Survey: Cybersecurity Threats In AI To Reveal PII & Confidential Information

CSI 4480
Harsirat Grewal

Abstract

In the rapidly advancing realm of artificial intelligence (AI), cybersecurity attacks have emerged as a formidable challenge, posing significant risks to the integrity, reliability, and functionality of AI systems. This paper surveys the landscape of cybersecurity attacks targeting AI, exploring the varied forms of these attacks and their mitigation strategies. It delves into the sophisticated methodologies employed by attackers which aim to exploit vulnerabilities in AI systems. The necessity of robust mitigation strategies is underscored, highlighting the critical role of research in developing advanced defensive mechanisms to safeguard AI against these evolving threats. This exploration is pivotal for understanding the vulnerabilities inherent in AI systems and for advancing the field towards more secure and resilient AI technologies.

Survey Goal

This survey aims to comprehensively explore the cybersecurity threats targeting Artificial Intelligence (AI) systems and the corresponding mitigation strategies. Initially, we delve into a broad spectrum of general cybersecurity challenges confronting AI, outlining prevalent threats and the current approaches to counter them. This foundational overview sets the stage for an in-depth discussion of escalating attacks that are specifically aimed at revealing PII or confidential information. A significant focus of the analysis is on model inversion attacks – a critical vulnerability in AI – where we dissect their nature and examine state-of-the-art mitigation techniques. A pivotal section of the study is dedicated to attacks on chatbots, particularly those that manipulate these systems into divulging confidential information or generating unethical responses. By investigating recent real-world instances, we illuminate the severity and complexity of these incidents. Subsequently, we evaluate targeted strategies to mitigate such specific threats. A crucial aspect of our survey is the identification of existing research gaps, particularly in areas where mitigation strategies are underdeveloped or absent. Finally, we venture into predictive analysis, hypothesizing potential future attacks on AI systems, and propose theoretical countermeasures. Our goal is not only to present a current landscape of AI cybersecurity but also to anticipate emerging challenges and suggest proactive defenses, thereby contributing to the fortification of AI systems against evolving cyber threats.

Cybersecurity Threats on Artificial Intelligence Systems

1. **Adversarial Attacks**: These involve manipulating an AI model's behavior by introducing carefully crafted input data, causing the model to make incorrect predictions.
   Techniques used:
   - Gradient-Based Attacks – Uses the gradient of model's loss function to determine how to modify the input to produce an incorrect input. Example – Fast Gradient

Sign Method. FGSM takes an input image and applies a small perturbation in the direction that increases the loss the most. This perturbation is calculated using the sign of the gradient of the loss function with respect to the input image. The result is a new image that appears almost identical to the original but is classified differently by the neural network.

- Optimization-Based Attacks – Frames the creation of adversarial examples as an optimization problem, often seeking the smallest change to the input that causes a misclassification. Example – Carlini & Wagner Attack. It formulates the attack as an optimization problem, where the goal is to find the smallest possible perturbation to the input that causes the target model to misclassify it. This method focuses on adjusting the input in a way that minimizes the difference from the original input while ensuring that the manipulated input is classified as a different class by the model.
- Decision-Based Attacks – These attacks iteratively modify the input based on model's output decisions. Example – The Boundary Attack. It operates by iteratively modifying an input image until the model misclassifies it, while keeping the changes minimal and imperceptible to the human eye. The attack starts with a large perturbation and reduces it gradually, staying close to the decision boundary of the classifier.
- Poisoning Attacks – Attackers inject malicious data into the model's training set to corrupt its learning process which leads to incorrect model behavior. For example, if an AI is trained to identify horses using images of brown horses and black-and-white cows, inserting a brown cow could lead the AI to misclassify it as a horse.
- Query Attack – These attacks repeatedly query a model to gather information about its behavior and use that information to craft adversarial attacks. For example, Oracle attack where attackers aimed to understand how the model works by analyzing input-output pairs provided by contradictory examples. Through this process, they could detect the model's sensitivity to changes in input data, potentially revealing the model's vulnerabilities to evasion attacks and infringing on intellectual property or the confidentiality of training data.
- Evasion Attacks – These are designed to evade detection by AI systems, such as fooling an automated image recognition system. Countermeasures include training models on adversarial examples and continuously updating the model with new data to recognize evasion tactics.

Mitigation Strategies:

- Input Validation – Before processing the inputs, they can be sanitized to remove potential adversarial perturbations. For example, in an image recognition system, this could involve ensuring that the input images conform to expected size, format, and range of values. Any input that doesn't meet these criteria can be rejected or normalized.

- Model Regularization – During the training, techniques such as weight decay or adding noise to input can make the model less sensitive to small perturbations in the input. For example, in a neural network, dropouts randomly deactivate a subset of neurons during each training iteration. This prevents the network from becoming overly dependent on specific features of the training data, thus making it more generalizable and less sensitive to small perturbations or adversarial inputs.
- Enhanced Adversarial Training – This involves training the model on detecting adversarial attacks. For example, in a facial recognition system, this could mean not only training the model with standard images of faces but also with modified images that have been altered using various adversarial attack techniques.
- Anomaly Detection Mechanism – This involves implementation of a system that monitors the AI's output for unusual or unexpected patterns. For example, in a network security system, an anomaly detector could analyze traffic patterns using the AI model. If the system observes deviations from normal behavior, such as unexpected data flow or unrecognized access patterns, it could flag these as potential security threats.
- Continuous Model Performance Training – This involves regularly updating the AI model with new data and monitoring its performance over time. For example, in a fraud detection AI system, the model could be continuously updated with new transaction data, including instances of detected fraud. By regularly retraining the model with the latest data, it stays up to date with evolving fraud patterns, which helps maintain its effectiveness in identifying fraudulent activities even as attackers adapt their methods.

2. **Model Inversion Attacks**: These attacks aim to extract sensitive information from a machine learning model, essentially reverse engineering the model to reveal private details about the training data.
   Techniques Used:
   - Reconstruction of Inputs – Attacks involving the reconstruction of inputs are an attack on a machine learning model used for text analysis or natural language processing. The attack's goal is to infer sensitive or private information that was part of the training data. The attacker submits various text inputs to the model and observes the outputs. They might input texts with specific themes or keywords and note the model's response in terms of sentiment classification, topic categorization, or other metadata. By analyzing the model's responses, the attacker can start to infer characteristics of the training data. If certain types of input consistently yield high confidence in sentiment classification, it might suggest that the model was extensively trained on similar texts. Using this knowledge, the attacker might attempt to reconstruct or approximate the original

training data, especially if it includes sensitive information. This could involve creating text that aligns closely with the inferred themes, styles, or sentiments of the training data. Such an attack could reveal private or sensitive information, particularly if the model was trained on confidential texts (like personal messages or proprietary documents).

- Exploiting Confidence Scores – Analyzing the model's confidence scores to infer properties about the training data or the model itself. For example, membership inference attack where the adversary aims to determine whether a particular data point was used in the training set of a machine learning model. The attacker can use the confidence scores (probability estimates) provided by the model for its predictions. If a model gives a very high confidence score for a specific input, it might indicate that the model has 'seen' this data or similar data during training. Conversely, lower confidence scores might suggest that the data was not part of the training set. By systematically querying the model with various inputs and observing the confidence scores, attackers can infer information about the data used to train the model, potentially leading to privacy breaches.
- Learning-Based Model Inversion – These attacks involve training a separate model to reconstruct private input data from the predictions made by the target model. They aim to achieve accurate inversion of individual private data rather than class-representative inversions. A multi-modal transposed CNN architecture has been developed to reconstruct private data by introducing explanations from the target model. This method improves both the inversion performance and the classification accuracy of the reconstructed data.
- Knowledge-Enriched Distributional Model Inversion Attack– In this attack, the attacker, without direct access to a model's private training data, uses publicly available data to understand general patterns related to the model's function. For example, in a medical prediction model, the attacker uses public health datasets to enhance their understanding of disease patterns. This knowledge helps them craft inputs to manipulate the model, aiming to infer or reconstruct aspects of the private training data, like specific health conditions.

Mitigation strategies:

- Limiting Model Exposure – Reduce the amount of information the model discloses. For example, instead of providing precise confidence scores, the model can return more general information like class labels.
- Limit Model Access – Restrict the number of requests a user can make to the model within a certain timeframe. Implement strict access controls, ensuring that only authenticated and authorized users can access the model. This could involve using OAuth tokens, API keys, or similar authentication methods.
- Query Filtering – Introduce filters that scrutinize the nature of the queries being made. If a query seems to be probing the model in a way that's indicative of an

attack (e.g., systematically varying inputs to map the model's response landscape), the system can flag or block such queries.

- Differential Privacy – Implementing differential privacy techniques in training can help protect the model against attacks aimed at reconstructing training data. It adds noise to the training process or the model's output, making it difficult to infer specifics about individual data points. For example, imagine a company is developing an AI model to recommend movies based on user viewing habits. To protect user privacy, during the training phase, they add random noise to the data. This noise makes it hard for anyone using the model to accurately deduce an individual's specific viewing history, even if they try model inversion attacks.
- Generalization – Improving model generalization through techniques like dropout, L1/L2 regularization, and data augmentation can prevent overfitting to the training data, making it harder to reconstruct specific inputs. For example, a company training an AI model on customer preferences enhances data privacy by incorporating a wide range of customer behaviors in the training set and applying regularization techniques. This broadens the model's learning, preventing it from being too specific to the training data.
- Homomorphic Encryption – For scenarios where data privacy is critical, applying homomorphic encryption allows computations on encrypted data without needing to decrypt it, significantly enhancing privacy.
- Data Governance and Policy – Implementing strict data governance policies and ensuring legal and ethical compliance can be a deterrent against misuse of model outputs.

3. **Other Attacks**: These are some few other attack types:
    - **Distributed Denial of Service (DDoS) Attacks**: Aimed at overwhelming AI systems with excessive data or requests, these can be mitigated by deploying network security solutions that can detect and block such traffic. Mitigation strategies include Rate limiting, traffic analysis, scalable infrastructure, and CDN (Content Delivery Networks).
    - **Transfer Attacks**: In these, adversarial examples crafted for one model are used to attack another model. Defenses include using model-specific features that make transferability of attacks more difficult. Mitigation strategies include model ensembling and continuous monitoring.
    - **Misuse of AI Assistants**: This includes manipulating AI assistants to perform unintended actions. Mitigation involves designing AI systems with strict command validation processes and ethical usage guidelines. Mitigation strategies include command validation, user authentication, and behavioral analysis.
    - **Data Manipulation Attacks**: These involve altering data used by AI systems to influence outcomes, which can be countered by implementing strong access

controls and data integrity checks. Mitigation strategies include strong access control, data integrity checks, and anomaly detection.

Personal Identifiable Information & Confidential Information

The most significant attacks on AI systems responsible for revealing confidential or Personally Identifiable Information (PII) are primarily focused on exploiting vulnerabilities in machine learning and deep learning models. These attacks are often carried out through adversarial methods or by manipulating the model's output to extract sensitive data.
A comprehensive review of the literature reveals various systematic studies addressing security threats and defensive techniques in machine learning, with a specific focus on adversarial attacks. These attacks target deep learning models, particularly in computer vision and supervised learning (classification), which have been extensively studied for their vulnerabilities to such attacks. These reviews highlight the ongoing need for robust defense mechanisms against adversarial attacks in deep learning, particularly where privacy and confidentiality are of utmost concern.
Another critical area of concern is the leakage of PII in language models. Research has shown that even when models are trained with defensive measures like differential privacy, there is still a risk of PII leakage. This is particularly concerning because language models, such as those used for text analysis or natural language processing, can inadvertently reveal sensitive information embedded in their training data. The study of PII leakage in undefended and differentially private models shows the complexity of the issue, emphasizing the need for more effective privacy-preserving techniques in AI.

Model Inversion (MI) attacks in the context of machine learning and AI are a type of cybersecurity threat where an attacker aims to reconstruct sensitive information from a machine learning model's output. This is particularly relevant when the model has been trained on datasets containing Personally Identifiable Information (PII). Here's how these attacks typically work:

1. **Targeting the Model's Outputs**: MI attacks exploit the relationship between a model's inputs and its outputs. The attacker uses the outputs (such as predictions or confidence scores) of the machine learning model to make inferences about the original input data.
2. **Reconstruction of Inputs**: By systematically querying the model with various inputs and analyzing the outputs, attackers can deduce patterns or features of the original training data. For instance, if a model consistently provides high confidence scores for specific types of data, it may indicate that the model was extensively trained on similar data.
3. **Exploiting Model Behaviors**: Attackers can use the model's behavior to infer private information. For example, in a facial recognition system, by observing how the model responds to different images, an attacker might reconstruct the features of faces the model was trained on.

4. **Learning-Based MI Attacks**: In some cases, attackers may train a separate model to infer private input data based on the predictions made by the target model. This approach can be more sophisticated and might involve using advanced machine learning techniques to achieve more accurate reconstructions.
5. **Use of Public Information for Enhancement**: Attackers might also use publicly available data to enhance their understanding of the kind of information the target model was trained on. This knowledge assists them in crafting inputs that are more likely to yield useful information about private training data.
6. **Extracting Direct Identifiers**: MI attacks can be particularly effective in extracting direct identifiers like names, phone numbers, or addresses. In some scenarios, even quasi-identifiers (like a combination of non-unique attributes) can be reconstructed, potentially leading to re-identification of individuals in anonymized datasets.

Mitigating Strategies & Gaps

Recent research on mitigating Model Inversion (MI) attacks has proposed innovative defense mechanisms. One notable approach is the Mutual Information Regularization based Defense (MID). This method aims to limit the information about the model input contained in the prediction, thereby reducing the adversary's ability to infer private training attributes from the model prediction. MID is model-agnostic and applicable to various models like linear regression, decision trees, and neural networks. It has been observed that MID leads to state-of-the-art performance against a variety of MI attacks, across different target models and datasets. However, this approach, while promising, has not fully addressed the challenge of balancing utility and privacy, a common issue in most defense mechanisms against MI attacks. Further research is needed to refine these strategies and achieve an optimal utility-privacy tradeoff.

Additionally, two primary methods have also emerged: using adversarial examples to enhance the robustness of models against MI attacks and employing prediction purification frameworks.

1. **Using Adversarial Examples**: This approach involves integrating adversarial examples into the training process to boost the robustness of models against MI attacks. While this method can enhance attack accuracy in a black box setting, it often requires detailed knowledge of the target model's parameters and architecture, which may not always be practical or feasible.
2. **Prediction Purification Framework**: Another approach is a unified purification framework that defends against both MI and membership inference attacks by reducing the dispersion of confidence score vectors. This method has been demonstrated to be effective in mitigating both types of attacks simultaneously. However, the challenge here lies in balancing the trade-off between maintaining model performance and ensuring adequate privacy protection.

Despite these advancements, there are limitations and gaps in the current research. For instance, many of these methods are designed for specific scenarios and may not be universally applicable across different models and datasets. Additionally, the effectiveness of these strategies can vary depending on the complexity of the model and the nature of the data it processes. More research is needed to develop universally applicable solutions that offer a better balance between model utility and privacy protection.

Chatbot Manipulation

Chatbot manipulation attacks, particularly prompt injection attacks, have emerged as significant cybersecurity risks. These attacks involve manipulating the language models underpinning chatbots to induce unintended responses, potentially leading to serious real-world consequences.

How Chatbot Manipulation Attacks Work:

1. **Exploitation of Chatbot AI**: Attackers manipulate the prompts or inputs given to a chatbot, which operates on AI and large language models (LLMs), to generate specific responses or actions.
2. **Prompt Injection**: This technique involves creating input or prompts that are designed to manipulate the behavior of the chatbot. For example, by inputting unfamiliar statements or exploiting word combinations, an attacker can override the chatbot's original script, causing it to perform unintended actions.
3. **Potential Consequences**: These manipulations can lead to a range of outcomes, from generating offensive content to disclosing confidential information, or even executing unauthorized transactions and data breaches.

Recent Cases:

- In February 2023, an experiment highlighted the vulnerability of LLMs, where a chatbot was manipulated to impersonate a scammer, soliciting sensitive bank account details from users. This instance demonstrated a new threat type: indirect prompt injection attacks, which could compromise sensitive data.
- The UK's National Cyber Security Centre (NCSC) has raised alerts about the increasing vulnerability of chatbots to these types of attacks. The NCSC emphasizes that while prompt injection attacks are challenging to detect and mitigate, a holistic system design considering machine learning risks can help prevent vulnerabilities exploitation.
- In June 2023, NVIDIA's AI software was reportedly manipulated to leak sensitive data. Researchers exploited vulnerabilities in the company's large language models (LLMs), enabling the extraction of embedded training data. The method involved feeding crafted inputs to the model, which in turn generated outputs containing the sensitive information. This incident raised significant concerns about the security of AI systems and the

potential for data leakage, highlighting the need for enhanced safeguards in AI and machine learning applications.

Mitigation Strategies:

1. **Rules-Based Systems**: Implementing systems alongside the machine learning model to counteract potentially damaging actions, even when prompted by attackers.
2. **Security in Design**: Prioritizing security in the application development process, especially in the integration of machine learning algorithms.
3. **Vigilance in AI Deployment**: Being cautious in deploying AI solutions, especially in sensitive areas like customer transactions.
4. **Awareness and Training**: Ensuring that users are aware of the potential risks associated with chatbot interactions and training them to recognize suspicious activities.

Recent research in mitigating chatbot manipulation attacks focuses on understanding and addressing vulnerabilities in chatbots, especially those based on large language models like GPT-3. The key challenges involve ensuring the chatbots respond appropriately to user inputs and preventing them from being manipulated to reveal sensitive information or perform unwanted actions.

One approach highlighted in the research includes developing more sophisticated natural language processing algorithms that better understand the context and intent of user queries. This can help in identifying and filtering out malicious or manipulative prompts. Another strategy involves implementing robust security measures and protocols within the chatbot framework to prevent unauthorized access and data breaches.

Despite these efforts, there are still gaps in research, particularly in developing universally effective solutions that balance chatbot functionality with security. The rapidly evolving nature of AI and language models also means that new vulnerabilities may emerge, requiring continuous research and development in this area.

<u>Predictive Analysis</u>

1. **Evolving Nature of Adversarial Attacks**: As AI continues to advance, adversarial attacks are expected to become more sophisticated. Attackers may develop new methods that are harder to detect and counter, such as more advanced gradient-based, optimization-based, and decision-based attacks. This evolution will necessitate the continuous development of advanced defensive mechanisms.
2. **Increasing Risks of Model Inversion Attacks**: Model inversion attacks, particularly those targeting the reconstruction of inputs and exploitation of confidence scores, are likely to become more prevalent. As AI models become more integrated into various sectors, the incentive for attackers to extract sensitive information increases.
3. **Chatbot Manipulation Tactics**: The risk of chatbot manipulation, including prompt injection attacks, is expected to rise as chatbots become more ubiquitous in services like

online banking and shopping. Attackers might develop more ingenious ways to manipulate chatbots into revealing confidential information or performing unintended actions.

4. **Enhanced Mitigation Strategies**: In response to these threats, future research will likely focus on developing more robust and sophisticated mitigation strategies. This could include advanced natural language processing algorithms to better understand and filter out malicious prompts, as well as more comprehensive security protocols within AI systems.
5. **Research Gaps and Challenges**: Despite advancements in mitigation strategies, there will be ongoing challenges in balancing the utility of AI systems with privacy and security. Research gaps might include the need for universally applicable security solutions and methods to adapt to the rapidly evolving AI landscape.
6. **Proactive and Predictive Security Measures**: Future efforts in AI cybersecurity may shift towards more proactive and predictive measures. This includes using AI itself to predict potential vulnerabilities and attacks, thereby allowing for preemptive action to strengthen AI systems against future threats.
7. **Legal and Ethical Considerations**: As AI becomes more entwined with daily life, the legal and ethical implications of AI security will become more complex. This might lead to new regulations and standards governing the development and deployment of AI systems.

Mitigation Strategies and Future Recommendations

1. **Proactive Cybersecurity Measures**: The NCSC (National Cybersecurity Centre) emphasizes the importance of a proactive approach to cybersecurity. This includes implementing strategies like rules-based systems alongside AI models to counteract damaging actions prompted by chatbot manipulation.
2. **Detection and Prevention Challenges**: Detecting and mitigating prompt injection and data poisoning attacks remain complex and challenging. Continuous research and development of more robust security measures are necessary to safeguard against these threats.
3. **Understanding Attacker Techniques**: Developing and maintaining security protocols involves understanding the methods attackers use and integrating security considerations into the design process of AI applications.

Research Gaps

Despite advancements, there are notable gaps in current research:

- **Universal Security Solutions**: Developing effective, universally applicable security measures that can adapt to the evolving nature of AI chatbots and the varying contexts in which they are deployed.
- **Balancing Functionality and Security**: Ensuring that security measures do not overly restrict the functionality and user experience of chatbots.

- **Continuous Evolution of Threats**: Keeping pace with the rapidly evolving techniques used by attackers in manipulating these AI systems.

Conclusion

The survey underscores the dynamic and evolving nature of cybersecurity threats in the AI landscape. As AI technologies become more integrated into various sectors, the sophistication and frequency of attacks targeting these systems are likely to increase. The survey identifies adversarial attacks, model inversion attacks, and chatbot manipulations as significant threats that can compromise the integrity and confidentiality of AI systems.

Adversarial attacks, including gradient-based, optimization-based, and decision-based methods, pose a challenge to the reliability of AI models. Model inversion attacks, which focus on extracting sensitive information from AI systems, are a growing concern, especially in applications involving PII. Similarly, chatbot manipulation, particularly through prompt injection techniques, presents a risk in terms of unauthorized access to confidential data and the generation of unethical responses.

The paper emphasizes the importance of robust and adaptive mitigation strategies. These include input validation, model regularization, enhanced adversarial training, anomaly detection mechanisms, and continuous model performance training. However, it also highlights that current mitigation strategies have limitations and might not be universally applicable across different AI models and contexts.

The survey points out the need for continuous research and development of advanced defensive mechanisms to safeguard AI systems. This includes developing more sophisticated natural language processing algorithms, implementing comprehensive security protocols, and adopting proactive and predictive security measures. Additionally, the legal and ethical implications of AI security, as AI becomes more pervasive in daily life, will necessitate new regulations and standards.

References

https://www.spiceworks.com/it-security/cyber-risk-management/news/china-backed-hackers-hit-us-infra/

https://dwfgroup.com/en/news-and-insights/insights/2023/9/cyber-security-risks-associated-with-ai-chatbots-being-manipulated-by-bad-actors

https://securityintelligence.com/articles/data-poisoning-ai-and-machine-learning/

https://www.riskinsight-wavestone.com/en/2023/06/attacking-ai-a-real-life-example/

https://www.softvire.co.nz/chatbot-attacks-what-are-they-and-how-to-prevent-them/

https://ar5iv.labs.arxiv.org/html/2306.09255

https://www.robustintelligence.com/blog-posts/nemo-guardrails-early-look-what-you-need-to-know-before-deploying-part-2

S. Shahriar, S. Allana, S. M. Hazratifard and R. Dara, "A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle," in IEEE Access, vol. 11, pp. 61829-61854, 2023, doi: 10.1109/ACCESS.2023.3287195