International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

Adversarial Attacks on AI Agents: Data Privacy Risks and Regulatory Implications

Date of Submission: 28-05-2025 Date of Acceptance: 08-06-2025

I. INTRODUCTION

Recent improvements in technology,alongside significant growth in computationalcapacities have made it easier for people to use AI approaches in many areas, including healthcare, finance, gaming, and self-driving cars(Papernot et al., 2016). Machine learning (ML) and deep learning (DL) models have made it possible to automate difficult processes and provide new features that were not possible with older technologies(Oseni et al., 2021). AI agents are now a key part of many applications, from virtual assistants to advanced data analysis systems. They show intelligent behaviour by being autonomous, reactive, and proactive. These agents, which are powered by large language models (LLMs) like GPT-4, have changed how jobs are done and are doing great work in areas like healthcare, customer service, and finance(Achiam et al., 2023). But as AI technologies get better and spread to important systems, they also make them more vulnerable. Since (Szegedy et al., 2013)groundbreaking paper showed that neural networks are vulnerable to adversarial attacks, there has been a lot of research into adversarial approaches that target AI models. These assaults can happen during training or testing and are very dangerous for data privacy, system integrity, and user trust. This shows how important it is to grasp what they mean and make good rules to protect people.

Building on their autonomous and proactive nature, AI agents are interconnected systems that constantly engage with their environment and users through cycles of perception, reasoning, and action(Liu et al., 2023). This is because they are autonomous and proactive. These agents are more than just big language models. They are complex systems that combine different parts, like perception modules that understand complicated, multi-modal user inputs, brain modules powered by LLMs that reason and plan, and action modules that carry out tasks or use outside tools(Moskal et al., 2023). There are several types of AI agents, such as chatbots that help with conversations, self-driving cars that can drive in changing surroundings, and recommendation engines that make material more relevant to each user. This multi-layered architecture lets AI agents provide smart and flexible services in important areas like healthcare, banking, and customer support(Yang et al., 2024).

However, the expanding deployment of AI agentsin important and sensitive areascreate new security lapses, especially when it comes to protecting and keeping data safe(Shestak & Tsyplakova, 2023). These agents are complicated because they have to deal with many steps of user contact, complicated internal processes, and work in different, sometimes untrusted contexts. This makes them vulnerable to many types of attacks. For example, adversarial assaults like prompt injection or jailbreak exploits can change the perception module, making the agent give out wrong or damaging information(Pöhler et al., 2024). Also, using outside tools or other agents may add even more hazards, such as data leaks or the unauthorised running of harmful code. Also, AI agents often use huge datasets, including fake data made by AI itself, which can accidentally reveal background knowledge that enemies use to break privacy(Chen et al., 2020). These problems show how important it is to have strong defence systems and rules that particularly deal with the security and privacy threats that AI agents bring when they work together in real-world, multi-agent ecosystems(Schwinn et al., 2023).

Given these vulnerabilities, it's very important to know about adversarial threats in AI systems so that strong defences can be built and rules can be followed. According to(Gupta et al., 2024), Adversarial Machine Learning (AML) techniques offer a way to find and reduce risks in AI models. They also stress that AML can be used for both defence and attack. However, it also adds complications, such false positives and high computing costs, that need to be properly managed.(Hazra et al., 2025) makes the same point about the specific hazards that agentic AI systems represent. Because they are autonomous and have access privileges, they are open to new attack vectors including prompt injection and goal manipulation(Chern et al., 2024). These assaults can put data privacy, system integrity, and user

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

confidence at risk, especially in important areas like finance and healthcare(Izacard & Grave, 2020). For example, evasion and poisoning assaults, which are widespread in classical AI, are even worse in agentic systems(Amerirad et al., 2023). This lets attackers take advantage of decision-making processes or leak sensitive data. To lower these risks, it is important to combine AI-based threat detection with proactive defence measures like continuous vulnerability scanning and adversarial(Balamurugan & Research, 2024). But both studies warn that ethical and legal frameworks need to change along with technology to deal with the problems of dual-use and make sure AI systems are used responsibly.

This paper provides a comprehensive review of adversarial attacks on AI agents, with a focus on the data privacy risks and regulatory issues that come up because of these threats. We start by looking at what adversarial assaults are and what kinds there are. Then we explain how they take advantage of weaknesses in AI models when they are being trained and used. After that, we look at the unique privacy issues that these assaults offer to different AI applications, such as healthcare, autonomous systems, natural language processing, and financial services. The study then looks at the current rules and regulations that deal with AI security and data privacy, pointing out where they fall short and what makes it hard to deal with threats from adversaries. We also talk about new ways to protect sensitive data and make AI more robust. Finally, we talk about where future research should go and how important it is for countries to work together to make sure that AI systems and users are safe from being used against them.

II. UNDERSTANDING ADVERSARIAL ATTACKS

Adversarial attacks are basically planned attempts to change how AI systems work. These manipulations entail putting false or carefully constructed prompts or data into the system with the clear intention of getting it to give wrong or biased results. According to (Lekota), these kinds of assaults take advantage of weaknesses in AI algorithms by making little changes to the input data, which makes the AI system give wrong or unexpected answers.

"Perturbations" are what these attacks are all about. (Qiu et al., 2019)credit (Szegedy et al., 2013) as the first to talk about adversarial samples. These are made by adding small, often invisible changes to the input samples of a machine learning model. The goal of these little adjustments is to have the model misclassify the input with a high level of confidence, even though a person looking at the original and the altered sample would not be able to tell the difference. For example, (Qiu et al., 2019)show that a small change to an image of a "panda" may make a complex model confidently misclassify it as a "gibbon" despite still seeming the same to people.

Adversarial instances are the consequent inputs derived from perturbations. According to (He et al., 2018), these are precisely engineered samples generated by including minute, human-invisible changes to original, accurately categorized inputs. Their sole objective is to deceive a machine learning model into misclassifying them with considerable certainty. When an original input, represented as x, is correctly classified by a model M as y_true (M(x) = y_true), an adversarial example x' is produced by introducing a minor perturbation to x, ensuring that x' retains perceptual similarity to x while resulting in $M(x') \neq y\_true$.

The presence of antagonistic cases has incited comprehensive investigation into their fundamental origins.(Mohammed, 2025)observe that researchers have presented multiple reasons. Certain theories ascribe their origin to the model's overfitting or inadequate regularization, resulting in a diminished ability to generalize when faced with novel data. Alternative viewpoints indicate that adversarial samples emerge from the pronounced nonlinearity intrinsic to deep neural networks. Nonetheless, a significant explanation proposed by (Goodfellow et al., 2014) asserts that the key factor is the linear behavior of deep neural networks in high-dimensional spaces. In such contexts, little perturbations, when applied cumulatively across multiple dimensions, might aggregate to produce a substantial, effective change in the output, resulting in misclassification. The intrinsic characteristics of adversarial attacks, which involve perturbations that are visually undetectable to humans but very successful in misleading AI, provide a considerable problem.

This phenomenon indicates that AI models, despite their sophisticated capabilities, function on principles that enable incremental, tiny alterations across numerous features to move an input across a decision border in a way that eludes human holistic awareness(Gupta et al., 2023). This trait signifies a significant vulnerability, indicating that even highly advanced visual or auditory AI systems might be deceived by inputs that seem innocuous to human operators. This has significant consequences for the reliability and security of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

human-AI interactions. If AI systems can be deceived by inputs that are indistinguishable from normal ones to humans, the efficacy of human oversight is inherently constrained(Zhang et al., 2019).This prompts significant inquiries about the dependability of AI in safety-critical domains, such as autonomous vehicles or medical diagnostics, where human verification is frequently regarded as the paramount safeguard. The observation indicates that a "human-in-the-loop" methodology may be inadequate for identifying these advanced attacks(Baluja & Fischer, 2017).

2.1 Types of Adversarial Attacks

2.1.1 Evasion Attacks

Evasion attacks occur during the testing or inference phase of an AI model. In this situation, an adversary directly alters input samples to evade detection systems or to provide incorrect outputs, critically without modifying the foundational model or its training data [1, 1, 1, 1]. According to (Oseni et al., 2021) these attacks leverage inherent flaws in the model to produce targeted, intended errors. Adversaries carefully design these inputs with micro-perturbations and utilize sophisticated optimization methods to disrupt the model's prediction skills.

Evasion attack objectives may differas they might be error-generic, intending to mislead classification irrespective of the specific output class, or error-specific, aimed at inducing misclassification into a designated category. These attacks significantly threaten the integrity of machine learning models during their operational runtime. Real-world instances highlight their gravity, such as the misunderstanding of road signs by autonomous vehicles. (Chen et al., 2024) reported the effective circumvention of antivirus software, and the manipulation of concealed voice commands in smart assistants.

2.1.2 Poisoning Attacks

Poisoning assaults transpire during the training phase of an AI model. Their main aim is to diminish the model's quality or to discreetly modify its decision bounds by introducing harmful data into the training dataset [1, 1, 1, 1]. (Barreno et al., 2006), referenced by (Qiu et al., 2019), initially used the phrase "poisoning attacks" and illustrated how an adversary, possessing knowledge of a machine learning model, can alter the original distribution of a training dataset by manipulating the training data. (Biggio et al., 2012) shows that randomly altering 40% of training labels significantly diminishes the performance of classifiers employing SVMs. This may entail the intentional introduction of inaccurately labeled data points, the calculated integration of hostile samples into the training dataset, or the malevolent alteration of existing labels or feature values.

Attackers utilizing poisoning strategies may pursue error-generic outcomes, intending to generate a broad array of misclassifications regardless of the target class, or error-specific outcomes, where the objective is to provoke specific, preset misclassifications. A notably deceptive type of poisoning is the "backdoor attack," in which particular "trigger" patterns are clandestinely integrated into the training data.(Polemi et al., 2024)assert that software developers may intentionally incorporate backdoors for illicit, commercial, or strategic reasons, and their existence can be uncovered by thorough acceptance testing initiatives. Upon training the model with this compromised dataset, it acquires the ability to link these triggers to specific, attacker-preferred outputs, hence enabling the attacker to manipulate the model's behavior subsequently by introducing inputs that contain these concealed triggers.

Poisoning assaults have a direct and serious impact, undermining the integrity and availability of AI models. Identifying these attacks is a considerable challenge because to the intricacy of the manipulations; the alterations may be minimal and hard to discern, particularly within extensive databases. Thus, the complete ramifications of these attacks may only be revealed post-deployment of the model in a practical environment. The notorious Microsoft Tay chatbot incident, in which malevolent individuals effectively used the AI to produce inappropriate material, exemplifies how data tampering during training can significantly influence model behavior and provoke public indignation(Hu & Tan, 2022).

2.1.3 Model Inversion Attacks

Model inversion attacks pose a significant risk to data privacy. These assaults allow an adversary to reconstruct confidential training inputs from a model's predictions. (Fredrikson et al., 2015) illustrated that these attacks employ machine learning (ML) application programming interfaces (APIs) to deduce sensitive attributes. The process entails carefully probing the model and rigorously examining its outputs to extract personal or secret information from the initial training dataset.

The immediate consequence of these attacks is a substantial privacy threat, since they may result in the disclosure of sensitive

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

information, promote identity theft, or allow for the exploitation of private data. Such assaults highlight significant privacy issues associated with granting API access to machine learning models trained on sensitive data.(Ma et al., 2023) emphasize that, although typically effective in black-box scenarios where the attacker lacks internal model knowledge, the computing complexity of such methods can increase exponentially with input size for non-linear models, presenting a realistic obstacle for the attacker.

2.1.4 Membership Inference Attacks (MIA)

Membership Inference Attacks (MIA) primarily target privacy infringements. The aim of MIA is to ascertain if a certain data item or an individual's record was incorporated in the model's training dataset. (Shokri et al., 2017) examined the distinctions across models to determine the presence of an input inside the training dataset for supervised models. This is often accomplished by examining temporal patterns in the learning model's behavior or by scrutinizing nuanced variations in the model's responses, occasionally through the training of a "shadow model" that replicates the behavior of the targeted model(Akhtar et al., 2017).

The immediate consequence of MIAs is a violation of privacy, enabling malevolent actors to deduce sensitive attributes of individuals whose data was utilized in the model's training. (Melis et al., 2019)shown how malevolent participants might deduce sensitive attributes in distributed learning. This may have significant consequences for data privacy, especially in dispersed learning contexts where data from several sources is consolidated for model training without the direct exchange of raw data.

2.1.5 Model and Functionality Stealing

Model and functionality theft attacks entail the unauthorized acquisition or reproduction of an AI model's intellectual property or functionalities.Model extraction, a variant of model theft, transpires when an adversary monitors input-output pairs derived from a target model's predictions. By utilizing these findings, the attacker can derive a set of parameters and develop a new model that closely mimics the behavior and performance of the original target model. (Tramèr et al., 2016)exhibited a successful model extraction attack on online machine learning service providers.Functionality extraction, a pertinent attack, seeks to develop "knock-off" models that replicate the target model's functionalities based exclusively on its design inputs and observed outputs, frequently enabled through Machine Learning as a Service (MLaaS) queries(Sarker & Privacy, 2023).

The repercussions of these attacks are complex: they can jeopardize the monetization strategies of model developers, infringe upon the privacy of training data by disclosing implicit patterns, and enable additional evasion attacks by equipping the attacker with a functional replica of the target system. Ultimately, these assaults facilitate the appropriation of intellectual property by permitting opponents to reconstruct or recreate valuable AI models, resulting in considerable economic and competitive disadvantages for the original developers(Hu et al., 2023).

2.1.6 Prompt Injection and Jailbreak

Malicious prompt injection uses multimodal prompts to control AI agents.The agent's "brain" must be influenced in order to do bad things that go against moderation and rules. According to (Khan et al., 2024) attackers can take advantage of prompt injection vulnerabilities by changing the AI's behavior to give false or harmful results in a database query environment.They might utilize "goal hijacking," which means changing the agent's instructions to their own, or "prompt leakage," which means getting the LLM to leak important, pre-planned instructions or internal settings(Hou et al., 2024).

Jailbreaking affects AI agents' LLMs in a way that is both similar and different.The model can make material that goes beyond its limitations if it doesn't follow its safety, ethical, or operational rules. (Deng et al., 2025) say that automatic tools like Jailbreaker can give jailbreak prompts in commercial AI systems.This is possible with manual design using single-step or multi-step prompts or automated jailbreak prompt systems. Prompt injection can put the security of systems and data at risk, commit crimes, and leak API keys and secrets. Prompt injection attacks can happen in any app and can lead to remote code execution, fake SQL queries, and biased AI bot outputs(H. Li et al., 2023).

It isimportant to give AI agents more ways to attack than only model-based flaws. Recent "agentic AI" assaults have gone from modifying what a model is supposed to do to quickly injecting and jailbreaking the agent's "perception" and "brain" modules(Jing et al., 2021).AI agents can access databases and carry out orders on their own, which means that quick injection can be used to change data and break into systems.AI agents

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

become entry points for hackers in networked systems, moving attacks away from single models.With this growth, cybersecurity that focuses on model robustness is not enough. AI agent lifetime, complex interactions, user input validation, secure tool integration, and agent behavior monitoring are all things that security solutions must now include. Successful attacks can automate hostile operations on a huge scale, which makes it harder to find and stop them. AI bots that work on their own need a complete security solution that sees AI agents as systems that are complicated and connected(Muça et al., 2024).

Table 2.1: Taxonomy of Adversarial Attacks on AI Agents

Attack Type	Primary Target	AI Lifecycle Phase	Attacker Goal	Key Characteristics/Examples	Relevant Snippet IDs
Evasion	Input Data	Inference/Testing	Misclassification, Bypass Detection	Imperceptible perturbations, mislead outputs without model alteration	[1, 1, 1, 1]
Poisoning	Training Data, Model Parameters	Training	Data Integrity Compromise, Biased Outcomes, Backdoor Insertion	Mislabeled data, adversarial samples in training, backdoor triggers	[1, 1, 1, 1]
Model Inversion	Model Parameters, Training Data	Inference/Testing	Data Privacy Breach, Re-identification, Sensitive Attribute Disclosure	Reconstruct training inputs from model predictions, infer personal info	
Membership Inference	Training Data	Inference/Testing	Data Privacy Breach, Membership Disclosure	Determine if specific data point was in training set, shadow models	
Model & Functionality Stealing	Model Parameters, Model Logic, Functionality	Inference/Testing, Deployment	IP Theft, Subvert Monetization, Facilitate Evasion	Extract model parameters from input-output pairs, create "knock-off" models	
Prompt Injection	Agent Behavior, Instructions, External Tools	Deployment/Interaction	Unauthorized Access, System Control, Data Exfiltration, Misinformation	Malicious prompts, goal hijacking, prompt leakage, malicious SQL queries	[1, 1, 1, 1]
Jailbreak	LLM Safety Guidelines, Agent Behavior	Deployment/Interaction	Bypass Safety Constraints, Generate Harmful Content, Malicious Actions	Manual or automated prompts, multi-modal inputs, domino effect in multi-agents	

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

III. COMMON TECHNIQUES USED IN ADVERSARIAL ATTACKS

3.1 Gradient-Based Techniques (White-Box)

White-box attacks provide the enemy full access to the inner workings of a target model, such as its architecture, parameters, and training data. This full access lets you make very effective adversarial instances with great care. (Goodfellow et al., 2014) came up with the Fast Gradient Sign Method (FGSM), which is one of the best white-box approaches. This method finds the gradient of the loss function with respect to the input data and then changes the input in the direction that causes the most loss to cause misclassification. People like FGSM because it is easy to use and doesn't take up a lot of computer power. Projected Gradient Descent (PGD) is an iterative version of FGSM(Liu et al., 2019). PGD makes modest changes to the adversarial example several times, each time getting better at it. This usually makes PGD stronger than FGSM, but it does need more computing power.

In addition to these basic tactics, there are other important white-box attacks that use different strategies. (Papernot et al., 2017)developed the Jacobian-based Saliency Map Attack (JSMA), which employs the Jacobian matrix to figure out how input features affect the output. It then uses saliency maps to find precise areas to change, usually by changing just a few pixels.(Moosavi-Dezfooli et al., 2016) came up with DeepFool, an iterative algorithm that finds the smallest norm adversarial perturbations. It is known for being fast and for being able to create perturbations that are hard to see. The Carlini & Wagner (C&W) attack, which was proposed by Carlini and Wagner in 2017, is widely thought to be one of the best ways to make adversarial examples. It does this by trying to lower a selected distance metric while making sure the example is still a valid input. There are numerous other gradient-based methods that add to the wide range of white-box adversarial strategies(Kiran & Kumar, 2023).

3.2 Black-Box Techniques

Black-box attacks assume that the attacker doesn't know anything about the target model, only what it does with its inputs and outputs. [1, 1, 1]. Using the transferability of adversarial examples is a key part of black-box assaults. According to (Qiu et al., 2019) an adversarial example produced for one model can often trick another, even if the two models have distinct topologies. Attackers take advantage of this by making examples with a local "substitute model" and then using them on the target model.

Another popular method is to ask the target model questions to find out where it is weak. (Bazaga et al., 2021) talk about how attackers can figure out the model's attributes and decision boundaries without having access to its internals by systematically submitting inputs and watching outputs. (Moosavi-Dezfooli et al., 2016) came up with the idea of Universal Adversarial Perturbations (UAP), which are supposed to provide one perturbation that works on a wide variety of inputs and doesn't depend on the image.

Finally, Adversarial Transformation Networks (ATN) are feed-forward neural networks that learn how to make adversarial instances. They can make changes to images that are specific to them or that don't depend on them in order to get over defenses.

Table 3.1: Common Adversarial Attack Techniques

Technique	Category	Attacker Knowledge	Principle/Mechanism	Pros (for attacker)	Cons (for attacker)	Relevant Snippet IDs
FGSM	Gradient-Based	White-Box	Perturbs input in gradient direction to maximize loss	Efficient, simple, computationally fast	Coarse approximation, less precise	[1, 1, 1, 1]
PGD	Gradient-Based	White-Box	Iterative FGSM, refining perturbations	Stronger, more effective adversarial examples	More computationally expensive, slower	[1, 1, 1, 1]
JSMA	Gradient-Based	White-Box	Uses Jacobian matrix to find sensitive input	Precise, few-pixel changes,	Computationally intensive	[1, 1]

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

			regions	effective for targeted attacks		
DeepFool	Optimization-Based	White-Box	Computes minimal norm perturbations iteratively	Computationally efficient, imperceptible perturbations	Relies on linearity assumption	[1, 1]
C&W	Optimization-Based	White-Box	Reformulates misclassification as objective function, minimizes distance	Very effective, strong against defenses	Computationally demanding	[1, 1]
One Pixel Attack	Black-Box Specific	Black-Box	Changes single pixel to cause misclassification	Extreme stealth, no model knowledge needed	Limited to image classification	[1, 1]
UAP	Black-Box Specific	Black-Box	Generates image-agnostic perturbation for widespread fooling	Universal effectiveness, efficient for many inputs	May not be perfectly imperceptible	[1, 1]
ATN	Black-Box Specific	Black-Box	Neural network trained to generate perturbations	Can bypass existing defenses, diverse outputs	Requires training a separate network	
Transferability Exploitation	Transferability-Based	Black-Box	Uses substitute model to create examples for target	No direct model access needed, practical	Effectiveness varies with model similarity	
Querying	Black-Box Specific	Black-Box	Interacts with model to infer vulnerabilities	No internal model knowledge needed	Can be slow, detectable by rate limiting	[1, 1, 1]

IV. DATA PRIVACY RISKS POSED BY ADVERSARIAL ATTACKS ON AI AGENTS

4.1. Direct Privacy Breaches

4.1.1. Model Inversion and Membership Inference Risks

Direct privacy breaches occur when adversarial directly cause sensitive personal data to be exposed or reconstructed. Model Inversion attacks are a great example since they let attackers rebuild sensitive training data, such facial photos, right from a model's predictions. (Fredrikson et al., 2015) showed that these attacks use machine learning APIs to figure out sensitive features, which means that they can reveal private information that the model was trained on, even if the original data was anonymized. The rebuilt inputs aren't exact copies, but they can provide typical values for a certain class. This raises big privacy issues when models trained on sensitive datasets are given access to APIs. In addition, Membership Inference Attacks (MIA), which (Shokri et al., 2017) looked at, let attackers figure out with a high degree of certainty whether a certain data item was part of a model's training dataset. A "shadow model" that looks like the target is sometimes used to carry out this kind of assault, which is a blatant invasion of privacy, especially for models trained on very private information like medical records. When these attacks happen together, they can make it possible to identify people again and reveal their private information, which makes standard data anonymization measures less effective(Huang et al., 2021).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

4.1.2 Data Reconstruction Attacks

Majeed and Hwang (2023) highlight that adversaries can leverage AI-generated synthetic data (SD) as a powerful form of background knowledge (BK) to reconstruct anonymized data back into its original, identifiable form with considerable accuracy. This reconstructed data can then be used to infer Quasi-Identifiers (QIDs) or Sensitive Attributes (SAs) of target individuals, leading to profound privacy breaches.

4.1.3 Exploitation of Synthetic Data as Background Knowledge

Despite its protective capabilities, AI has a "dark side" in which adversaries can aggregate and exploit AI-generated knowledge, particularly high-quality synthetic data (SD), as potent background knowledge to compromise individual privacy, as argued by (Majeed & Hwang, 2023). They say that most of the anonymization solutions that are now available have not taken this SD-based knowledge into account in the past, making them weak. The ways this abuse works are subtle but effective: high-quality synthetic data can closely match the statistical distributions of sensitive variables and often keep main categories within Quasi-Identifier (QID) values. This makes it much more likely that someone will be able to identify someone else or break the privacy of a group, which goes against the idea that synthetic data always provides strong privacy protections, as shown by (Rocher et al., 2019). This new way of attacking can find identities or private information in datasets that were previously anonymized. This is a big threat that is often underestimated and goes against what we thought we knew about data privacy. It also means that we need new privacy models that can handle attacks that use advanced synthetic data(Di Vaio et al., 2020).

4.2 Indirect Privacy Risks

4.2.1 Unauthorized Data Access and Data Exposure through Compromised AI Agents

In addition to direct breaches, adversarial attacks create a number of indirect ways to undermine privacy, frequently taking advantage of the built-in features of AI agents. (Khan et al., 2024) say that AI agents make the attack surface much bigger because they can directly access database systems. This makes them possible entry points for bad actors. If an AI agent is hacked, for example by a successful prompt injection attack, it can be used to get access to private information without permission. This can cause serious privacy infractions and big legal problems for businesses. AI systems are already quite complicated, which makes it much more likely that data may leak or be hacked, either by accident or on purpose by someone trying to hurt the system(Azhar et al., 2023).

4.2.2 Misuse of Personal or Confidential Data

AI agents that have been hacked or influenced through several types of assaults can accidentally reveal private information in their outputs, which can lead to the release of sensitive information that wasn't meant to be shared. (Khan et al., 2024) point out that the fact that AI agents can act on their own makes this risk worse. If an enemy gains control of them, they can carry out commands that make it easier to steal or corrupt vast amounts of data. This includes the ability to change AI-generated searches to get to, change, or even delete sensitive data in associated databases.

4.2.3 Impact on User Trust and Data Integrity

The public perception that AI systems have unrestricted access to personal data, particularly if these systems are known to be vulnerable to compromise, can substantially impede the widespread adoption of AI-powered services and severely erode user confidence, as noted by (Park et al., 2023). (Lekota)also stresses that adversarial attacks directly damage the integrity of the information that AI systems process and create by changing data or how models work. This could lead to making bad decisions based on data that has been tampered with, which could have serious effects in important situations.

4.2.4 API Usage Risks

Companies that integrate Large Language Model (LLM) APIs suffer indirect privacy risks because submitting queries including personal or confidential data to external providers may mistakenly reveal sensitive information to third parties, causing major compliance difficulties (Hu et al., 2021). Also, AI agents' natural usefulness and independence, which often need direct access to large datasets and the ability to take action, make data security a double-edged sword. These features are useful, but they also make the agent a major security risk if they are hacked, because a successful attack can lead to automated data theft, manipulation, or coordinated large-scale attacks. The rise from model prediction compromise to systemic threat shows how important it is to have strong security measures in place for AI agents, especially when they are working with sensitive

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 06 June 2025, pp: 34-47 www.ijaem.net ISSN: 2395-5252

data or external APIs(Bhuiyan et al., 2018). These measures should include strict access constraints and regular audits.

4.3 Broader Impact on Confidentiality, Integrity, and Availability (CIA) of AI Systems

The essential security principles of Confidentiality, Integrity, and Availability (CIA) that underpin robust AI systems are fundamentally compromised by adversarial attacks. Attacks such as model inversion, membership inference, and prompt leakage can compromise confidentiality by revealing sensitive training data, identifying individuals, or exposing proprietary instructions and API keys(Samonas & Coss, 2014). The integrity of model outputs and connected databases is at risk due to evasion attacks and data manipulation, while poisoning attacks that corrupt training data and models result in biased and unreliable predictions. Therefore, integrity is significantly compromised. Finally, attacks that result in system slowdowns, denial-of-service (DoS), or resource exhaustion can have a substantial impact on availability, thereby preventing legitimate users from accessing AI functionalities(Yee & Zolkipli, 2021). The critical interconnectedness of security and privacy is emphasized by this threat landscape. Robust security measures are necessary to ensure true privacy, as any security vulnerability undoubtedly affects privacy.

V. REGULATORY IMPLICATIONS AND GOVERNANCE FRAMEWORKS

5.1 Existing and Emerging Regulations

The European Union's AI Act is a groundbreaking piece of legislation that is swiftly changing the regulatory landscape for AI. This law protects basic rights from AI systems that are too risky. It makes it illegal for apps that are considered "unacceptable risk" (like those that violate basic rights or manipulate people) and sets strict rules for "high-risk" AI when it comes to data governance, transparency, and security (Polemi et al., 2024). The General Data Protection Regulation (GDPR) and the NIS 2 Directive are two important pieces of cybersecurity law that set rigorous rules for how personal data must be handled and make key sectors more resilient to cyberattacks. As AI threats become more advanced, laws need to change to make it clear who is responsible for AI-related accidents, lower risks, and build public trust (Ludlow et al., 2015).

5.2 Key Principles for Trustworthy AI

AI trustworthiness is a multifaceted notion that encompasses features such as cybersecurity, transparency, robustness, correctness, data quality and governance, human oversight, and meticulous record keeping (Polemi et al., 2024). The NIST AI Risk Management Framework (AI RMF 1.0) also lists traits such being valid, trustworthy, secure, explainable, privacy-enhanced, fair, accountable, and open. These traits are all connected and heavily affected by human factors, such as biases and mistakes made during the creation and use of AI. User comprehension and acceptance are equally important for public trust.Therefore, the EU AI Act says that risk management systems must take into account both technological and human-related concerns(B. Li et al., 2023).

5.3 Frameworks and Guidelines for AI Security Governance

There are a number of frameworks that provide systematic ways to control AI security. The NIST AI Risk Management Framework (AI RMF) is one of the most important. It has a whole set of steps for governing, mapping, measuring, and managing AI-related risks (De Almeida et al., 2021). The OWASP Top 10 for Machine Learning Security and LLM Applications lists important weaknesses in ML security (like Data Poisoning and Membership Inference) and LLM applications (like Prompt Injection and Model Denial of Service) and gives ways to avoid them and factors that make them more likely to happen (Wirtz et al., 2022). MITRE ATLAS (Adversarial Threat Landscape for AI Systems) is also an important resource that lists the tactics and techniques that adversaries use against AI. However, it also points out that many LLM-related attack techniques still don't have clear mitigation measures (Gianni et al., 2022; Birkstedt et al., 2023). The quick changes in AI agent-specific dangers, such swift injection and jailbreak, show a big gap between general AI governance concepts and specific, enforceable rules. This shows that regulatory authorities need to be able to quickly react to this "arms race."

VI. MITIGATION AND DEFENSE STRATEGIES AGAINST ADVERSARIAL ATTACKS ON AI AGENTS

6.1 Technical Defenses

Developing robust AI systems against adversarial attacks involves several proactive training and data-centric defenses. Adversarial Training purposefully adds adversarial cases to the

learning process to make a model more robust and able to generalize (Tramèr et al., 2016; Bonawitz et al., 2019). Data Sanitization and Input Preprocessing work together to carefully remove contaminated samples from training datasets and make inputs stronger to stop bad data from getting in (Baniecki & Biecek, 2024). Regularization and Defensive Distillation are two methods that make models even more robust. They do this by adding penalty terms to cost functions or making output surfaces smoother, which makes them less sensitive to changes (Biggio et al., 2012; Papernot et al., 2016). Feature Squeezing and other noise reduction methods make data representation easier to get rid of adversarial noise. Ensemble Methods, on the other hand, integrate many AI models to make them more resilient overall, which makes it more difficult for attackers to break into all of them at the same time.

Advanced cryptography and operational security measures are just as important as training improvements. Cryptography, such as Homomorphic Encryption (HE) and Secure Multiparty Computation (SMPC), protects privacy well by letting people do computations on encrypted data or learn together without giving away raw inputs. However, this often requires a lot of extra computing power (Alotaibi & Rassam, 2023). Robust Aggregation approaches like Krum stop bad clients from ruining the aggregated model in distributed learning (Blanchard et al., 2017). Differential Privacy (DP) adds noise to data or gradients to protect privacy. It is especially good at stopping membership inference attacks (Chen et al., 2019). Isolated Sandboxes and Emulators are very important for AI agents to use to simulate interactions and check for dangers before they happen in the actual world (Deng et al., 2025). A key factor to keep in mind for all these solutions is the built-in trade-off between privacy and utility. Many defenses may lower accuracy on "clean" data or add a lot of extra work, therefore AI system design and deployment need to be carefully balanced(Bountakas et al., 2023).

6.2 Organizational and Policy Measures

Comprehensive organizational strategies and policy frameworks are essential components of effective AI security, in addition to technical implementations. It is very important to have Comprehensive Risk Management and Incident Response Plans. These plans need to include a proactive security posture and set procedures for identifying and mitigating AI risks throughout the development lifecycle(Ilahi et al., 2021). Continuous Monitoring and Anomaly Detection Systems are essential for finding threats early on. They check inputs carefully and look for strange AI activity that could mean an assault (Brundage et al., 2018; Ijiga et al., 2024). To lower the risks that come from relying on outside sources, you need to do Supply Chain Auditing and follow a Secure Development Lifecycle. This will make sure that trustworthy tools are used and that protections are built in from the start (Deng et al., 2025; Lekota, 2024). Finally, educating and making users aware of security issues is essential for creating a strong human defense layer. This gives users the tools they need to spot suspicious activity and improves incident management throughout the business (Khan et al., 2024; Mohammed, 2025).

In addition to these operational steps, Ethical Guidelines and Accountability for Developers are very important for encouraging ethical AI development that is fair, open, and proactive in dealing with bias and privacy issues (Mohammed, 2025; Lekota, 2024). The fact that AI security is so broad shows how important it is to have a multi-faceted, interdisciplinary approach. Securing AI agents isn't only a computer science problem; it needs people from cybersecurity, AI, social psychology, behavior, and ethics to work together (Polemi et al., 2024). Organizations need to create a security culture that takes into account all aspects of security, including ethical design, ongoing legal compliance, and managing human behavior as part of the threat landscape. This is in line with a socio-technical approach to future AI security research and development.

VII. CONCLUSION

Dynamically evolving adversarial attacks on AI agents pose a substantial and dynamic hazard to the integrity, confidentiality, and availability of AI systems. These assaults have gone beyond typical model-based weaknesses and are now taking use of the larger attack surface that comes with autonomous, interacting AI agents. There are a lot of different types of these assaults, from little changes that cause misclassification to complex prompt injections that can help steal data and completely take over a system. The use of AI-generated synthetic data is a particularly important and new threat vector that, ironically, can give enemies a lot of background knowledge to use to break privacy.

Because adversarial AI is always changing, there is a constant "arms race" between attackers and defenders. As Ma et al. (2023) point out, new attack methods are always being

developed to get around old protections. This means that security must be proactive and adaptable. Because of this continuing struggle, no one protection mechanism can solve the problem for good. Instead, constant awareness and adaptation are the most important things to do.

Deng et al. (2025) say that future research should focus on making AI agents' input inspection mechanisms more accurate and efficient, building stronger LLM-based agents that can handle more advanced manipulations, and making sure that complex AI systems have secure memory management. Mohammed (2025) also stresses the importance of coming up with new ways to undertake adversarial training and strong sampling designs in order to effectively fight against the changing tactics used by enemies. Polemi et al. (2024) also stress the importance of creating clear ethical rules for the creation and use of AI in cybersecurity. This should be done together with encouraging deep collaboration between technical specialists, ethicists, and social scientists to create an AI ecosystem that is truly strong and trustworthy and fits with what society wants and expects. [1, 1].

REFERENCES

[1]. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L.,…Anadkat, S. J. a. p. a. (2023). Gpt-4 technical report.

[2]. Akhtar, N., Liu, J., & Mian, A. J. a. p. a. (2017). Defense against Universal Adversarial Perturbations. arXiv.

[3]. Alotaibi, A., & Rassam, M. A. J. F. I. (2023). Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. 15(2), 62.

[4]. Amerirad, B., Cattaneo, M., Kenett, R. S., & Luciano, E. J. R. (2023). Adversarial artificial intelligence in insurance: from an example to some potential remedies. 11(1), 20.

[5]. Azhar, S., Yunus, N. H. M., Yusof, H. N. M., Hussin, N. A., Norhisham, N. Z., & Zainuddin, A. J. J. J. o. E. T. (2023). Big Data Security Issues: A Review. 11(1), 9-17.

[6]. Balamurugan, M. J. I. J. o. S., & Research. (2024). Guardians at Risk: The Challenge of Adversarial Attacks on Authentication Systems and Artificial Intelligence.

[7]. Baluja, S., & Fischer, I. J. a. p. a. (2017). Adversarial transformation networks: Learning to generate adversarial examples.

[8]. Baniecki, H., & Biecek, P. J. I. F. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. 102303.

[9]. Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006). Can machine learning be secure? Proceedings of the 2006 ACM Symposium on Information, computer and communications security,

[10]. Bazaga, A., Gunwant, N., & Micklem, G. J. S. R. (2021). Translating synthetic natural language to database queries with a polyglot deep learning framework. 11(1), 18462.

[11]. Bhuiyan, T., Begum, A., Rahman, S., Hadid, I. J. I. J. o. E., & Technology. (2018). API vulnerabilities: Current status and dependencies. 7(2.3), 9-13.

[12]. Biggio, B., Nelson, B., & Laskov, P. J. a. p. a. (2012). Poisoning attacks against support vector machines.

[13]. Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. J. I. R. (2023). AI governance: themes, knowledge gaps and future agendas. 33(7), 133-167.

[14]. Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. J. A. i. n. i. p. s. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. 30.

[15]. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V.,…systems. (2019). Towards federated learning at scale: System design. 1, 374-388.

[16]. Bountakas, P., Zarras, A., Lekidis, A., & Xenakis, C. J. C. S. R. (2023). Defense strategies for adversarial machine learning: A survey. 49, 100573.

[17]. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B.,…Filar, B. J. a. p. a. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.

[18]. Chen, J., Chen, C., Hu, J., Grundy, J., Wang, Y., Chen, T., & Zheng, Z. (2024). Identifying smart contract security issues in code snippets from stack overflow. Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis,

[19]. Chen, L., Chen, P., & Lin, Z. J. I. A. (2020). Artificial intelligence in education: A review. 8, 75264-75278.
[20]. Chen, T., Liu, J., Xiang, Y., Niu, W., Tong, E., & Han, Z. J. C. (2019). Adversarial attack and defense in reinforcement learning-from AI security view. 2, 1-22.
[21]. Chern, S., Fan, Z., & Liu, A. J. a. p. a. (2024). Combating adversarial attacks with multi-agent debate.
[22]. De Almeida, P. G. R., dos Santos, C. D., Farias, J. S. J. E., & Technology, I. (2021). Artificial intelligence regulation: a framework for governance. 23(3), 505-525.
[23]. Deng, Z., Guo, Y., Han, C., Ma, W., Xiong, J., Wen, S., & Xiang, Y. J. A. C. S. (2025). Ai agents under threat: A survey of key security challenges and future pathways. 57(7), 1-36.
[24]. Di Vaio, A., Palladino, R., Hassan, R., & Escobar, O. J. J. o. B. R. (2020). Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. 121, 283-314.
[25]. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC conference on computer and communications security,
[26]. Gianni, R., Lehtinen, S., & Nieminen, M. J. F. i. C. S. (2022). Governance of responsible AI: From ethical guidelines to cooperative policies. 4, 873437.
[27]. Goodfellow, I. J., Shlens, J., & Szegedy, C. J. a. p. a. (2014). Explaining and harnessing adversarial examples.
[28]. Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. J. I. A. (2023). From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. 11, 80218-80245.
[29]. Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. J. I. J. o. I. M. D. I. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. 4(1), 100232.
[30]. Hazra, S., Majumder, B. P., & Chakrabarty, T. J. a. p. a. (2025). AI Safety Should Prioritize the Future of Work.
[31]. He, W., Li, B., & Song, D. (2018). Decision boundary analysis of adversarial examples. International Conference on Learning Representations,
[32]. Hou, Y., Tamoto, H., & Miyashita, H. (2024). " my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems,
[33]. Hu, B., Zhao, C., Zhang, P., Zhou, Z., Yang, Y., Xu, Z., & Liu, B. J. a. p. a. (2023). Enabling intelligent interactions between an agent and an LLM: A reinforcement learning approach.
[34]. Hu, W., & Tan, Y. (2022). Generating adversarial malware examples for black-box attacks based on GAN. International Conference on Data Mining and Big Data,
[35]. Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y.,…Li, K. J. A. C. S. (2021). Artificial intelligence security: Threats and countermeasures. 55(1), 1-36.
[36]. Huang, C., Chen, S., Zhang, Y., Zhou, W., Rodrigues, J. J., & De Albuquerque, V. H. C. J. I. I. o. T. J. (2021). A robust approach for privacy data protection: IoT security assurance using generative adversarial imitation learning. 9(18), 17089-17097.
[37]. Ijiga, O. M., Idoko, I. P., Ebiega, G. I., Olajide, F. I., Olatunde, T. I., & Ukaegbu, C. J. J. S. T. (2024). Harnessing adversarial machine learning for advanced threat detection: AI-driven strategies in cybersecurity risk assessment and fraud prevention. 11, 001-024.
[38]. Ilahi, I., Usama, M., Qadir, J., Janjua, M. U., Al-Fuqaha, A., Hoang, D. T., & Niyato, D. J. I. T. o. A. I. (2021). Challenges and countermeasures for adversarial attacks on deep reinforcement learning. 3(2), 90-109.
[39]. Izacard, G., & Grave, E. J. a. p. a. (2020). Leveraging passage retrieval with generative models for open domain question answering.
[40]. Jing, H., Wei, W., Zhou, C., & He, X. (2021). An artificial intelligence security framework. Journal of Physics: Conference Series,
[41]. Khan, R., Sarkar, S., Mahata, S. K., & Jose, E. J. a. p. a. (2024). Security Threats in Agentic AI System.

[42]. Kiran, A., & Kumar, S. S. (2023). Synthetic data and its evaluation metrics for machine learning. In Information Systems for Intelligent Systems: Proceedings of ISBM 2022 (pp. 485-494). Springer.
[43]. Lekota, N. F. Governance Considerations of Adversarial Attacks on AI Systems. Proceedings of the International Conference on AI Research,
[44]. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J.,…Zhou, B. J. A. C. S. (2023). Trustworthy AI: From principles to practices. 55(9), 1-46.
[45]. Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., & Song, Y. J. a. p. a. (2023). Multi-step jailbreaking privacy attacks on chatgpt.
[46]. Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X.,…Zheng, Y. J. a. p. a. (2023). Prompt Injection attack against LLM-integrated Applications.
[47]. Liu, Y., Mao, S., Mei, X., Yang, T., & Zhao, X. (2019). Sensitivity of adversarial perturbation in fast gradient sign method. 2019 IEEE symposium series on computational intelligence (SSCI),
[48]. Ludlow, K., Bowman, D. M., Gatof, J., & Bennett, M. G. J. N. (2015). Regulating emerging and future technologies in the present. 9, 151-163.
[49]. Ma, C., Li, J., Wei, K., Liu, B., Ding, M., Yuan, L.,…Poor, H. V. J. P. o. t. I. (2023). Trusted ai in multiagent systems: An overview of privacy and security for distributed learning. 111(9), 1097-1132.
[50]. Majeed, A., & Hwang, S. O. J. I. A. (2023). When AI meets information privacy: The adversarial role of AI in data sharing scenario. 11, 76177-76195.
[51]. Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. 2019 IEEE symposium on security and privacy (SP),
[52]. Mohammed, A. J. A. P. (2025). Artificial Intelligence-Powered Cyber Attacks: Adversarial Machine Learning.
[53]. Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition,
[54]. Moskal, S., Laney, S., Hemberg, E., & O'Reilly, U.-M. J. a. p. a. (2023). Llms killed the script kiddie: How agents supported by large language models change the landscape of network threat testing.
[55]. Muça, M., Kapçiu, R. J. I. J. o. A. C. S., & Applications. (2024). Dimensionality Reduction: A Comparative Review using RBM, KPCA, and t-SNE for Micro-Expressions Recognition. 15(1).
[56]. Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., & Vasilakos, A. J. a. p. a. (2021). Security and privacy for artificial intelligence: Opportunities and challenges.
[57]. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. Proceedings of the 2017 ACM on Asia conference on computer and communications security,
[58]. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. 2016 IEEE symposium on security and privacy (SP),
[59]. Park, D., An, G.-t., Kamyod, C., & Kim, C. G. J. J. o. w. e. (2023). A study on performance improvement of prompt engineering for generative AI with a large language model. 22(8), 1187-1206.
[60]. Pöhler, L., Schrader, V., Ladwein, A., & von Keller, F. J. a. p. a. (2024). A technological perspective on misuse of available AI.
[61]. Polemi, N., Praça, I., Kioskli, K., & Bécue, A. J. F. i. b. D. (2024). Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. 7, 1381163.
[62]. Qiu, S., Liu, Q., Zhou, S., & Wu, C. J. A. S. (2019). Review of artificial intelligence adversarial attack and defense technologies. 9(5), 909.
[63]. Rocher, L., Hendrickx, J. M., & De Montjoye, Y.-A. J. N. c. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. 10(1), 3069.
[64]. Samonas, S., & Coss, D. J. J. o. I. S. S. (2014). The CIA strikes back: Redefining confidentiality, integrity and availability in security. 10(3).
[65]. Sarker, I. H. J. S., & Privacy. (2023). Multi-aspects AI-based modeling and

adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. 6(5), e295.

[66]. Schwinn, L., Dobre, D., Günnemann, S., & Gidel, G. (2023). Adversarial attacks and defenses in large language models: Old and new threats. Proceedings on,

[67]. Shestak, V., & Tsyplakova, A. J. B. L. J. (2023). Countering Cyberattacks on the Energy Sector in the Russian Federation and the USA. 10(4), 35-52.

[68]. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. 2017 IEEE symposium on security and privacy (SP),

[69]. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. J. a. p. a. (2013). Intriguing properties of neural networks.

[70]. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction {APIs}. 25th USENIX security symposium (USENIX Security 16),

[71]. Wirtz, B. W., Weyerer, J. C., & Kehl, I. J. G. i. q. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. 39(4), 101685.

[72]. Yang, W., Bi, X., Lin, Y., Chen, S., Zhou, J., & Sun, X. J. A. i. N. I. P. S. (2024). Watch out for your agents! investigating backdoor threats to llm-based agents. 37, 100938-100964.

[73]. Yee, C. K., & Zolkipli, M. F. J. J. o. I. i. E. (2021). Review on confidentiality, integrity and availability in information security. 8(2), 34-42.

[74]. Zhang, J., Li, C. J. I. t. o. n. n., & systems, l. (2019). Adversarial examples: Opportunities and challenges. 31(7), 2578-2593.