IJICTR
International Journal of Information & Communication Technology Research

Volume 17- Number 2 – 2025 (9-17)

Defending Against Adversarial Attacks in Artificial Intelligence Technologies

Seyed Hadi Sajjadi*
ICT Research Institute
Tehran, Iran
h.sadjadi@itrc.ac.ir

Sasan Karamizadeh
ICT Research Institute
Tehran, Iran
s.karamizadeh@itrc.ac.ir

Amirmansour Yadegari
ICT Research Institute
Tehran, Iran
amyadegari@itrc.ac.ir

Received: 13 February 2024 – Revised: 8 March 2024 - Accepted: 31 April 2024

Abstract—**The rapid adoption of artificial intelligence (AI) technologies across diverse sectors has exposed vulnerabilities, particularly to adversarial attacks designed to deceive AI models by manipulating input data. This paper comprehensively reviews adversarial attacks, categorising them into training-phase and testing-phase types, with testing-phase attacks further divided into white-box and black-box categories. We explore defence mechanisms such as data modification, model enhancement, and auxiliary tools, focusing on the critical need for robust AI security in healthcare and autonomous systems sectors. Additionally, the paper highlights the role of AI in cybersecurity, offering a taxonomy for AI applications in threat detection, vulnerability assessment, and incident response. By analysing current defence strategies and outlining potential research directions, this paper aims to enhance the resilience of AI systems against adversarial threats, thereby strengthening AI's deployment in sensitive applications.**

Keywords: Adversarial Attacks, AI Security, Defense Mechanisms, Machine Learning, Cybersecurity

Article type: Research Article

I. INTRODUCTION

Artificial Intelligence (AI) technologies are increasingly used in various fields, such as image classification, object detection, voice control, and machine translation. Additionally, they are employed in more advanced areas like drug compound analysis, brain circuit reconstruction, particle accelerator data analysis, and the analysis of the impact of DNA mutations [1]. Defensive strategies against adversarial attacks include adversarial training, certified defences, feature transformations, and ensemble methods, each with varying effectiveness and limitations in enhancing model robustness [2]. Classification ML/DL techniques are employed to defend against adversarial attacks on artificial neural networks, improving their robustness and security in critical applications [3].

Research into adversarial AI technologies has gained popularity as neural networks are vulnerable to attacks. Attacks can be categorised into training-phase attacks, which alter the training dataset or data labels, and testing-phase attacks, which can be divided into white-box and black-box attacks. In white-box attacks, adversaries access the target model's parameters, algorithms, and structure[4]. In contrast, adversaries

Cannot obtain information about the target model in black-box attacks, but can train a substitute model by querying the target model.

* Corresponding Author

Types of Adversarial Attacks Adversarial attacks can be divided into two categories: training-phase and testing-phase attacks. Each will be explained below.

1- Training-phase adversarial attacks involve modifying training datasets, input features, and data labels. ·

- Modifying the training dataset: Changing the original distribution of the training dataset by modifying or removing training data ·
- Manipulating labels: Reducing model performance by randomly flipping a certain percentage of training labels ·
- Manipulating input features: Injecting carefully crafted malicious data into the training dataset to change the decision boundary

2- Testing-phase adversarial attacks include white-box and black-box attacks

- White-box attacks: Adversaries have access to the target model's parameters, algorithms, and structure.
- Black-box attacks: Adversaries cannot obtain information about the target model but can train a substitute model by querying it.

This paper comprehensively reviews adversarial attacks within artificial intelligence technologies, categorising them into two primary types: training-phase and testing-phase attacks. The latter is divided into white-box and black-box categories based on the adversary's access to the model's internal details. We explore defence mechanisms, including data modification, model enhancement, and auxiliary tools, highlighting the need for robust AI security measures in critical sectors such as healthcare and autonomous systems. Our classification builds upon existing frameworks but focuses on adversarial attacks that target AI models, specifically within cybersecurity applications. Unlike previous works, our approach emphasizes the growing complexity of attacks in modern AI systems. It categorises the primary defence strategies into data alteration, model modification, and the application of supplementary technologies. Moreover, we propose a taxonomy that organises the various AI applications in cybersecurity, such as threat detection, vulnerability assessment, and incident response, based on the nature of adversarial threats. By classifying adversarial attacks and defence mechanisms, we aim to provide a clearer understanding of the vulnerabilities in AI systems and suggest future research directions to enhance the resilience of AI technologies against adversarial threats.

II. Literature Review

Adversarial attacks have been extensively explored in computer vision, with numerous methods and techniques developed primarily for image recognition tasks. Researchers have highlighted the serious public safety risks posed by adversarial attacks, such as in self-driving cars, where a stop sign could be misclassified, potentially leading to fatal accidents. Similarly, adversarial attacks significantly affect network security, particularly in areas like intrusion detection and malware detection, where machine learning has rapidly been adopted [5].

Despite the growing research in adversarial machine learning within network security, a comprehensive survey has yet to cover the expanding body of work in this field. Existing surveys include Akhtar et al. [6], which reviewed adversarial attacks on deep learning in computer vision [7] and provided a broad overview of adversarial attacks in artificial intelligence with brief discussions on cloud security, malware detection, and intrusion detection. The paper [8] focuses on security threats and defensive techniques in machine learning, particularly regarding vulnerabilities in learning algorithms. Rosenberg et al. [9] offered a general review of adversarial attacks in cybersecurity domains such as intrusion detection systems, URL detection systems, biometric systems, CPSS (Cyber-Physical Systems), and industrial control systems. However, unlike these works, our review focuses exclusively on network security and employs a distinct approach to classify adversarial attacks and defense mechanisms [10].

In [11], adversarial machine learning is discussed in the context of cyberwarfare, including attacks on malware classifiers. Zhang et al. [12] addressed the limitations of deep learning in mobile and wireless networks but did not focus on network security. Buczak et al. [13] surveyed machine learning-based cybersecurity intrusion detection but did not examine adversarial attacks. A paper [14] provided a historical overview of adversarial machine learning in computer vision and cybersecurity, though it did not delve deeply into network security. In [15], the security of machine learning in malware detection, particularly in Call and Control (C&C) detection techniques, focuses on identifying weaknesses and limitations in secure machine learning algorithms.

DNNGuard offers a hybrid architecture that efficiently executes DNN networks alongside detection algorithms, enhancing robustness against adversarial attacks through dynamic resource scheduling and support for sparse and dense workloads [16]. TXAI-ADV employs deep learning and machine learning classifiers to detect and defend against adversarial attacks in Consumer Internet of Things devices [17]. A novel four-component methodology enhances model efficiency and adversarial resistance through optimized parameter tuning, weight compression, and a multi-expert architecture, improving performance across various attack scenarios [18]. Latent adversarial training (LAT) enhances robustness against unforeseen adversarial attacks by utilising structured latent representations, improving performance on clean data compared to traditional adversarial training methods [19].

A. Types of Attacks and Threats in Machine Learning

This section examines the threats and attacks that machine learning systems face.

- **Poisoning the Training Dataset**

  The malicious manipulation of a training dataset to mislead the prediction of a machine-learning model is called a poisoning attack. The goal of

poisoning attacks is to misclassify and lose samples or test data by using poisoned samples, also known as adversarial examples, in the training datasets.

- **Backdoor in the Training Dataset**

  Recently, researchers have shown that attackers can create a hidden backdoor in the training data or a pre-trained model. A backdoor attack, also known as a Trojan attack, is an adversarial attack where a malicious attacker injects a backdoor into models during the training phase. As a result, the backdoored model operates normally on clean samples but can be activated by a backdoor pattern to misclassify backdoor samples as the target label specified by the backdoor attacker.

- **Random Threats**

  This type of threat is generated randomly or through legitimate components. Human errors represent a common random threat. They typically occur during the configuration or operation of devices, information systems, or the execution of processes.

- **Adversarial Example Attacks**

  In this attack, the attacker carefully crafted the disruption in input data to cause the machine-learning model to make mistakes. Adversarial example attacks can be divided into two categories: 1—general error attacks, which cause the model to make mistakes, and 2—specific error attacks, where the goal is for the model to misidentify the adversarial example as coming from a particular class.

- **Model Stealing Attacks**

  In this type of attack, an adversary can steal a machine-learning model by observing the output labels and confidence levels for chosen inputs. This attack, also known as model extraction or model stealing attack, has become an emerging threat.

- **Environmental Threats**

  Environmental threats include natural disasters (floods, earthquakes), human-made disasters (fires, explosions), and the failure of supporting infrastructure (power outages, communication breakdowns).

- **Recovery of Sensitive Training Data**

  In addition to the model extraction or stealing attacks described above, other related attacks in machine learning include: 1—membership inference attack, where the attacker attempts to determine whether a specific sample's data was used during model training, and 2—inversion attack, where the attacker infers information about the training data.

- **Hostile Threats**

  This type of threat is created with malicious intent (such as denial of service attacks, unauthorised access, and identity hiding) for individuals, groups, organisations, or nations.

- **Vulnerability**

  This is an existing weakness that an attacker may exploit.

B. *Type of robustness of neural networks against adversarial attacks*

To enhance the robustness of neural networks against adversarial attacks, researchers have proposed various defence methods, which are detailed below:

- **Data Modification Methods**

  Data modification methods include adversarial training, gradient hiding, block transferability, compression, and randomisation.

- **Gradient Hiding** conceals the target model's gradient information from adversaries. However, this approach can be easily circumvented by training a proxy black-box model with gradients and using adversarial examples generated by this model.

- **Blocking Transferability**

  Blocking Transferability prevents the transferability of adversarial examples by labelling adversarial inputs as blank instead of classifying them as their original labels.

- **Data Compression**

  Data compression improves robustness by compressing the data. However, excessive compression can reduce the accuracy of classifying the original image.

- **Data Randomisation**

  removes potential adversarial perturbations in the image and applies data augmentation operations during the training process to slightly improve the robustness of the target model.

C. *Models for Defending Against Adversarial Attacks*

In machine learning, ensuring the reliability of models against adversarial attacks is a critical concern. Adversarial attacks involve subtle modifications to input data to deceive the model into making incorrect predictions, which can be particularly problematic in sensitive applications like security and healthcare [20]. To mitigate these risks, several defensive mechanisms have been developed. Here are some key models and technologies for defending against adversarial attacks:

• Regularisation:

Incorporates regularisation techniques to enhance the generalisation ability of the target model and limit vulnerabilities.

• Defensive Distillation:

Produces a model with smoother output levels and less sensitivity to perturbations, improving model robustness and reducing the success rate of adversarial attacks by up to 90 percent.

• Feature Squeezing:

This technique aims to reduce the complexity of data representation and minimise adversarial interference, due to its lower sensitivity.

• Deep Contractive Network:

Utilises a denoising autoencoder to mitigate adversarial noise.

• Mask Layer:

This layer encodes the difference between the original images and the output features of previous network layers to enhance robustness against adversarial inputs.

D. Key Adversarial Features

Adversarial examples possess three critical features, which are explained below [21]:

• Transferability:

This allows attacks on one model to be effective on another with similar training.

• Adversarial Instability:

Physical transformations can cause adversarial examples to lose their effectiveness.

• Use of Adversarial Training:

This regularization method can reveal model defects and improve robustness, but it is costly.

Adversaries have various capabilities and objectives, which are classified based on their access to the training and testing phases. In the training phase, adversaries may inject, alter, or corrupt data and learning algorithms [22]. In the testing phase, adversaries can enforce incorrect outputs through white-box or black-box attacks, with varying knowledge about the target model.

D. *Adversarial Capabilities and Objectives*

The security of machine learning models is evaluated based on the capabilities and objectives of adversaries. This section explores the potential actions and intentions of adversaries [23]. Adversarial capabilities refer to the extent of information an adversary can access and utilise regarding the target model. These capabilities are generally categorised into two types:

1- **Training Phase Capabilities**

In the training phase, adversaries may try to directly manipulate or compromise the target model by modifying the dataset used for training.

2- **Testing Phase Capabilities**

In the testing phase, adversarial attacks do not manipulate the target model during the testing phase but instead force it to produce incorrect outputs.

III. Applications of AI, Machine Learning, and Large Language Models in Cybersecurity

AI and machine learning identify potential threats by examining digital footprints like network traffic, system activity logs, and user actions. These technologies can recognize unusual patterns or anomalies that might signal a malicious attack. This proactive approach allows organisations to anticipate and counteract threats before they cause significant damage [24].

A. *Vulnerability Assessment and Penetration Testing*

Machine learning helps uncover weaknesses in IT systems by analysing software code, system configurations, and network setups. Attackers can exploit these vulnerabilities. Additionally, AI-powered tools simulate cyberattacks to assess a system's defences and identify areas for improvement.

B. *Forensics and Incident Response*

When a cyberattack occurs, AI can expedite the investigation and recovery process. AI reduces the attack's impact by analysing security alerts and automating response actions. Furthermore, AI can examine various data sources to uncover crucial details about the incident, aiding recovery and preventing future attacks.

C. *Malware Analysis and Classification*

Machine learning is essential for understanding and combating malware. By examining numerous malware samples, AI can categorise different types of malware and identify new threats. This knowledge enables organisations to develop effective countermeasures to protect their systems.

D. *Fake News Detection*

AI in fake news detection leverages machine learning technologies to analyse various data sources, including text, media content, social context, and network structure, to identify potentially false or misleading information. AI algorithms can effectively flag potential fake news by recognising patterns and anomalies in these data sources [25].

E. *Fraud Detection*

AI plays a vital role in fraud detection by analyzing large volumes of data, identifying patterns and anomalies, and flagging suspicious activities in real time. It can detect unusual patterns, identify fraudulent transactions, and prevent identity theft [26].

IV. Explanation of Need

A. *Securing AI Technologies*

AI can potentially revolutionise society by addressing pressing issues, but its cyberattack vulnerabilities constitute a significant concern. Despite advancements, many AI systems can be manipulated by sophisticated hacking techniques. To protect AI systems, researchers are developing countermeasures. However, we need a comprehensive understanding of these attacks, including the attackers' goals and methods. This knowledge is crucial for creating effective defences, especially in critical areas like industry and healthcare. Overcoming these challenges

requires focused research on AI security and privacy [27].

B. Address Security Risks in AI Deployment

When a multitude of adverse events occurs in a field, public outcry arises, gradually revealing issues and problems [28, 29]. This often leads to a wave of regulatory demands. The emergence of such conditions marks the beginning, and it takes years for a regulatory body to be established and begin regulating the industry. This process takes a considerable amount of time.

Regulating AI in terms of its application in critical industries and sectors has multiple dimensions. A review of strategic documents from various countries shows that regulation is a crucial component of all such documents. Table 1 illustrates the regulatory focuses highlighted by different countries in their strategic documents:

TABLE I. REGULATION IN THE OPINION OF DIFFERENT COUNTRIES

No.	Country	Key Regulatory Features
1	Germany	Algorithmic transparency, explainability, data protection, predictability, transparency, validation to prevent abuse and discrimination, harmonizing copyright laws for ML purposes
2	China	Harmony and friendship, fairness and justice, privacy, security, controllability, emphasis on IP rights and ethical values
3	France	AI ethics, accountability of AI developers for harmful applications, and AI system transparency
4	Lithuania	Human-centred AI, ethical goals, trustworthiness, transparency, privacy, security
5	Mexico	Privacy protection, personal data protection, flexible IP system
6	Qatar	Ethical framework, accountability in sensitive sectors like health and judiciary, cybersecurity protection, and privacy
7	Sweden	Ethical considerations, safety, security, trust-building, transparency, emphasis on human and social dimensions, privacy protection
8	Turkey	Data and algorithm regulation in sensitive fields like national security and health, strict AI ethics policies, protection of AI-related IP, and privacy protection
9	UAE	Ensuring data privacy, ethical AI use, data integrity and security, high emphasis on citizen protection, and privacy
10	USA	Designing trustworthy and secure AI systems, developing AI aligned with ethical, legal, and social goals, enhancing fairness, transparency, and accountability, IP rights, ensuring stability, fairness, and security

Table I shows that security and privacy are essential components of AI regulation across all countries. Therefore, thorough studies on the security dimensions of AI deployment are necessary to develop regulatory strategies and policies for AI usage. Regulation, awareness, supervision, and transparent development programs can only be effectively planned with an understanding of these dimensions.

Rapid developments in AI and the increasing adoption of AI in areas such as autonomous vehicles, lethal weapon systems, robotics, and similar domains are significant. Governments face challenges in managing the scale and pace of social and technical transitions. While substantial literature on various aspects of AI is emerging, AI governance must be more developed [30]. New AI applications offer opportunities to enhance economic efficiency and quality of life, but also introduce unintended and undesirable consequences, creating new forms of risk that must be addressed [31].

To maximize AI's benefits while minimizing unwanted risks, governments worldwide must better understand the scope and depth of these risks and develop regulatory and governance processes and structures to address these challenges. We face multifaceted challenges in AI governance, including emerging governance approaches to AI, policy capacity-building, legal and regulatory challenges of AI and robotics, and unresolved issues and gaps that require attention.

V. HIGH-LEVEL ARCHITECTURE OF AI SECURITY TECHNOLOGIES

In this paper, we have proposed several strategies to defend against adversarial attacks on AI systems, categorized into three main groups: data alteration, model modification, and the use of auxiliary tools, as shown in Fig.1

A. Data Alteration

1) Adversarial Training

In adversarial training, models are trained on modified data samples designed to deceive them, enhancing their robustness against adversarial attacks. For instance, a neural network trained to distinguish between cats and dogs can be exposed to slightly altered images of cats misclassified as dogs. By incorporating these adversarial examples into the training set, the model learns to correctly classify these deceptive inputs, improving its resilience against similar future attacks.

2) **Obfuscation**

Obfuscation involves hiding or concealing sensitive information within the data to prevent attackers from extracting meaningful insights. For example, sensitive information can be encrypted using cryptographic algorithms in textual data. Encrypting personal identification numbers (PINs) in a database makes it easier for attackers to interpret the data with decryption keys.

3) **Blocking Transferability**

By blocking certain features or attributes during data transmission, such as IP addresses or metadata, attackers are limited in exploiting vulnerabilities. For example, blocking GPS coordinates from being sent with uploaded images on social media platforms reduces the risk of disclosing user locations to attackers.

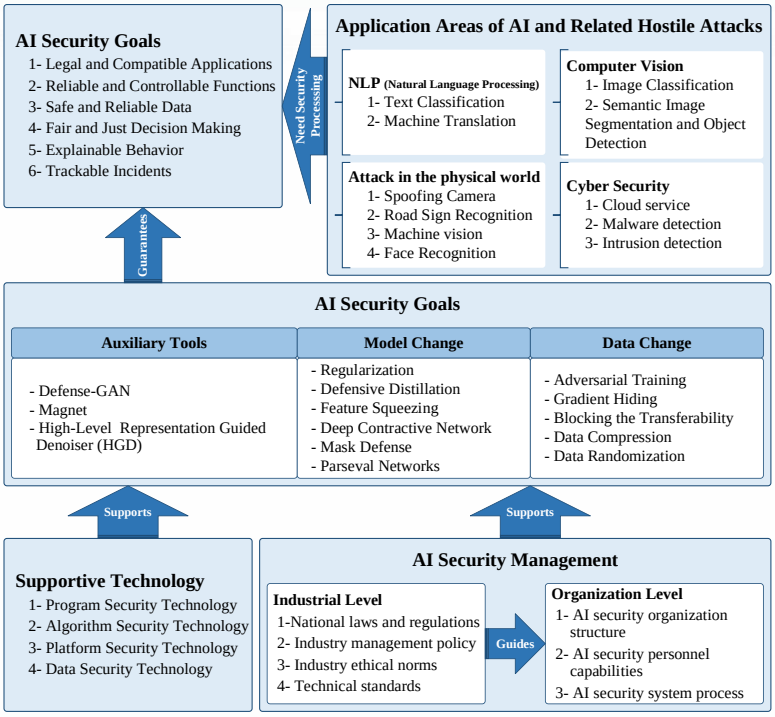

Figure 1. High-Level Architecture of AI Security Technologies

4) Data Aggregation

Aggregating data from multiple sources can help obscure individual data points, making it harder for attackers to identify patterns or extract sensitive information. For instance, combining health data from different hospitals for research can protect patient privacy while allowing researchers to analyze trends across a larger population.

5) Data Randomization

Introducing randomness into data, like altering image pixel values or adding noise to textual data, complicates adversarial attacks. For example, adding small amounts of noise to image pixel values can disrupt the patterns that adversarial attacks rely on, making it harder for attackers to generate effective perturbations.

B. Model Modification

1) Regularization:

Regularization techniques such as L1 or L2 prevent overfitting by adding a penalty term to the loss function. This encourages the model to learn simpler patterns and reduces its dependence on noisy or irrelevant features, thus improving its generalization performance. For instance, L2 regularization in a classification task prevents the model from overly relying on a small subset of features, enhancing its robustness against unseen data.

2) Defensive Distillation:

Defensive distillation involves training a smaller, more robust model using the predictions of a larger, more complex model as soft labels. By learning from the uncertainties and mistakes of the larger model, the smaller model becomes more resilient to adversarial attacks. For example, a smaller neural network trained on the predictions of a pre-trained deep convolutional network learns to resist adversarial perturbations better.

3) Feature Squeezing:

Feature squeezing reduces the impact of adversarial perturbations by lowering the bit depth of input features. Small perturbations that might deceive the model are magnified by quantizing input features to a lower bit depth, facilitating their detection and defense. For instance, reducing the color depth of input images from 32-bit to 8-bit helps identify and defend against subtle adversarial changes.

4) Contractive Neural Networks:

Contractive neural networks impose constraints on hidden representations during training, encouraging the model to be less sensitive to small changes in input space. This helps the model learn more robust features and reduces its sensitivity to adversarial perturbations. For example, contractive networks can encode textual data in natural language processing to make them more resistant to adversarial manipulation.

5) Masking Defense:

Masking defense involves restructuring the model to hide sensitive information or specific features during inference. By concealing certain features, the model prevents attackers from exploiting vulnerabilities associated with them, enhancing its security against adversarial attacks. For instance, a speech recognition

model might obscure specific phonetic features during prediction to defend against targeted adversarial attacks.

6) Sparse Convolutional Networks:

Sparse convolutional networks apply the sparse coding framework to neural network weight matrices, promoting robust learning and improving generalization performance. These regularization techniques enhance the neural network's resilience to adversarial attacks by encouraging more stable and meaningful representations. For instance, sparse networks impose constraints on weight matrices in image classification, aiding in stable training and better generalization.

C. Use of Auxiliary Tools

1) Defensive Generative Adversarial Models:

Defensive GANs can be employed in AI defense strategies by generating adversarial examples to augment the training dataset. Exposing the model to these adversarial examples during training helps it better distinguish between natural and adversarial inputs, thus improving its resistance to adversarial attacks. For instance, defensive GANs can generate adversarial images to train the AI model in image classification tasks, enabling it to recognize and defend against similar adversarial inputs.

2) Magnets:

In AI algorithms, "magnets" may refer to robust detection and tracking mechanisms within AI systems to neutralize erroneous inputs or attacks. These mechanisms use advanced algorithms to identify and mitigate potential threats, ensuring the security and reliability of AI systems. Magnets can represent sophisticated anomaly detection systems that protect AI models from adversarial attacks by identifying unusual patterns or behaviors in data streams.

3) Guided Noise Injection:

Guided noise injection is a defensive AI mechanism that injects carefully designed noise into model input data at a high level. By incorporating high-level feature representations from data, this noise injection helps the model ignore irrelevant or misleading features, increasing its resilience against adversarial attacks while maintaining performance on factual inputs. For example, injecting noise into the input data of a deep learning model for image recognition, guided by high-level features extracted from the data, helps the model ignore insignificant details, enhancing its resistance to adversarial attacks while preserving accuracy on factual inputs.

D. Reinforcements and Supporters of Defensive Strategies

1) AI Security Technologies

a) Application Security Technologies:

These technologies ensure the security of AI applications, protecting them from vulnerabilities and adversarial attacks.

b) Algorithm Security Technologies:

These focus on securing the algorithms, making them robust against manipulation and exploitation.

c) Data Security Technologies:

These technologies safeguard the data used in AI systems, ensuring its integrity and confidentiality.

d) Platform Security Technologies:

These ensure the security of the platforms on which AI systems run, protecting them from external threats.

2) AI Security Management

a) Organizational Level:

- **AI Security Organizational Structure**
  Establishing a dedicated structure within organizations to manage and oversee AI security.
- **Capabilities of AI Security Personnel**
  Enhancing the skills and capabilities of personnel responsible for AI security.
- **AI Security System Processes**
  Implementing robust processes to manage and maintain AI security within organizations.

b) Industry Level:

- **National Laws and Regulations**
  Developing and enforcing national laws and regulations to govern AI security.
- **Industry Management Policies**
  Establishing policies to manage AI security across various industries.
- **Industry Ethical Norms**
  Promoting ethical norms to guide the responsible use of AI technologies.
- **Technical Standards:**
  Setting technical standards to ensure the security and integrity of AI systems across industries.

The industry-level measures guide organisational-level practices, ensuring comprehensive and cohesive AI security management.

VI. Developing Program Foundations and Vision

The following overarching AI security ontology has been derived to establish program foundations and a roadmap for research in AI security, as shown in Fig 2.

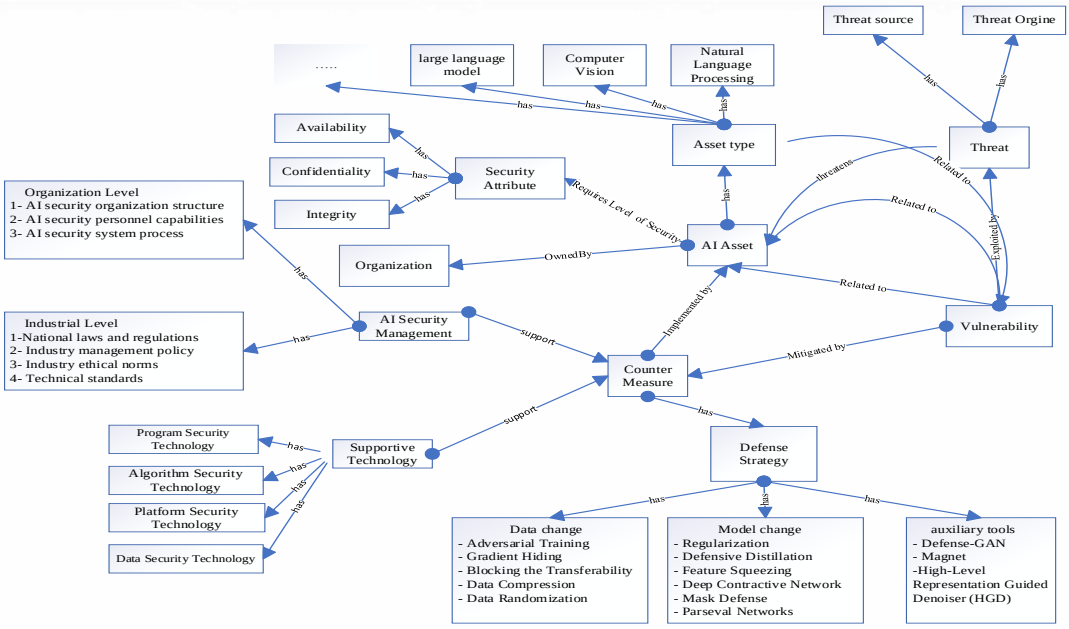

Figure 2. High-Level Ontology of Computational Security

According to the ontology graph shown in Fig 2, each organisation possesses a set of AI assets. These assets encompass various types, such as computer vision, natural language processing, large language models, and more. Each asset requires appropriate security features, such as accuracy, confidentiality, and integrity. These security levels can be compromised due to different types of vulnerabilities. Threat actors may exploit these vulnerabilities to breach the targeted security policies of these assets. Threats may originate from various human and machine sources or natural and artificial origins.

To mitigate vulnerabilities in AI assets, suitable defensive strategies must be implemented. These strategies include data alteration, model modification, and auxiliary tools. AI security management supports these defensive strategies, covering both organisational and industry levels, as well as supportive components and strategic defences.

VII. Conclusion

In conclusion, this paper underscores the pressing vulnerabilities of AI systems to adversarial attacks and provides a detailed taxonomy of these attacks. We reviewed various defence mechanisms and identified their strengths and limitations, emphasising the necessity for adaptable security strategies to protect AI applications in high-stakes sectors. By examining AI's role in cybersecurity, the study demonstrated the technology's potential in proactive threat detection, vulnerability assessment, and response management. As AI integrates into critical areas, ensuring its security against adversarial threats remains essential. Future research should prioritise developing advanced defence strategies, establishing regulatory frameworks, and implementing comprehensive security measures to safeguard the reliable deployment of AI technologies across diverse applications.

References

[1] Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences, 9(5), 909.

[2] Amyn, Lakhani., Naikade, Rohit. (2024). 1. Securing Machine Learning: Understanding Adversarial Attacks and Bias Mitigation. International journal of Innovative Science and research technology, doi: 10.38124/ijisrt/ijisrt24jun1671

[3] (2024). 9. Defense artificial neural networks from adversarial attacks using deep learning classification techniques. International Research Journal of Modernization in Engineering Technology and Science, doi: 10.56726/irjmets50791

[4] Chaudhary, P. K. AI, ML, AND LARGE LANGUAGE MODELS IN CYBERSECURITY. International Research Journal of Modernization in Engineering Technology and Science (Peer-Reviewed, Open Access, Fully Refereed International Journal),2024

[5] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," arXiv preprint arXiv:1801.00553, 2018.

[6] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," Applied Sciences, vol. 9, no. 5, p. 909, 2019.

[7] X. Liu, Y. Lin, H. Li, and J. Zhang, "Adversarial examples: Attacks on machine learning-based malware visualization detection methods," arXiv preprint arXiv:1808.01546, 2018

[8] V. Duddu, "A survey of adversarial machine learning in cyber warfare," Defence Science Journal, vol. 68, no. 4, pp. 356–366, 2018.

[9] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," ACM Computing Surveys (CSUR), vol. 54, no. 5, pp. 1–36, 2021.

[10] J. Gardiner and S. Nagaraja, "On the security of machine learning in malware c&c detection: A survey," ACM Computing Surveys (CSUR), vol. 49, no. 3, p. 59, 2016.

[11] Xingbin, Wang., Boyan, Zhao., Yu-Lan, Su., Sisi, Zhang., Fengkai, Yuan., Jun, Zhang., Dan, Meng., Rui, Hou. (2024). 2. A Hybrid Sparse-dense Defensive DNN Accelerator Architecture against Adversarial Example Attacks. ACM Transactions in Embedded Computing Systems, doi: 10.1145/3677318

[12] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," IEEE Communications Surveys & Tutorials, 2019.

[13] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2015.

[14] Stephn, Ojo., Moez, Krichen., Meznah, A., Alamro., Alaeddine, Mihoub. (2024). 6. TXAI-ADV: Trustworthy XAI for Defending AI Models against Adversarial Attacks in Realistic CIoT. Electronics, doi: 10.3390/electronics13091769

[15] S., Belhaouari. (2024). 7. Defense against adversarial attacks: robust and efficient compressed optimized neural networks.. Dental science reports, doi: 10.1038/s41598-024-56259-z

[16] Ibitoye, O., Abou-Khamis, R., Shehaby, M. E., Matrawy, A., & Shafiq, M. O. (2019). The Threat of Adversarial Attacks on Machine Learning in Network Security--A Survey. arXiv preprint arXiv:1911.02621.

[17] Stephen, Casper., Lennart, Schulze., Oam, Patel., Dylan, Hadfield-Menell. (2024). 10. Defending Against Unforeseen Failure Modes with Latent Adversarial Training. arXiv.org, doi: 10.48550/arxiv.2403.05030

[18] D. Le, B. Vo, and G. Nguyen, "A Survey on the Applications of Artificial Intelligence in Cybersecurity," arXiv preprint arXiv:1905.06233, 2019.

[19] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[20] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.

[21] S. Krasser, D. Carrel, and R. Beyah, "Automated Threat Intelligence and Machine Learning-based Network Intrusion Detection System," Proceedings of the 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), pp. 1-8, 2017.

[22] R. Shams, C. Fung, and T. Menzies, "Predicting Threats to Cybersecurity Using Time-Series Models," IEEE Transactions on Information Forensics and Security, vol. 14, no. 9, pp. 2343-2358, 2019.

[23] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp. 77- 91, 2018.

[24] Sangwan, R. S., Badr, Y., & Srinivasan, S. M. (2023). Cybersecurity for AI systems: A survey. Journal of Cybersecurity and Privacy, 3(2), 166-190.

[25] Oseni, Ayodeji, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. "Security and privacy for artificial intelligence: Opportunities and challenges." arXiv preprint arXiv:2102.04661 (2021).

[26] "Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk'". NPR.org. Retrieved 27 November 2017. Taeihagh, Araz. "Governance of artificial intelligence." Policy and Society 40, no. 2 (2021): 137-157.

[27] Syed, R. (2020). Cybersecurity vulnerability management: A conceptual ontology and cyber intelligence alert system. Information & Management, 57(6), 103334.

[28] Kong, Z., Xue, J., Wang, Y., Huang, L., Niu, Z., & Li, F. (2021). A survey on adversarial attack in the age of artificial intelligence. Wireless Communications and Mobile Computing, 2021(1), 4907754.

[29] https://dideban-ict.com/wp,2021

[30] Park, C., Lee, J., Kim, Y., Park, J. G., Kim, H., & Hong, D. (2022). An enhanced AI-based network intrusion detection system using generative adversarial networks. IEEE Internet of Things Journal, 10(3), 2330-2345.

[31] Taeihagh, A. (2021). Governance of artificial intelligence. Policy and society, 40(2), 137-157.

Seyyed Hadi Sajadi obtained his B.Sc. in Computer Engineering—software from Shahid Beheshti University, his M.Sc. in Socioeconomic Systems Engineering from Mazandaran University of Science and Technology, and his Ph.D. in Computer Engineering from Sharif University of Technology. He possesses research and executive experience in cybersecurity, social networks, and cyberspace regulation. Additionally, he has overseen the execution of numerous research projects in information technology and cybersecurity at various organisational and national levels.

Sasan Karamizadeh received his M.Sc. and Ph.D. in Computer Science from the Universiti Teknologi Malaysia(UTM) in 2012 and 2017, respectively. He spent one year as a postdoctoral researcher at the Iran Telecommunication Research Centre. Currently, he is with the Department of Security at the ICT Research Institute. He is a researcher and lecturer specialising in artificial intelligence, deep learning, machine learning, and information security, with professional experience in both roles within these fields.

Amirmansour Yadegari received a B.Sc. and M.Sc. in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran. In his work experience, he has directed several interdisciplinary research projects concerning Network Security, Internet Governance, Cyber Policymaking, and Spatial Planning. He is now a Faculty Member at the ICT Security Research Faculty, ICT Research Institute.