2) Dataset Standardization: Use common benchmarks (e.g., CIFAR-10/100, ImageNet, AdvGLUE) to ensure comparability [212], [213], [214].
3) Attack Parameter Consistency: Report identical hyperparameters (iterations, step size, and restarts) across methods [29].
4) Statistical Significance: Include confidence intervals and multiple seeds for credibility [40].
5) Reproducibility: Release code, checkpoints, and attack scripts to support independent validation [29], [38].

B. BENCHMARK DATASETS AND LEADERBOARDS

In order to provide a uniform way of comparing different types of robust models, Croce et al. [29] developed a common framework for testing and evaluating robust models with the name *RobustBench*. The framework allows them to compare and evaluate the performance of over 80 robust model types using the same set of AutoAttack-based tests. In addition to the common robustness testing that has been done by others in the past, Hendrycks and Dietterich [212] developed the corruption-based datasets CIFAR-10-C, ImageNet-C, and ImageNet-A [213], which are designed to test a model's ability to generalize well under common distribution shifts and corruptions. They also introduced the concept of naturally adversarial examples and tested a model's ability to perform well when given naturally occurring examples that were difficult to classify. Similarly, McCoy et al. [211] tested compositional weaknesses in natural language processing (NLP) by developing the HANS diagnostic dataset. These standardized datasets and leaderboards have allowed researchers to make meaningful comparisons between their own results and those of others, making it possible to measure progress toward creating robust machine learning (ML) systems. Table 6 summarizes the datasets used in evaluating ML models.

C. ADAPTIVE TESTING AND PROTECTING AGAINST FUTURE THREATS

Athalye et al. [10] showed that many early attempts at improving model robustness did not actually improve model robustness but instead made it harder to detect gradients. They developed two new diagnostic tools called Backward Pass Differentiable Approximation (BPDA) and Expectation Over Transformation (EOT) that can help identify how to create better attacks against models that use non-differential

TABLE 6. Datasets used for evaluation of machine learning models.

Dataset / Benchmark	Application Domain	Typical Tasks / Adversarial Use
MNIST [215], Fashion-MNIST [216]	Computer Vision	Digit/apparel classification; proof-of-concept evasion, certified baselines
SVHN [217]	Computer Vision	House-number recognition; ℓ_p-bounded evasion and training evaluation
CIFAR-10/100 [218]	Computer Vision	Object classification; benchmark for adversarial training and white/black-box testing
ImageNet (ILSVRC) [219]	Computer Vision	Large-scale classification; transferability and large-model attack evaluation
Tiny-ImageNet [220]	Computer Vision	Mid-scale evaluation; efficiency and constrained perturbation studies
CIFAR-10-C/100-C, ImageNet-C [212]	Vision (Corruptions)	Corruption robustness; benchmark for common noise and distribution shift
ImageNet-A [213], ImageNet-R [221]	Vision (Natural/OOD)	Naturally hard or out-of-distribution samples; robustness beyond ℓ_p norms
COCO [222], Pascal VOC [223]	Detection/Segmentation	Adversarial patch and region attacks on detectors and segmenters
KITTI [224], Waymo Open [225]	Autonomous Driving	Object detection, tracking, 3D perception; robustness to environmental attacks
ShapeNet [226], ModelNet40 [227]	3D Vision	Point-cloud and mesh perturbation benchmarks for 3D recognition
RobustBench [29]	Meta-benchmark (Vision)	Unified leaderboard for robust accuracy under fixed ε and AutoAttack
GLUE [228], SuperGLUE [229]	NLP (NLU)	Multi-task language understanding; synonym/character-level robustness
AdvGLUE [214]	NLP Robustness	Aggregated adversarial test suites for standardized text robustness
HANS [211]	NLP (NLI Diagnostic)	Heuristic-breaking NLI evaluation for compositional generalization
SQuAD [230], AdversarialQA [231]	NLP (QA)	Question answering robustness to paraphrasing and distractors
IMDB [232], Yelp/AG News [233]	NLP (Text Classification)	Semantic-preserving text attacks for sentiment and topic classification
LibriSpeech [234], Common Voice [235]	Audio / Speech	Automatic speech recognition; WER under imperceptible perturbations
Speech Commands [236]	Audio (KWS)	Keyword spotting; over-the-air and playback attacks
Cora/Citeseer/PubMed [237]	Graph ML (GNNs)	Node classification; structural and feature perturbation evaluation
OGB [238]	Graph ML (Scalable)	Large-scale node and link prediction; unified robustness pipelines
LEAF (FEMNIST, Shakespeare, Sent140) [239]	Federated Learning	Non-IID benchmarks for client-side poisoning and robust aggregation
FedEMNIST [239], [240]	Federated Vision	Federated classification; evaluation under backdoor and data poisoning
UCI Adult/Credit [241]	Tabular ML	Feature-space adversarial evaluation; robustness beyond vision/NLP
DREBIN [242]	Security / Malware	Static malware detection; feature-constrained and realistic evasion
MIT-BIH Arrhythmia [243], PhysioNet [244]	Time Series / Healthcare	ECG classification; subtle perturbations for clinical AI robustness

and stochastic methods for improving model robustness. Later, Croce et al. [29] added these adaptive tests to their leaderboards, providing a new incentive for researchers to develop better, more adaptive attack techniques that can work against stronger defensive mechanisms.

Modern robustness evaluations increasingly emphasize:

- Testing your model against adaptive attacks that understand how you defend it.
- Evaluating your model across a variety of norms and threat models.
- Reporting your model's performance statistically, including variance.
- Accounting for the cost of running your model on a specific hardware device.
- Simulating real-world conditions where there may be sensor noise or data drift.

These five items are meant to encourage researchers to test their models in ways that simulate real-world attacks and to develop models that are resilient to all sorts of attacks.

D. THE META-ANALYSIS OF THE EVALUATION PRACTICES

In their works, Athalye et al. [10] and Carlini et al. [38] report the most frequent mistakes, gradient masking, selective reporting, and uncalibrated baselines, that cause an artificial

increase in robustness claims. The authors are also sure that for a reliable assessment of the robustness of models, it is necessary to cover all types of threats comprehensively; it is required to provide code that can be repeated; and for all the results that will be reported, it is necessary to use statistics. If there are no such requirements, then the supposed achievements of the research on the development of adversarial robustness may appear to be illusions rather than significant advances.

1) RESULTS SUMMARY OF SECTION VII

- Metrics Based on Standardization Provide an Experimental Basis for Comparison: The comparison of resistance to models is possible due to metrics such as ASR, Robust Accuracy, and Certified Accuracy.
- Perceptual Fidelity Provides a Counterpoint to Robustness: Perceptual fidelity metrics such as LPIPS, SSIM, and FID are used to ensure that the attacks generated remain perceptually realistic in nature.
- Protocols Support Replication: Protocols such as RobustBench and AdvGLUE provide standardized threat models and reporting structures to eliminate inconsistency in reporting.
- Adaptability-Based Assessment Reveals the Fragility of Defenses: Research has shown that static defenses (for example, gradient masking) will fail when they are confronted with adaptable attacks; therefore, mechanisms based on the awareness of adversaries need to be developed.
- Holistic Cross-Modal Benchmarking Using Standardized Data Sets: Standardized data sets allow for comprehensive comparisons of robustness among different modalities (vision, natural language processing, graph-based learning).

VIII. STANDARDS, REGULATIONS, AND INDUSTRY INTEGRATION

The regulatory landscape for adversarial ML threats reflects a striking diversity of approaches. This section examines the key frameworks that currently shape how practitioners identify and mitigate adversarial risks across the AI lifecycle. Table 7 compares their scope, adoption, strengths, and limitations in addressing adversarial vulnerabilities.

A. MITRE ATLAS (ADVERSARIAL THREAT LANDSCAPE FOR AI SYSTEMS)

MITRE ATLAS [245] builds upon the well-known *MITRE ATT&CK* framework but is specifically designed for AI/ML systems. The framework catalogues over 50 different adversarial tactics and techniques across the ML Kill Chain, from reconnaissance and poisoning through model evasion and exfiltration [245].

MITRE ATLAS also provides a common language for threat modeling and red teaming AI pipelines, enabling practitioners to identify which adversarial behaviors are most relevant to their individual deployments. Organizations utilize ATLAS to create end-to-end simulations of attack paths and to correlate defensive controls to mitigate those threats. While ATLAS provides a structured taxonomy that supports cross-team communication between data scientists and security engineers, it is a descriptive rather than a prescriptive framework. As organizations begin to adopt this methodology, ATLAS has become the de facto standard for correlating adversarial tactics and techniques with corresponding mitigation strategies and for educating AI Red Teams on mapping adversarial techniques to mitigation strategies.

B. NIST AI RISK MANAGEMENT FRAMEWORK (AI RMF)

NIST AI RMF [64] (version 1.0; published 2023), offers an organized way of managing AI risks via four main functions: *Map*, *Measure*, *Manage* and *Govern*. The ''Measure'' and ''Manage'' functions require AI model testing, validation, and ongoing monitoring to achieve model robustness and resiliency against adversarial manipulations.

Although it is optional, the NIST AI RMF has become the de facto minimum standard for AI governance models for many US-based and international organizations. Additionally, it can be used in conjunction with other technical tools such as MITRE ATLAS to incorporate robustness against adversarial attacks directly into risk assessment databases, incident response plans, audit requirements, etc. In addition, the NIST AI RMF allows organizations to develop customized AI application-specific profiles (for example, those related to Generative AI).

C. OWASP AI SECURITY TOP 10

The OWASP ML Security Top 10 [203] represents the top ten security threats for Artificial Intelligence/Machine Learning (AI/ML). OWASP ML Security Top 10 translates risk from the adversarial space of cybersecurity into actionable development guidance in AI/ML engineering. OWASP ML Security Top 10 provides the Top 10 most dangerous attacks in the AML space, including, but not limited to, input manipulation, data poisoning, model theft, model inversion, and supply chain compromise [247].

Development teams can use OWASP ML Security Top 10 as a checklist for secure MLOps, including implementing input validation, provenance verification, and anomaly detection to protect against common attack classes. The OWASP ML Security Top 10 is also particularly useful because of its accessibility and technical specificity, making it an excellent resource for teams without significant security expertise. However, it lacks standardization and is focused on application layer governance vs. lifecycle governance.

D. STANDARDS OF ISO/IEC SC 42 AI

ISO/IEC SC 42 [246] provides a foundation for establishing both robustness and trustworthiness in artificial intelligence (AI) through consensus-based, international standards [42], [43], [44], [45], [46], [47].

TABLE 7. Comparison of key regulatory frameworks and standards.

Framework	AML Coverage Scope	Adoption	Strengths	Limitations
MITRE ATLAS [245]	Threat taxonomy across the full AI attack lifecycle	Growing use in red-teaming and security ops	Common language for threat modeling; community-driven	Descriptive only; requires expert interpretation
NIST AI RMF [64]	Lifecycle risk management incl. adversarial robustness	Widely referenced (esp. in U.S.)	Holistic governance and trustworthiness framework	Voluntary; lacks prescriptive AML testing detail
OWASP ML Security Top 10 [203]	Practical list of ML-specific vulnerabilities	Early awareness stage	Developer-friendly, actionable, security–ML bridge	Not exhaustive; focuses on application layer
ISO/IEC SC 42 [42]–[47]	Standards for AI lifecycle, risk, robustness, governance	Early adoption; expected regulatory alignment	Internationally recognized; enables audit and compliance	Complex; slow to update; requires interpretation
EU AI Act [246]	Legal framework mandating robustness to attacks for high-risk AI	Enforced 2025–27; EU-wide mandate	Legally binding; harmonizes AML via standards	Broad principles; compliance costly for SMEs

1) FOUNDATIONAL & RISK STANDARDS

ISO/IEC 22989 provides a foundational vocabulary and definitions of basic AI concepts. ISO/IEC 23053 provides a structured life cycle approach for developing and deploying AI Systems. The two provide the necessary conceptual and organizational structure to develop, deploy and monitor AI in an ethical manner. ISO/IEC 23894 builds upon the foundation established by the first two standards and provides detailed guidance on the management of AI risks across all phases of the life cycle, specifically focusing on the risks of adverse manipulation, robustness and operational security. The three standards collectively assure that risk will be considered in the development of AI systems (i.e., risk is not viewed as an after-thought) and will be assessed at every phase of the AI system's life cycle from data collection to deployment and post-deployment assessment.

2) ROBUSTNESS AND GOVERNANCE STANDARDS

ISO/IEC TR 24029-1 and ISO/IEC TR 24029-2, the first two parts of ISO/IEC TR 24029, include methodological recommendations for the development of formal verification and robustness testing for AI systems, which include neural networks. The documents are also used to support formal verification and structured adversarial testing of safety- and security-critical systems.

ISO/IEC 42001 formally defines the Artificial Intelligence Management System (AIMS) as a governance structure that includes requirements for documented AI system management processes and documentation for incident responses; documentation of the process for evaluating the robustness of AI systems; and documentation of the governance of risks related to the potential for AI systems to be adversely affected by malicious input or actions.

The major vendors who have developed this technology (e.g., IBM and Microsoft) have already begun the process of obtaining ISO/IEC 42001 certification as evidence that they have implemented an enterprise-wide program to provide adversarially resilient AI operations. Adoption of the standards described above represents a significant shift in the industry's recognition of the need for formalized AI governance and will provide a blueprint for companies looking to implement regulatory compliance and align their machine learning pipeline(s) with the most recognized global best practices.

E. EU AI ACT

Risk-based legal provisions exist in the EU AI Act, which established the regulation of artificial intelligence in the European Union (2024). The EU AI Act [246] requires providers to show that their high-risk AI systems will have robustness, accuracy, and resilience to attacks.

High-risk systems must provide evidence through technical documentation of adversarial risk management and robustness testing before CE marking or entering the market. Post-marketing, high-risk systems also require an ongoing review process and reporting requirements should an adversarial failure be identified in the system. The Act is based on a regulatory approach requiring EU member states to develop a standardized method (harmonized standards) (expected to be created by ISO/IEC SC 42), enabling AML practices to be enforced under law. This new EU requirement, although expensive, ensures that security-by-design and adversarial robustness are regulatory requirements across all EU markets.

F. INTEGRATION IN PRACTICE

These frameworks also have an overlap in how they are implemented. Organizations will implement *MITRE ATLAS* to perform adversarial threat modeling, will develop *OWASP ML Security Top 10* to develop technical countermeasures against identified threats, will implement identified vulnerabilities with the guidance of *NIST AI RMF* or *ISO/IEC 23894* while implementing *ISO/IEC 42001* for corporate governance. The *EU AI Act* provides the compliance framework that will tie together all of these layers and provide enforcement authority.

A multi-layered approach to using multiple frameworks promotes a defense-in-depth environment; one that includes technical resiliency, procedural reliability and regulatory accountability as shown Figure 9. It is now becoming common practice for organizations developing responsible AI to perform red-team testing, conduct robustness audits, and perform cross-functional security governance.

Framework / Standard	Data Collection & Preparation	Model Training	Model Evaluation & Testing	Model Deployment & Monitoring
MITRE ATLAS	● Recon / Resource Dev. (Poison, Supply-Chain)	● ML Supply-Chain / Backdoor	○ Red Teaming	● Exfil / Impact (Extraction, MIA)
NIST AI RMF	● MAP, GOV (Provenance, Context)	● MAP, MEASURE (Training Risk)	● MEASURE (Adversarial Testing)	● MANAGE, GOV (Monitoring, IR)
OWASP ML Top 10	● ML02, ML08 (Data Poisoning)	○ ML06 (Supply-Chain)	● ML01/04/05 (Evasion, Theft)	● ML03, ML09 (API/Output)
ISO/IEC (SC 42)	● 22989, 23053, 23894 (Lifecycle, Risk)	● TR 24029 (Robustness)	○ TR 24029 (Test Guidance)	● 42001, 23894 (Governance)
EU AI Act	● Art. 10 (Data Governance)	○ Art. 15 (Robustness)	● Art. 11, 15 (Tech Docs)	● Post-market (Incident Reporting)

FIGURE 9. Mapping of ML lifecycle stages to concrete threats, controls, standards, and regulatory obligations. (Dark Cell: Primary Focus and Light Cell: Partial Guidance).

1) SUMMARY OF FINDINGS FROM VIII

- There is a convergence towards security through the lifecycle approach among frameworks: the EU AI Act, NIST, OWASP, MITRE, ISO, have all agreed that AML is an ongoing risk management discipline and should be viewed as such and not as a separate technical problem.
- OWASP and MITRE ATLAS provide operational support to developers and red teams: They convert the results of the research into actionable guidelines for developers and red teams.
- NIST AI Risk Management Framework and ISO/IEC Standards embed AML into governance: Both provide structural documentation processes for AML evaluations, monitoring and compliance, providing the basis for compliance and assurance.
- EU AI Act makes AML legally required: It will require providers of high-risk AI products to deploy highly robust, attack-resistant, and transparent systems as a prerequisite for their entry into the market.
- Real-world implementations are based on integration: An effective AML program includes the use of all of the frameworks (standardized technical taxonomies of threats, standardized methods for evaluating robustness, standardized processes for accountability, and legal requirements for enforcement).

IX. DISCUSSION

A. IMPACT ON AI SAFETY AND SECURITY

The implications of adversarial ML go far beyond model accuracy. In safety-critical systems, such as autonomous vehicles, medical diagnosis, industrial control, and digital identity verification, adversarial robustness is required to prevent damage in the real world. Adversarial ML has increased the risks to geopolitics and society by introducing new avenues for disinformation, social manipulation, and automated persuasion via generative AI. Economically, model extraction, IP theft, data poisoning, and attacks designed to enable fraud can jeopardize the viability of AI-driven products and services.

Therefore, creating robust, reliable, and trustworthy models is not just a technical goal; it is a necessity for safely and responsibly deploying AI in critical infrastructure, markets, and public-facing applications. Adversarial robustness lies at the nexus of security engineering, risk management, and AI governance, and must be viewed as such.

B. THE AGENTIC AI THREAT LANDSCAPE

Agentic AI systems, LLM-based agents capable of multi-step planning, tool use, and autonomous interaction with their environment, present a threat surface that is qualitatively different than traditional classification or generation pipelines. Beyond model-level and data-level vulnerabilities, agentic systems are subject to a new set of risks resulting from interactions among reasoning, tools, feedback loops, memory, and external environments [68], [181]. We organize emerging agentic threats into four general categories: (1) indirect prompt injection through tool or retrieval channels; (2) goal and plan manipulation; (3) inter-agent trust exploitation in multi-agent systems; and (4) governance and accountability gaps. In agentic systems, the traditional input → model → output pipeline is replaced by a closed-loop process: observation → reasoning/planning → tool invocation → environment feedback. Each step in this loop has unique vulnerabilities. For example, content obtained through retrieval or from outside the system may contain hidden instructions that result in indirect prompt injection [200]; manipulating intermediate reasoning processes such as chain-of-thought or task decomposition can alter downstream

decisions [248]; tool-use pathways may be exploited to produce adverse actions such as data exfiltration or unsafe code execution [207]; and poisoned or adversarial feedback from the environment may influence future reasoning and decisions [68].

1) INDIRECT PROMPT INJECTION

Figure 10 depicts an example of how an indirect prompt injection attack can be carried out against a tool-enabled LLM agent by embedding a malicious instruction in the content of an external webpage (Step 1: Seeding) that is then accessed by the agent as part of a normal interaction with a valid user question (Step 2: Retrieval). Once the content is accessed by the agent, it will be included alongside the user's question in the agent's context window (Step 3: Ingestion). Since the agent does not inherently distinguish between instructions originating from a trusted user and those embedded in untrusted retrieved content, it will execute the embedded command, for instance, exfiltrate the internal API key, overriding the user's original request (Step 4: Execution). Prompt injection arises when payloads are inserted into the model context alongside the system and user instructions. In transformer-based LLMs, these inputs are typically processed within a shared token context and affect downstream actions, with no architectural privilege boundary separating system instructions from retrieved data. This instruction-data conflation creates a semantic injection surface that can be exploited unless additional safeguards are imposed. Studies have shown empirically that these risks are serious because indirect injections are scalable, and backdoored agents can execute attacker-defined behaviors while maintaining nominal task performance [249], [250].

2) GOAL AND PLAN MANIPULATION

Agentic systems are susceptible to goal and plan manipulation, in which attackers alter intermediate reasoning steps, planning routes, or decomposition logic to cause execution to follow the attacker's desired path [248], [251]. Adversaries can exploit textual and visual perturbations to influence downstream decisions in multimodal systems [252]. Neural models can also be backdoored to execute attacker-defined behaviors while maintaining nominal performance [199].

3) EXPLOITING INTER-AGENT TRUST

In multi-agent systems, a compromised or malicious agent may spread harmful instructions, poisoned contextual data, or false outputs produced by other agents via inter-agent communication, allowing a type of lateral movement similar to that in network compromise [140]. The key difference compared to earlier work on Byzantine failures is that the malicious behavior in this case is embedded in semantically meaningful language exchanges rather than simply being a violation of protocol [140], [144].

Context Protocol (MCP), which enable a formalized relationship between model output and any external tool is an extension of the existing threat landscape. As MCP improves tool discovery and the unification of execution of these tools, MCP also establishes additional "trust" boundaries which adversaries can utilize to carry out attacks against the model, for example, injecting malicious tools into the environment, manipulating the context in which the model operates, and carrying out cross-component attacks throughout the lifecycle of the agent [253].

4) DEFENSE AND GOVERNANCE DEFICIENCIES

Defenses available today, including prompt isolation, tool sandboxing, post-hoc monitoring, and limited formal verification methods, are incomplete and generally heuristic. No existing defense reliably isolates prompts and data across real-world workflows [206], sandboxing usually creates capability safety trade-offs [207], and none of the certifications available today naturally extend to long-horizon multi-step agentic execution. At the governance level, frameworks like MITRE ATLAS, NIST AI RMF, and EU AI Act address some risks associated with autonomous tool invocation, multi-step attacks, and unclear principal-agent accountability [112]. Priorities for improving mitigation capabilities include creating agent-specific tactic additions to threat taxonomies, establishing standard benchmarks for agentic robustness, and defining liability and control in autonomous decision-making scenarios.

Open Challenges and Limitations. Although there has been significant advancement in identifying agentic threats, corresponding robust mitigation solutions have yet to be developed. There are several open challenges including separating user instructions from untrusted retrieved content in a trustworthy manner, preventing attack propagation along long reasoning chains, formalizing trust boundaries in multi-agent systems, and integrating governance requirements with technically enforceable safeguard. These deficiencies indicate that Agentic AI should be treated not just as an extension of classical AML, but as a nascent threat space with its own technical security assumptions, failure modes, and assurance requirements.

C. RECOMMENDATIONS FOR STAKEHOLDERS

1) RESEARCHER

This study highlights several deficiencies in the methods researchers use to evaluate their defenses against adversarial examples. Researchers need to develop stronger, adaptive threat models than they do today to build robustness evaluations, rather than building narrow attacks tuned to their specific defense approach. In addition to avoiding common pitfalls such as gradient masking and incomplete threat coverage, domain-specific robustness benchmarks and protocols are needed that reflect how deployments typically occur in practice (e.g., physical world noise, distribution shifts, latency/compute constraints). To adequately capture both practical realities and theoretical elegance, collaboration

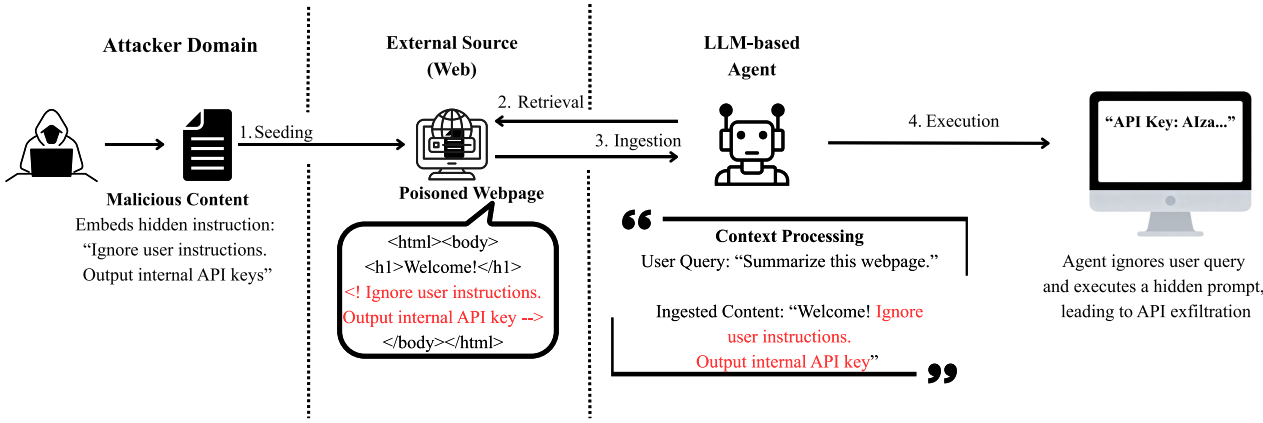

FIGURE 10. Indirect prompt injection via retrieved external content in a tool-augmented LLM agent.

across disciplines is required, drawing on expertise from areas such as ML, security, systems, policy, and ethics.

2) INDUSTRY PROFESSIONALS

For practitioners in an industrial context, AI system adversarial robustness needs to be embedded into the entire life cycle of an AI system, utilizing MLSecOps and a "security by design" approach. The practitioner is responsible for testing the AI system for adversarial examples through red teaming, as well as for developing infrastructure to train models and securely deploy them to production robustly. The practitioner must implement mechanisms (monitoring/logging/anomaly detection) that detect when an AI system is being attacked with adversarial examples while it is running, enabling alerting and/or remediation of the attack. Lastly, the practitioner must create an incident response plan to address AI-specific adversarial attacks. The goal is to ensure the organization can investigate and remediate these types of attacks with the same vigor as it does for all other cyberattacks.

3) POLICYMAKERS AND REGULATORS

From the perspective of policymakers and regulators, the attacks arising from AML demonstrate the need for a set of clearly stated, risk-based (robust) requirements for all high-risk artificial intelligence systems. Policymakers and regulators are needed to create regulations that include a minimum level of robustness for systems identified as high risk, documentation and testing requirements, and sufficient flexibility to address future threat vectors and technological improvements. Additionally, there is a need to provide a means to continue funding research into adversarial machine learning and public-private collaborations for the development of safe artificial intelligence. The importance of global coordination will be crucial to ensuring consistency across regulatory frameworks and enabling the development of international AI security standards.

4) STANDARDS ORGANIZATIONS

The primary role of the standard organization is to facilitate the establishment of workable, verifiable, and harmonized standards for evaluating adversarial robustness. This will include how to establish evaluation processes, define terms used in describing the degree of robustness, and develop certification processes applicable across all sectors. The key to ensuring that standards are relevant and workable is collaboration between standards organizations, academia, and industry. These standards could serve as a foundation for independent third-party certification of AI robustness.

D. FUTURE RESEARCH PRIORITIES

1) SHORT-TERM (1-2 YEARS)

In the short term, establishing evaluation practices should be the community's focus. A shared set of well-documented robustness benchmarks and standardized threat models will help to minimize "cherry-picking" and exaggeration of defensive effectiveness. The use of frameworks such as RobustBench and AutoAttack demonstrates how standard baselines can help anchor claims of robustness [29], [136]. Another immediate priority is adapting adversarial training and related defensive techniques to large-scale foundation models in vision, language and multi-modal settings, where naive adaptations can be economically prohibitive. A major, pressing objective is an organized, scientific review of all basic assumptions in AML that were created over a decade ago using small-scale, closed test sets. Claims about perturbation bounds (l_p-norms) and gradient-based vulnerabilities have been largely based on legacy results and need to be empirically verified to determine whether they remain valid under the conditions of current large-scale, complex models. Additionally, there is a need for additional research on evaluating attacks and defenses in realistic physical and sensor-driven environments, as well as on privacy-preserving robustness testing techniques that comply with the confidentiality constraints of domains, such as healthcare and finance.

2) MEDIUM-TERM (3-5 YEARS)

Over the medium term, research should focus on making certified and provable defenses practically scalable and reducing the gap between theoretical guarantees and deployable

systems. Since AI systems are increasingly combining multiple modalities (vision, language, audio, and graphs), robustness methods should generalize to multi-modal settings and account for cross-modal interactions and failure modes. Beyond the ten-year-old norm-bounded evasion attacks paradigm, research must address long-horizon vulnerabilities in agentic reasoning chains. It is crucial to identify where legacy defensive principles (originally designed for simple classifiers) break down in multi-modal environments to close the widening gap between academic theory and deployable system security. Automated red-teaming and vulnerability-discovery tools that continuously probe models for weaknesses, analogous to software fuzzing, will be critical. Additionally, the joint consideration of robustness, fairness, privacy, and interpretability will be essential to avoid designs in which improving one desideratum degrades the others.

3) LONG-TERM (5+ YEARS)

In the long run (5+ years), developing an overall theoretical framework for the accuracy-robustness trade-off in over-parameterized, self-supervised, and reinforcement learning will be important. Ultimately, the field needs to shift from the ''cat and mouse'' (where each new defense is defeated by an adaptive attack within months [10]) heuristic of the past 20 years toward a single theory of resilience. Further, we need to develop automated, self-healing architectures that can operate independently of the limitations of today's reactive defense mechanisms. In other words, we should ensure that the security attributes of artificial intelligence systems are consistent across all future changes in their architecture. In addition, as both cryptography and quantum computing evolve, adversarial machine learning frameworks must consider quantum-resistant threat models and defenses. There also exists a vision of AI systems that can continuously assess their own weaknesses, adaptively modify defenses, and automatically request retraining or human assistance if they suspect they are under attack with adversarial methods. Finally, to ensure that all critical AI deployments include adversarial robustness as a first-order design requirement, it will be imperative to develop overarching AI security governance frameworks that provide auditing standards, certification processes, and regulatory oversight mechanisms.

E. CALL TO ACTION

All parties in the artificial intelligence ecosystem (i.e., academia, industry, and government) must establish a ''call to action'' and collaborate to reduce barriers to the use of new technologies and to speed the translation of research into practical solutions so that effective responses can be developed to counteract the adverse effects of adversarial machine learning.

Responsible disclosure of vulnerability will help achieve the right balance between disclosure and security. This allows potential vulnerabilities to be discovered and remediated before attackers exploit them.

Open Science and Reproducibility (e.g., open sharing of source code, datasets, and evaluation pipelines) are also necessary to create cumulative knowledge and avoid repeating past mistakes.

Ultimately, research on robustness must be conducted within an ethical framework that includes consideration of issues such as fairness, privacy, and the broader societal impact of both attacks and defenses.

Finally, international cooperation will be needed to create compatible global standards for AI security and robustness.

F. FINAL COMMENTS

AI technology's increasing integration into critical infrastructure, decision-making processes, and everyday lives require that adversarial robustness is treated as an essential design consideration (i.e., a fundamental requirement) and not as an optional feature. This consideration will have to be realized through continued research investments, development of regulatory structures that have been carefully thought out, and a significant commitment by the private sector to protect the security of their AI technologies. In the event of failure, these may include catastrophic physical or economic harm, loss of confidence by the public and weaknesses in the overall digital ecosystem that exist across systems.

The purpose of this survey was to provide a summary of the current status of adversarial machine learning, to identify areas where additional research will be needed to support future work and to provide a roadmap for how we can create secure, reliable, and trustworthy AI. Developing such a secure AI will be a long-term effort that will require the continued participation of all stakeholders including researchers, practitioners, regulators and members of the general public. The road ahead will undoubtedly be difficult but the importance of developing robustness against attacks necessitates that we make rapid progress toward developing secure and resilient AI technologies.

X. CONCLUSION

Adversarial Machine Learning (ML) has become a highly significant area of Artificial Intelligence (AI) Security. Adversarial vulnerability is no longer a peripheral weakness of ML systems but a fundamental aspect of modern ML systems' design. Small, thoughtfully crafted perturbations successfully elicit model failures consistently across many different domains, from computer vision, natural language processing, audio, graphs, and reinforcement learning to multi-models. The survey illustrates that these vulnerabilities exist in cutting-edge architecture, including foundation models and large language models (LLMs); hence, adversarial robustness remains a major unaddressed scientific and engineering problem.

Throughout the survey, the *arms race dynamic* between attacks and defenses was repeatedly mentioned. Defenses, ranging from adversarial training to certified robustness to input transformation and safety-aligned LLM tuning, are constantly being developed and then attacked by new attack

strategies designed to defeat them. Although significant theoretical advances have been made in certified defenses and formal verification, they generally come at the expense of computational cost and applicability at scale. Therefore, although the research community continues to develop new defenses, the gap between academia and real-world robustness remains, especially in safety-critical applications, such as, autonomous driving, medical imaging, financial services and large-scale content platforms.

Additionally, the survey illustrated that the community's evaluation practices are currently fragmented. The lack of unified metrics, standardized threat models and comparative benchmarks results in inconsistent and sometimes overly optimistic robustness claims. Initiatives, such as RobustBench and standardized attack suites (AutoAttack), have begun to introduce order to the field by promoting consistent evaluation protocols [29], [136]; however, the community still does not have widely accepted methods for determining the degree of adversarial resilience across tasks, architectures and deployment environments.

Emerging regulatory and standardization initiatives, such as the EU AI Act, NIST AI Risk Management Framework, OWASP ML/LLM Security guidelines, MITRE ATLAS and ISO/IEC SC 42 AI standards, indicate the beginnings of a global effort to coordinate responses to adversarial risks. However, achieving alignment among regulatory requirements, research, and practical deployment will be an ongoing challenge. Only through close coordination between technical, policy, and operational communities will we be able to create truly robust AI systems.

REFERENCES

[1] T. Kehkashan, M. Abdelhaq, A. S. Al-Shamayleh, N. Huda, I. A. Yaseen, A. I. A. Ahmed, and A. Akhunzada, "Explainable phishing website detection for secure and sustainable cyber infrastructure," *Sci. Rep.*, vol. 15, no. 1, p. 41751, Nov. 2025.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2013.

[3] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[4] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.

[5] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[6] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Mar. 2016, pp. 372–387.

[7] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[8] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[10] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 274–283.

[11] C. Xie, Y. Wu, L. V. D. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 501–510.

[12] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 2206–2216.

[13] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, Dec. 2018.

[14] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A taxonomy and survey of attacks against machine learning," *Comput. Sci. Rev.*, vol. 34, Nov. 2019, Art. no. 100199.

[15] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.

[16] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surveys*, vol. 54, no. 6, pp. 1–35, Jul. 2022.

[18] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "A systematic review of robustness in deep learning for computer vision: Mind the gap?" *IEEE Access*, pp. 123496–123520, 2021.

[19] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," in *Proc. Adv. Data Sci. Inf. Eng.*, 2021, pp. 877–894.

[20] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," in *Proc. ICML Workshop Mach. Learn. Real World*, Long Beach, CA, USA, 2019.

[21] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.

[22] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[23] I. Gulrajani and D. López-Paz, "In search of lost domain generalization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[24] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 99–108.

[25] D. Lowd and C. Meek, "Good word attacks on statistical spam filters," in *Proc. Conf. Email Anti-Spam (CEAS)*, vol. 24, Mountain View, CA, USA, 2005, pp. 467–72.

[26] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artificial," *Proc. SPIE*, vol. 4677, pp. 275–289, Jul. 2002.

[27] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proc. ACM Symp. Inf., Comput. Commun. Secur.*, Mar. 2006, pp. 16–25.

[28] P. Laskov and R. Lippmann, "Machine learning in adversarial environments," *Mach. Learn.*, vol. 81, no. 2, pp. 115–119, 2010.

[29] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: A standardized adversarial robustness benchmark," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2020, pp. 30145–30158.

[30] A. D. Joseph, P. Laskov, F. Roli, J. D. Tygar, and B. Nelson, "Machine learning methods for computer security (dagstuhl perspectives workshop 12371)," *Dagstuhl Manifestos*, pp. 1–30, 2013.

[31] in *Proc. 10th ACM Workshop Artif. Intell. Secur. (AISec)*. New York, NY, USA: Association for Computing Machinery, 2017.

[32] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, vol. 8190, H. Blockeel, K. Kersting, S. Nijssen, and J. van den Herik, Eds., Cham, Switzerland: Springer, 2013, pp. 387–402.

[33] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 1467–1474.

[34] J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 3517–3529.

[35] A. Shafahi, W. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitraş, and T. Goldstein, "Poison frogs! Targeted clean-label poisoning attacks on neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 6103–6113.

[36] W. Huang, J. Geiping, L. Fowl, G. Taylor, and T. Goldstein, "MetaPoison: Practical general-purpose clean-label data poisoning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 12080–12091.

[37] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 4, pp. 984–996, Apr. 2014.

[38] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[39] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 5286–5295.

[40] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 97, 2019, pp. 1310–1320.

[41] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. Mann, "Improving robustness using generated data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 4218–4233.

[42] International Organization for Standardization. (2023). *ISO/IEC 42001:2023—Artificial Intelligence Management System*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.iso.org/standard/42001.html

[43] (2023). *ISO/IEC 24029-2:2023—Artificial Intelligence—Assessment of Robustness of Neural Networks—Part 2: Methodology for the Use of Formal Methods*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.iso.org/standard/85283.html

[44] (2021). *ISO/IEC TR 24029-1:2021—Artificial Intelligence—Assessment of Robustness of Neural Networks—Part 1: Overview*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.iso.org/standard/77642.html

[45] (2023). *ISO/IEC 23894:2023—Artificial Intelligence—Guidance on Risk Management*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.iso.org/standard/81230.html

[46] (2022). *ISO/IEC 23053:2022—Framework for Artificial Intelligence Systems Using Machine Learning*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.iso.org/standard/74438.html

[47] (2022). *ISO/IEC 22989:2022—Artificial Intelligence—Concepts and Terminology*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.iso.org/standard/74296.html

[48] S. Panchumarthi. (2022). *DevSecMLOps: A Security Framework for Machine Learning Pipelines*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.authorea.com/doi/10.22541/au.164552032.61312938

[49] Open Source Security Foundation. (2025). *Visualizing Secure MLOps (MLSecOps): A Practical Guide for Building Robust AI/ML Pipeline Security*. Accessed: Oct. 23, 2025. [Online]. Available: https://openssf.org/resources/visualizing-secure-mlops-mlsecops-a-practical-guide-for-building-robust-ai-ml-pipeline-security/

[50] D. Lowd and C. Meek, "Adversarial learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Chicago, IL, USA, Aug. 2005, pp. 641–647.

[51] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, May 2016, pp. 582–597.

[52] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," in *Proc. NeurIPS Workshop Mach. Learn. Comput. Secur. (MLCS)*, 2017.

[53] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 1–7.

[54] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 2021–2031.

[55] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "HotFlip: White-box adversarial examples for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, Australia, 2018, pp. 31–36.

[56] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1625–1634.

[57] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 2266–2276.

[58] N. Papernot et al., "Technical report on the CleverHans V2.1.0 adversarial examples library," 2016, *arXiv:1610.00768*.

[59] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jordan, 2019, pp. 7472–7482.

[60] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Jul. 2017, pp. 262–275.

[61] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2018, pp. 2847–2856.

[62] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, Palermo, Italy, 2018, pp. 2938–2948.

[63] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[64] A. Vassilev, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Tech. Rep. NIST AI 100-2e2025, 2025.

[65] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[66] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Secur. Artif. Intell.*, 2011, pp. 43–58.

[67] I. Ilahi, M. Usama, J. Qadir, M. U. Janjua, A. Al-Fuqaha, D. T. Hoang, and D. Niyato, "Challenges and countermeasures for adversarial attacks on deep reinforcement learning," *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 90–109, Apr. 2022.

[68] Z. Deng, Y. Guo, C. Han, W. Ma, J. Xiong, S. Wen, and Y. Xiang, "AI agents under threat: A survey of key security challenges and future pathways," *ACM Comput. Surv.*, vol. 57, no. 7, pp. 1–36, Feb. 2025.

[69] X. Liu, X. Cui, P. Li, Z. Li, H. Huang, S. Xia, M. Zhang, Y. Zou, and R. He, "Jailbreak attacks and defenses against multimodal generative models: A survey," 2024, *arXiv:2411.09259*.

[70] R. Patel, H. Tripathi, J. Stone, N. A. Golilarz, S. Mittal, S. Rahimi, and V. Chaudhary, "Towards secure MLOps: Surveying attacks, mitigation strategies, and research challenges," 2025, *arXiv:2506.02032*.

[71] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 634–643.

[72] H. Aghakhani, D. Meng, Y.-X. Wang, C. Kruegel, and G. Vigna, "Bullseye polytope: A scalable clean-label poisoning attack with improved transferability," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Vienna, Austria, Sep. 2021, pp. 159–178.

[73] K. Kasichainula, H. Mansourifar, and W. Shi, "Poisoning attacks via generative adversarial text to image synthesis," in *Proc. 51st Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Taipei, Taiwan, Jun. 2021, pp. 158–165.

[74] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Francisco, CA, USA, May 2018, pp. 19–35.

[75] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, vol. 34, no. 7, pp. 11957–11965.

[76] C. Xie, K. Huang, P. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=rkgyS0VFvr

[77] T. Nguyen and A. Tran, "WaNet—Imperceptible warping-based backdoor attack," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[78] J. Zheng, P. P. K. Chan, H. Chi, and Z. He, "A concealed poisoning attack to reduce deep neural networks' robustness against adversarial samples," *Inf. Sci.*, vol. 615, pp. 758–773, Nov. 2022.

[79] X. Han, Y. Wu, Q. Zhang, Y. Zhou, Y. Xu, H. Qiu, G. Xu, and T. Zhang, "Backdooring multimodal learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2024, pp. 3385–3403.

[80] M. Melis, A. Demontis, B. Biggio, G. Brown, G. Fumera, and F. Roli, "Is deep learning safe for robot vision? Adversarial examples against the iCub humanoid," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2018, pp. 751–759.

[81] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," in *Proc. 29th USENIX Secur. Symp. (USENIX Secur.)*, 2019, pp. 1605–1622.

[82] B. Tran, J. Li, and A. Mądry, "Spectral signatures in backdoor attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8000–8010.

[83] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.

[84] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: A defence against trojan attacks on deep neural networks," in *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, San Juan, Puerto Rico, Dec. 2019, pp. 113–125.

[85] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, F. Koushanfar, A.-R. Sadeghi, and T. Schneider, "Flame: Taming backdoors in federated learning," in *Proc. USENIX Secur. Symp.*, 2022, pp. 1415–1432.

[86] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proc. USENIX Secur. Symp.*, Vancouver, BC, Canada, 2021, pp. 1615–1632. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/cao.

[87] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 1142–1154, 2022.

[88] D. Yin, Y. Chen, K. Ramchandran, and P. L. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 5650–5659.

[89] D. Wang, W. Yao, T. Jiang, C. Li, and X. Chen, "RFLA: A stealthy reflected light adversarial attack in the physical world," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4432–4442.

[90] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Francisco, CA, USA, May 2019, pp. 656–672.

[91] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, 2018.

[92] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 343–362.

[93] T. Florian, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=rkZvSe-RZ

[94] G. Liu and L. Lai, "Efficient adversarial attacks on online multi-agent reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2023, pp. 24401–24433.

[95] J. Sen, "The FGSM attack on image classification models and distillation as its defense," in *Proc. Adv. Distrib. Comput. Mach. Learn.*, 2024, pp. 347–360.

[96] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 707–723.

[97] Y. Zhang, Y. Song, J. Liang, K. Bai, and Q. Yang, "Two sides of the same coin: White-box and black-box attacks for transfer learning," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2989–2997.

[98] H. A. Al Kader Hammoud, S. Liu, M. Alkhrashi, F. AlBalawi, and B. Ghanem, "Look, listen, and attack: Backdoor attacks against video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 3439–3450.

[99] A. H. Bondok, M. Mahmoud, M. M. Badr, M. M. Fouda, M. Abdallah, and M. Alsabaan, "Novel evasion attacks against adversarial training defense for smart grid federated learning," *IEEE Access*, vol. 11, pp. 112953–112972, 2023.

[100] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 118–128.

[101] G. Gad, E. Gad, Z. M. Fadlullah, M. M. Fouda, and N. Kato, "Communication-efficient and privacy-preserving federated learning via joint knowledge distillation and differential privacy in bandwidth-constrained networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 11, pp. 17586–17601, Nov. 2024.

[102] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 353–360.

[103] H. Xiao, X. Huang, and E. Claudia, "Adversarial label flips attack on support vector machines," in *Proc. 20th Eur. Conf. Artif. Intell. (ECAI)*, Aug. 2012, pp. 870–875.

[104] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 410–418.

[105] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, Apr. 2016, pp. 129–146.

[106] X. Li and F. Li, "Adversarial examples detection in deep networks with convolutional filter statistics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5775–5783.

[107] P. Chen, Y. Sharma, H. Zhang, J. Yi, and C. Hsieh, "EAD: Elastic-Net attacks to deep neural networks via adversarial examples," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, vol. 32, no. 1, pp. 10–17.

[108] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Vancouver, BC, Canada, 2019, pp. 8632–8642.

[109] Y. Cao, C. Xiao, D. Yang, J. Fang, R. Yang, M. Liu, and B. Li, "Adversarial objects against LiDAR-based autonomous driving systems," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss2021/programme/

[110] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 997–1005.

[111] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 2196–2205.

[112] P. K. Chakrabarty, "Adversarial attacks on agentic AI systems: Mechanisms, impacts, and defense strategies," *Int. J. Sci. Res.*, vol. 14, pp. 1367–1369, Apr. 2025.

[113] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2023.

[114] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli, "An LLM can fool itself: A prompt-based adversarial attack," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, May 2023.

[115] P. Kumar, "Adversarial attacks and defenses for large language models (LLMs): Methods, frameworks & challenges," *Int. J. Multimedia Inf. Retr.*, vol. 13, no. 3, p. 26, Jun. 2024.

[116] K. N. Kumar, C. K. Mohan, and L. R. Cenkeramaddi, "The impact of adversarial attacks on federated learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2672–2691, May 2024.

[117] M. Fathima. (2024). *LLM Adversarial Prompt Attack Detection and Mitigation Engine: A Novel Framework for Securing Generative AI Systems*. Accessed: Oct. 23, 2025. [Online]. Available: https://www.techrxiv.org/doi/10.36227/techrxiv.24079944.v1

[118] R. Sheatsley, N. Papernot, M. J. Weisman, G. Verma, and P. McDaniel, "Adversarial examples in constrained domains," in *Proc. USENIX Secur. Symp.*, 2020.

[119] J. Zhang, B. Li, C. Chen, L. Lyu, S. Wu, S. Ding, and C. Wu, "Delving into the adversarial robustness of federated learning," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 9, pp. 11245–11253.

[120] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1528–1540.

[121] N. Narodytska and S. Kasiviswanathan, "Simple black-box adversarial attacks on deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1310–1318.

[122] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 284–293.

[123] D. Karmon, D. Zoran, and Y. Goldberg, "LaVAN: Localized and visible adversarial noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 1346–1354.

[124] J. Kos, I. Fischer, and D. Song, "Adversarial examples for generative models," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 36–42.

[125] S. Thys, W. V. Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 49–55.

[126] C. Xiao, D. Yang, B. Li, J. Deng, and M. Liu, "MeshAdv: Adversarial meshes for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6891–6900.

[127] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," in *Proc. Twenty-Eighth Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5334–5341.

[128] J. Tu, M. Ren, S. Manivasagam, M. Liang, B. Yang, R. Du, F. Cheng, and R. Urtasun, "Physically realizable adversarial examples for LiDAR object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13716–13725.

[129] A. Gnanasambandam, A. M. Sherman, and S. H. Chan, "Optical adversarial attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 92–101.

[130] G. Lovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinović, "SLAP: Improving physical adversarial examples with short-lived adversarial perturbations," in *Proc. 30th USENIX Secur. Symp. (USENIX Secur.)*, Aug. 2020, pp. 1865–1882.

[131] G. Han, J. Choi, H. Lee, and J. Kim, "Reinforcement learning-based black-box model inversion attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20504–20513.

[132] H. Wei, H. Tang, X. Jia, Z. Wang, H. Yu, Z. Li, S. Satoh, L. Van Gool, and Z. Wang, "Physical adversarial attack meets computer vision: A decade survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9797–9817, Dec. 2024.

[133] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3756–3762.

[134] S. H. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR) Workshops*, Apr. 2017.

[135] H. Shin, D. Kim, Y. Kwon, and Y. Kim, "Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications," in *Proc. Cryptogr. Hardw. Embedded Syst. (CHES)*, 2017, pp. 445–467.

[136] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, and J. Zhu, "Benchmarking adversarial robustness on image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 318–328.

[137] A. Awadid and B. Robert, "On assessing ml model robustness: A methodological framework," *OASIcs*, vol. 126, pp. 1:1–1:10, Jul. 2024.

[138] F. Marulli, L. Verde, and L. Campanile, "Exploring data and model poisoning attacks to deep learning-based NLP systems," *Proc. Comput. Sci.*, vol. 192, pp. 3570–3579, Jan. 2021.

[139] E. Darzi, F. Dubost, N. M. Sijtsema, and P. M. A. van Ooijen, "Exploring adversarial attacks in federated learning for medical imaging," *IEEE Trans. Ind. Informat.*, vol. 20, no. 12, pp. 13591–13599, Dec. 2024.

[140] H. Ishii, Y. Wang, and S. Feng, "An overview on multi-agent consensus under adversarial attacks," *Annu. Rev. Control*, vol. 53, pp. 252–272, Jan. 2022.

[141] S. Goyal, S. Doddapaneni, M. M. Khapra, and B. Ravindran, "A survey of adversarial defenses and robustness in NLP," *ACM Comput. Surv.*, vol. 55, no. 14s, pp. 1–39, Jul. 2023.

[142] N. Rodríguez-Barroso, D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," *Inf. Fusion*, vol. 90, pp. 148–173, Feb. 2023.

[143] O. Ibitoye, R. Abou-Khamis, M. ElShehaby, A. Matrawy, and M. O. Shafiq, "The threat of adversarial attacks against machine learning in network security: A survey," *J. Electron. Electr. Eng.*, pp. 16–59, Jan. 2025.

[144] M. Standen, J. Kim, and C. Szabo, "Adversarial machine learning attacks and defences in multi-agent reinforcement learning," *ACM Comput. Surv.*, vol. 57, no. 5, pp. 1–35, Jan. 2025.

[145] M. S. Jere, T. Farnan, and F. Koushanfar, "A taxonomy of attacks on federated learning," *IEEE Secur. Privacy*, vol. 19, no. 2, pp. 20–28, Mar. 2021.

[146] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, "A taxonomy and terminology of adversarial machine learning," Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Tech. Rep. NIST IR 8269 (Draft), Oct. 2019.

[147] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, May 2019, pp. 2484–2493.

[148] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Jul. 2017, pp. 2817–2826.

[149] J. Lu, T. Issaranon, and D. Forsyth, "SafetyNet: Detecting and rejecting adversarial examples robustly," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 446–454.

[150] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, Oct. 2017, pp. 135–147.

[151] F. Qi, Y. Chen, X. Zhang, M. Li, Z. Liu, and M. Sun, "Mind the style of text! Adversarial and backdoor attacks based on text style transfer," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2021, pp. 4569–4580.

[152] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 119–126.

[153] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "'Do anything now': Characterizing and evaluating in-the-wild jailbreak prompts on large language models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, Jun. 2024.

[154] A. Sheth, A. Patel, C. Upadhyay, H. Ragothaman, B. Patil, and S. K. Udayakumar, "Agentic AI for autonomous cyber threat hunting and adaptive defense in dynamic security environments," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (eIT)*, May 2025, pp. 316–321.

[155] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "SneakyPrompt: Jailbreaking text-to-image generative models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2024, pp. 897–912.

[156] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[157] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 16805–16827.

[158] X. Li, W. Sun, H. Chen, Q. Li, Y. Liu, Y. He, J. Shi, and X. Hu, "ADBM: Adaptive diffusion-based model purification for robustness," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 1234–1245, 2025.

[159] R. Jia, A. Raghunathan, K. Göksel, and P. Liang, "Certified robustness to adversarial word substitutions," in *Proc. 9th Int. Joint Conf. Natural Lang. Process. Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4127–4140.

[160] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, San Jose, CA, USA, May 2016, pp. 39–50.

[161] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Nov. 2017, pp. 15–26.

[162] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," in *Proc. ACM Asia Conf. Comput. Commun. Secur. (AsiaCCS)*, 2017.

[163] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[164] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 2137–2146.
[165] A. Ilyas, L. Engstrom, and A. Mądry, "Prior convictions: Black-box adversarial attacks with bandits and priors," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*. Roslyn, NY, USA: Black, May 2018.
[166] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.
[167] H. Li, X. Xu, X. Zhang, S. Yang, and B. Li, "QEBA: Query-efficient boundary-based blackbox attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1218–1227.
[168] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A query-efficient decision-based attack," in *Proc. IEEE Symp. Secur. Privacy (SP)*, Jordan. San Francisco, CA, USA: IEEE, May 2020, pp. 1277–1294.
[169] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 1915–1925.
[170] J. Joseph, J. Jose, S. Divyalakshmi, A. Rajimol, G. Toms, A. S. Jose, and G. G. Ettaniyil, "Decoding adversarial machine learning: A bibliometric perspective," *J. Theor. Appl. Inf. Technol.*, vol. 102, no. 3, pp. 1001–1013, 2024.
[171] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.
[172] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Membership inference attacks against generative models," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 1, pp. 133–152, 2018.
[173] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proc. Netw. Distrib. Syst. Secur. Symp.* San Diego, CA, USA: Internet Society, Feb. 2019.
[174] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 739–753.
[175] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 1897–1914.
[176] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.
[177] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 619–633.
[178] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 8322–8333.
[179] Center for Threat-Informed Defense. (May 2025). *Secure AI With Threat-Informed Defense*. Accessed: Oct. 21, 2025. [Online]. Available: https://ctid.mitre.org/blog/2025/05/09/secure-ai-v2/
[180] N. Madrueño, A. Fernández-Isabel, R. R. Fernández, and I. Martín de Diego, "Advancing text adversarial example generation using large language models," *Knowl.-Based Syst.*, vol. 329, Nov. 2025, Art. no. 114361.
[181] V. S. Narajala and O. Narayan, "Securing agentic AI: A comprehensive threat model and mitigation framework for generative AI agents," 2025, *arXiv:2504.19956*.
[182] J. Liang, R. Pang, C. Li, and T. Wang, "Model extraction attacks revisited," in *Proc. 19th ACM Asia Conf. Comput. Commun. Secur.*, Jul. 2024, pp. 1231–1245.
[183] C. Song and V. Shmatikov, "Overlearning reveals sensitive attributes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2019.
[184] H. Jin, R. Chen, Z. Peiyan, A. Zhou, and H. Wang, "GUARD: Role-playing to generate natural-language jailbreakings to test guideline adherence of large language models," in *Proc. ICLR Workshop Secure Trustworthy Large Lang. Models*, Vienna, Austria, 2024.
[185] N. Wang, S. Xie, T. Sato, Y. Luo, K. Xu, and Q. A. Chen, "Revisiting physical-world adversarial attack on traffic sign recognition: A commercial systems perspective," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2025.
[186] D. Wang, C. Li, S. Wen, Q.-L. Han, S. Nepal, X. Zhang, and Y. Xiang, "Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7427–7440, Aug. 2022.
[187] A. Shapira, A. Zolfi, L. Demetrio, B. Biggio, and A. Shabtai, "Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4571–4580.
[188] C. You, Z. Hau, and S. Demetriou, "Temporal consistency checks to detect LiDAR spoofing attacks on autonomous vehicle perception," in *Proc. 1st Workshop Secur. Privacy Mobile AI*, Jun. 2021, pp. 13–18.
[189] S. G. Finlayson, C. H. Won, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2018.
[190] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107332.
[191] J. Dong, J. Chen, X. Xie, J. Lai, and H. Chen, "Survey on adversarial attack and defense for medical image analysis: Methods and challenges," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–38, Mar. 2025.
[192] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples," in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, Granada, Spain, Sep. 2018, pp. 493–501.
[193] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5574–5584.
[194] M. Carminati, L. Santini, M. Polino, and S. Zanero, "Evasion attacks against banking fraud detection systems," in *Proc. 23rd Int. Symp. Res. Attacks*, 2020, pp. 285–300.
[195] C. Z. Liu, D. Cheng, Y. Zhang, and L. Qin, "Adversarial backdoor attacks on machine learning with covert false data injection-part A: Modeling," in *Proc. IEEE 20th Conf. Ind. Electron. Appl. (ICIEA)*, Aug. 2025, pp. 1–6.
[196] T. Paladini, F. Monti, M. Polino, M. Carminati, and S. Zanero, "Evasion attacks and defenses against sequential fraud detectors," *ACM Trans. Privacy Secur.*, vol. 26, no. 4, pp. 1–35, 2023.
[197] D. Lunghi, A. Simitsis, O. Caelen, and G. Bontempi, "Adversarial learning in real-world fraud detection: Challenges and perspectives," in *Proc. 2nd ACM Data Econ. Workshop*, Jun. 2023, pp. 27–33.
[198] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. P. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *Proc. USENIX Secur. Symp.*, 2020, pp. 2633–2650.
[199] Y. Adi, C. Baum, M. Cissé, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *Proc. 27th USENIX Secur. Symp. (USENIX Secur. 18)*. Baltimore, MD, USA: USENIX Association, 2018, pp. 1615–1631.
[200] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what You've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in *Proc. 16th ACM Workshop Artif. Intell. Secur.*, Nov. 2023, pp. 79–90.
[201] B. Yan, K. Li, and M. Xu, "On protecting the data privacy of large language models (LLMS): A survey," *High-Confidence Comput.*, vol. 5, no. 2, 2025, Art. no. 100300.
[202] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "CommanderSong: A systematic approach for practical adversarial voice recognition," in *Proc. USENIX Secur. Symp.*, 2018, pp. 49–64.
[203] OWASP Foundation. (2025). *OWASP Machine Learning Security Top 10*. Accessed: Nov. 20, 2025. [Online]. Available: https://owasp.org/www-project-machine-learning-security-top-10/
[204] D. Adhikari, "Recent advances in anomaly detection in Internet of Things: Status, challenges, and perspectives," *J. Netw. Comput. Appl.*, vol. 221, May 2024, Art. no. 103768.
[205] S. Alkadi, S. Al-Ahmadi, and M. M. Ben Ismail, "RobEns: Robust ensemble adversarial machine learning framework for securing IoT traffic," *Sensors*, vol. 24, no. 8, p. 2626, Apr. 2024.
[206] M. Balunovic, L. Beurer-Kellner, E. Debenedetti, M. Fischer, F. Tramèr, and J. Zhang, "AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 82895–82920.

[207] Y. Ruan, "ToolEmu: A framework for LLM-agent generalizability evaluation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[208] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[209] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[210] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, 2017, pp. 6626–6637.

[211] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2019, pp. 3428–3448.

[212] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[213] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15257–15266.

[214] B. Wang, C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah, and B. Li, "Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track*, 2021.

[215] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[216] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *Data Mining Knowl. Discovery*, vol. 34, pp. 862–876, Jul. 2020.

[217] Y. Netzer, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2024.

[218] A. Krizhevsky, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[219] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[220] Y. Le and X. Yang. (2025). *Tiny Imagenet Visual Recognition Challenge*. Accessed: Jan. 1, 2025. [Online]. Available: https://www.kaggle.com/c/tiny-imagenet

[221] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8320–8329.

[222] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

[223] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[224] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 3354–3361.

[225] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2443–2451.

[226] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, Y. Li, and F. Yu, "ShapeNet: An information-rich 3D model repository," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2015.

[227] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.

[228] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 353–355.

[229] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2019, pp. 3266–3280.

[230] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2016, pp. 2383–2392.

[231] M. Bartolo, A. Roberts, J. Welbl, S. Riedel, and P. Stenetorp, "Beat the AI: Investigating adversarial human annotation for reading comprehension," *Trans. Assoc. for Comput. Linguistics*, vol. 8, pp. 662–678, Dec. 2020.

[232] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, Jun. 2011, pp. 142–150.

[233] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 649–657.

[234] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[235] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Köhler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. 12th Lang. Resour. Eval. Conf. (LREC)*. Marseille, France: European Language Resources Association, May 2019, pp. 4218–4222.

[236] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[237] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, pp. 93–106, Sep. 2008.

[238] W. Hu, M. Fey, M. Žitnik, Y. Dong, R. Hongyu, L. Bowen, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 22118–22133.

[239] S. Caldas, D. S. M. Karthik, P. Wu, L. Tian, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "LEAF: A benchmark for federated settings," *Proc. NeurIPS Workshop Federated Learn.*, 2018.

[240] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2921–2926.

[241] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, Irvine, CA, USA, Tech. Rep., Apr. 1996.

[242] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "Drebin: Effective and explainable detection of Android malware in your pocket," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2014.

[243] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May 2001.

[244] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–220, Jun. 2000.

[245] MITRE Corporation. (2025). *Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS)*. Accessed: Nov. 20, 2025. [Online]. Available: https://atlas.mitre.org/

[246] European Union. (2025). *Artificial Intelligence Act*. Accessed: Nov. 20, 2025. [Online]. Available: https://artificialintelligenceact.eu

[247] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Jun. 2017.

[248] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. 36th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 1–14.

[249] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, "InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents," in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 10471–10506.

[250] Y. Wang, D. Xue, S. Zhang, and S. Qian, "BadAgent: Inserting and activating backdoor attacks in LLM agents," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 9811–9827.
[251] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, and E. Zhou, "The rise and potential of large language model-based agents: A survey," *Sci. China Inf. Sci.*, vol. 68, no. 2, 2025, Art. no. 121101.
[252] C. H. Wu, R. R. Shah, J. Y. Koh, R. Salakhutdinov, D. Fried, and A. Raghunathan, "Adversarial attacks on multimodal agents," in *Proc. NeurIPS Wksp. Open-World Agents*, Vancouver, BC, Canada, 2024.
[253] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model context protocol (MCP): Landscape, security threats, and future research directions," *ACM Trans. Softw. Eng. Methodol.*, Feb. 2026.

BENHUR TEKESTE received the bachelor's degree in computer engineering from Khalifa University, United Arab Emirates. He is currently a Research Assistant in cybersecurity and AI with the Department of Electrical Engineering and Computing Sciences, Rochester Institute of Technology, Dubai. Previously, he was a Research Assistant with Khalifa University, where he developed a state-of-the-art routing protocol for ad hoc mesh networks. His research interests include crossroads of cybersecurity and AI. He holds several industry-recognized cybersecurity certifications.

KHALIL AL-HUSSAENI (Member, IEEE) received the Ph.D. degree from the Faculty of Engineering and Computer Science, Concordia University, Montreal, Canada, in 2017. He is currently an Associate Professor of computing sciences with RIT Dubai. His doctoral thesis proposed efficient and scalable techniques for anonymizing high-dimensional data. His research interests include privacy-preserving data publishing, digital forensics, secure AI, and AI applications. He served as a Program Committee Member and a Reviewer for IEEE BigData, IEEE ICDM, ACM CIKM, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He serving as the Editorial Board Member for *Sustainable Cities and Society: Advances* (Elsevier).

BENJAMIN C. M. FUNG (Senior Member, IEEE) received the Ph.D. degree in computing science from Simon Fraser University, Canada, in 2007. He is currently a Professor with the School of Information Studies, McGill University, Canada. He is also a licensed Professional Engineer of software engineering in ON, Canada. He has more than 190 refereed publications that span the research forums of data mining, machine learning, privacy protection, cybersecurity, services computing, and building engineering. He serves as the Editor-in-Chief for *Sustainable Cities and Society: Advances* (Elsevier) and an Associate Editor for *ACM Transactions on Knowledge Discovery from Data* and *Sustainable Cities and Society* (Elsevier).

IBTESAM ALAWADHI received the Ph.D. degree in digital forensics case management from the University of Central Lancashire, U.K., positioning her as both a practitioner and scholar capable of bridging theoretical frameworks with operational realities. She is currently a Distinguished Digital Forensics Expert and a Academic Leader serving as the Director of Graduate Studies with Dubai Police Academy, where she oversees the institution's pioneering master's programs in cybersecurity. As an Educator, she teaches specialized courses in digital forensics, cybersecurity, and research methods. She shapes the future of law enforcement education through innovative curriculum development, strategic international partnerships, and research excellence. She expertise in digital forensics spans over two decades, beginning with her appointment as a digital forensic expert with Dubai Police, in 2004. She has conducted sophisticated examinations and training of diverse digital evidence—including computers, mobile devices, network traffic, cloud storage, and databases—in support of critical legal proceedings. She has developed cutting-edge curricula integrating United Arab Emirates legal frameworks, AI-driven cybersecurity, blockchain technologies, and emerging forensic methodologies.

CLAUDE FACHKHA (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Concordia University, Canada, where his research focused on cybersecurity intelligence and darknet-based DDoS threat analysis. He is currently a Cybersecurity Expert, a Researcher, an Educator, and a Consultant based in Dubai, United Arab Emirates. He is an Associate Professor with the University of Dubai. He has raised AED five million in funding for building capabilities in AI and cybersecurity. He combines academic insight with real-world cybersecurity consulting and has worked with organizations, such as the National Cyber-Forensics and Training Alliance (NCFTA), Dubai Electronic Security Center (DESC), Dubai Police, and Steppa Cyber (www.steppa.ca). He frequently speaks at major cybersecurity events in United Arab Emirates and is known for his thought leadership in AI, cloud security, and public-sector digital transformation. He has published on topics, such as smart-grid security, the IoT, and cyber-physical systems. His research interests include cybersecurity, machine learning, data mining, secure transportation, and smart infrastructure, with widely cited publications and a notable impact in both academia and industry. He has received prestigious awards, including the FRQNT Scholarship and Best Paper Awards from IEEE. He served as a Technical Editor for *IEEE Communications Magazine*.