Research Article

Cognitive Honeypots: Leveraging Logical Contradictions to Detect and Analyze Adversarial AI Behavior

Kush Janani[1]

1. College of Computing and Digital Media, DePaul University, United States

Traditional honeypots are designed to attract human attackers by mimicking vulnerable systems, but as artificial intelligence (AI) becomes increasingly sophisticated, new security paradigms are needed. This paper introduces the concept of ”cognitive honeypots” - a novel approach that leverages logical contradictions to detect and analyze adversarial AI behavior. Unlike conventional security measures that focus on patching vulnerabilities or making them difficult to exploit, cognitive honeypots intentionally present logical inconsistencies designed to attract adversarial AI systems. By analyzing how AI attackers might engage with these cognitive traps, defenders could discover new classes of adversarial reasoning, biases, and vulnerabilities embedded in model logic. We present a theoretical framework for cognitive honeypots, propose an implementation architecture, and discuss their potential effectiveness against various types of adversarial AI. Our analysis suggests that cognitive honeypots could enable unprecedented proactive security measures against emerging AI threats and contribute to the development of more robust AI systems.

Corresponding author: Kush Janani, kjanani@depaul.edu

I. Introduction

The rapid advancement of artificial intelligence (AI) has transformed numerous domains, from healthcare and finance to cybersecurity and autonomous systems. However, this progress has been accompanied by the emergence of sophisticated adversarial techniques designed to manipulate, deceive, or exploit AI systems[1]. As AI becomes more integrated into critical infrastructure and decision-making processes, the security implications of these vulnerabilities grow increasingly significant.

 doi.org/10.32388/DE644X

Traditional cybersecurity approaches have relied on honeypots—systems designed to mimic vulnerable targets—to attract, detect, and analyze human attackers[2]. These systems have proven valuable for understanding attack patterns, developing defensive strategies, and gathering intelligence on threat actors. However, as the threat landscape evolves to include adversarial AI, conventional honeypot approaches must adapt to address the unique characteristics of AI-based attacks.

A. Motivation and Research Gap

Current defensive approaches against adversarial AI typically focus on either patching weaknesses in trained models or making it difficult to compute adversarial examples that exploit them[3][4]. While these methods have shown some success, they often engage in an endless cat-and-mouse game with attackers, where each new defense is eventually circumvented by more sophisticated attacks. Furthermore, these approaches provide limited insight into the reasoning patterns and decision processes of adversarial AI systems.

A significant research gap exists in proactive defense mechanisms that can not only detect adversarial AI but also reveal fundamental vulnerabilities in AI reasoning. Understanding how adversarial AI systems approach problem-solving, identify exploits, and make decisions could lead to more robust defensive strategies and the development of inherently more secure AI architectures.

B. Proposed Approach: Cognitive Honeypots

In this paper, we introduce the concept of ”cognitive honeypots”—a novel approach that applies the honeypot paradigm to the domain of AI security. Unlike traditional honeypots that expose technical vulnerabilities, cognitive honeypots intentionally present logical contradictions designed to attract and trap adversarial AI systems. These contradictions serve as cognitive traps that exploit the reasoning patterns and optimization processes commonly used by AI systems.

By analyzing how adversarial AI might engage with these cognitive traps, defenders could:

- Detect adversarial AI behavior with high accuracy
- Discover new classes of adversarial reasoning and attack patterns
- Identify biases and vulnerabilities embedded in model logic
- Develop proactive security measures against future AI threats

Our approach builds upon recent work in adversarial machine learning, particularly the concept of "trapdoors" introduced by Shan et al.[1], which demonstrated the effectiveness of intentionally injected weaknesses in neural networks for detecting adversarial examples. We extend this concept beyond classification models to address a broader range of AI systems and reasoning capabilities.

C. Research Contributions

The primary contributions of this paper are:

1. Introduction and formalization of the cognitive honeypot concept for AI security
2. Development of a taxonomy of logical contradictions potentially effective against different AI architectures
3. Proposal of a framework for analyzing adversarial reasoning patterns through cognitive honeypots
4. Theoretical analysis of cognitive honeypot effectiveness against various types of adversarial AI
5. Design principles and implementation guidelines for deploying cognitive honeypots in real-world systems

D. Paper Organization

The remainder of this paper is organized as follows: Section II provides background information on traditional honeypots, adversarial attacks on AI systems, and related work in AI security. Section III presents our conceptual framework for cognitive honeypots, including theoretical foundations and key components. Section IV describes our methodology, including the design principles, implementation architecture, and evaluation metrics. Section V presents a theoretical analysis of the potential effectiveness of cognitive honeypots against various adversarial AI systems. Section VI discusses the implications of our findings, limitations of the approach, and potential countermeasures. Section VII outlines directions for future work, and Section VIII concludes the paper.

II. Background and Related Work

This section provides essential background information on traditional honeypots, adversarial attacks on AI systems, and current defensive approaches. We also review related work in deception techniques and logical reasoning in AI systems.

A. Traditional Honeypots in Cybersecurity

Honeypots have been a valuable tool in cybersecurity for decades, serving as decoy systems designed to attract attackers and study their behavior[2]. These systems typically mimic vulnerable targets, such as servers, databases, or networks, while incorporating monitoring capabilities to record and analyze attack patterns. Honeypots are generally classified based on their level of interaction:

- **Low-interaction honeypots** simulate only basic services and have limited functionality, making them easier to deploy but providing less detailed information about attacks.
- **Medium-interaction honeypots** offer more realistic services and can capture more sophisticated attack behaviors without exposing a complete operating system.
- **High-interaction honeypots** provide fully functional systems that allow attackers to interact with real operating systems and applications, enabling the collection of comprehensive information about attack techniques.

The primary benefits of honeypots include their ability to detect new attack vectors, gather intelligence on threat actors, and divert attackers from legitimate targets. However, traditional honeypots are primarily designed to attract and analyze human attackers or automated tools created by humans, rather than adversarial AI systems.

B. Adversarial Attacks on AI Systems

Adversarial attacks on AI systems involve manipulating input data to cause the AI to make errors or behave in unintended ways[5]. These attacks exploit vulnerabilities in the learning algorithms, decision boundaries, or optimization processes used by AI systems. Common types of adversarial attacks include:

- **Evasion attacks** modify inputs at test time to cause misclassification or incorrect outputs, such as adding imperceptible perturbations to images to fool image recognition systems[6].
- **Poisoning attacks** contaminate training data to introduce backdoors or vulnerabilities that can be exploited later[7].
- **Model extraction attacks** attempt to steal or replicate a model's functionality through query access[8].
- **Membership inference attacks** determine whether specific data was used to train a model, potentially revealing sensitive information[9].

- **Logic-based attacks** exploit logical inconsistencies in AI reasoning to generate adversarial examples[10].

These attacks can be further categorized based on the attacker's knowledge (white-box, black-box, or gray-box), the attack's goal (targeted or untargeted), and the attack's scope (global or local). As AI systems become more complex and widespread, the sophistication and impact of adversarial attacks continue to grow.

C. Current Defensive Approaches

Numerous approaches have been proposed to defend AI systems against adversarial attacks. These can be broadly categorized into the following strategies:

- **Adversarial training** incorporates adversarial examples into the training process to make models more robust[3].
- **Gradient masking/obfuscation** techniques modify the model to hide gradient information, making it harder to generate adversarial examples[11].
- **Input preprocessing/purification** methods attempt to remove adversarial perturbations before feeding data to the model[12].
- **Certified robustness** approaches provide mathematical guarantees about a model's behavior under certain types of perturbations[13].
- **Ensemble methods** combine multiple models to improve robustness through diversity[14].
- **Detection-based approaches** aim to identify adversarial examples before they cause harm[15].

While these approaches have shown some success, they often focus on specific types of attacks or models and may not generalize well to new attack vectors. Additionally, many defenses have been shown to be vulnerable to adaptive attacks specifically designed to circumvent them[16].

D. Honeypots for Neural Networks

Recent work by Shan et al.[1] introduced the concept of "trapdoors" as honeypots for neural network models. This approach intentionally injects specific weaknesses into the classification manifold that attract adversarial attacks. By comparing the neuron activation signatures of inputs to those of known trapdoors, the system can detect adversarial examples with high accuracy.

The trapdoor approach demonstrated effectiveness against various state-of-the-art attacks, including Projected Gradient Descent (PGD), Carlini-Wagner (CW), Elastic Net, and Backward Pass Differentiable Approximation (BPDA). Importantly, the presence of trapdoors had negligible impact on normal classification performance, making this a practical defense mechanism.

While the trapdoor concept provides a valuable foundation for our work, it primarily focuses on classification models and does not address the broader range of AI systems and reasoning capabilities that cognitive honeypots aim to protect.

E. Logical Reasoning and Contradictions in AI

Logical reasoning remains a significant challenge for AI systems, despite recent advances in large language models and reasoning frameworks. Research by Nakamura et al.[10] demonstrated that even state-of-the-art natural language inference (NLI) models are vulnerable to attacks based on logical contradictions. Their LogicAttack approach leverages inference rules from propositional logic, such as Modus Tollens and Bidirectional Dilemma, to generate adversarial examples that expose logical inconsistencies in AI reasoning.

These findings highlight the potential for using logical contradictions as a basis for cognitive honeypots. By designing traps that exploit known weaknesses in AI reasoning, defenders can create effective mechanisms for detecting and analyzing adversarial AI behavior.

F. Deception Techniques in Cybersecurity

Deception has long been recognized as a powerful tool in cybersecurity, with applications ranging from honeypots and honeytokens to moving target defense and disinformation campaigns[17]. These techniques leverage the asymmetry of knowledge between defenders and attackers, creating uncertainty and increasing the cost of successful attacks.

Recent work has begun to explore the application of deception techniques specifically to AI security. For example, research on adversarial examples has shown that subtle modifications to input data can cause AI systems to make confident but incorrect predictions[5]. This suggests that AI systems may be particularly vulnerable to certain forms of deception, especially those that exploit the statistical patterns and optimization processes they rely on.

Cognitive honeypots build upon this foundation by designing deceptive elements specifically tailored to the reasoning patterns and vulnerabilities of AI systems. By understanding how adversarial AI

approaches problem-solving and decision-making, defenders can create more effective traps and gain valuable insights into attack methodologies.

III. Cognitive Honeypots: Conceptual Framework

This section presents our conceptual framework for cognitive honeypots, including the theoretical foundations, key components, and operational principles. We define cognitive honeypots as intentionally designed logical contradictions or reasoning traps embedded within AI systems to attract, detect, and analyze adversarial AI behavior.

A. Theoretical Foundations

The cognitive honeypot approach is grounded in several theoretical domains, including deception theory, adversarial machine learning, and cognitive science. These foundations provide the basis for understanding how and why cognitive traps can be effective against adversarial AI systems.

1. Deception Theory in AI Security

Deception has long been recognized as an asymmetric advantage in security contexts, where defenders can manipulate attackers' perceptions and decision-making processes[17]. In the context of AI security, deception can be particularly effective due to several characteristics of AI systems:

- **Optimization bias:** AI systems typically rely on optimization algorithms that follow gradient paths toward apparent solutions, making them susceptible to deliberately crafted attractive regions in the solution space.
- **Pattern recognition limitations:** While AI excels at identifying patterns, it often struggles with distinguishing genuine patterns from deceptive ones, especially when the deception aligns with statistical regularities in the training data.
- **Context sensitivity:** AI systems may fail to recognize when they are operating in adversarial contexts, making them vulnerable to environments specifically designed to elicit particular behaviors.
- **Reasoning constraints:** Current AI systems exhibit limitations in logical reasoning, causal inference, and handling contradictions, creating exploitable vulnerabilities.

These characteristics create opportunities for defensive deception through cognitive honeypots that can attract adversarial AI into revealing its presence and techniques.

2. Cognitive Biases in AI Systems

Research in cognitive science has identified numerous biases that affect human reasoning, many of which have analogues in AI systems. These biases can be leveraged in the design of cognitive honeypots:

- **Anchoring bias:** AI systems may over-rely on initial information or features, which can be exploited by presenting misleading initial conditions.
- **Confirmation bias:** AI tends to interpret new information in ways that confirm existing patterns or predictions, making it susceptible to traps that reinforce incorrect assumptions.
- **Availability heuristic:** AI may prioritize readily available features or patterns, which can be manipulated by making certain paths or solutions artificially prominent.
- **Framing effects:** The way problems are presented can significantly impact AI decision-making, allowing defenders to guide adversarial AI toward specific traps.

By understanding these biases, we can design cognitive honeypots that effectively attract and trap adversarial AI systems while minimizing impact on legitimate operations.

B. Key Components of Cognitive Honeypots

Our framework identifies four essential components of effective cognitive honeypots:

1. Logical Contradiction Design

The core of a cognitive honeypot is the logical contradiction or reasoning trap that serves as the attractor for adversarial AI. These contradictions must be:

- **Attractive to adversarial optimization:** The contradiction should create regions in the feature or decision space that appear promising to optimization algorithms commonly used in adversarial attacks.
- **Distinguishable from normal operations:** The contradiction should be designed so that legitimate inputs and operations are unlikely to trigger it, minimizing false positives.
- **Robust against detection:** The contradiction should not be easily identifiable as a trap by sophisticated adversaries attempting to avoid detection.
- **Informative when triggered:** When an adversarial AI engages with the contradiction, the interaction should reveal useful information about the adversary's techniques and capabilities.

Based on our analysis of current AI vulnerabilities, we propose a taxonomy of effective logical contradictions for cognitive honeypots, as shown in Table I.

Category	Contradiction Type	Effectiveness Against
Feature-space Contradictions	Classification Manifold Trapdoors	Gradient-based attacks, Optimization-based attacks
Logical Reasoning Contradictions	Propositional Logic Violations	Language models, Reasoning systems
Temporal Contradictions	Sequential Inconsistencies	Reinforcement learning agents, Planning systems
Contextual Contradictions	World Knowledge Violations	Multimodal systems, Knowledge-based AI

Table I. Taxonomy of Logical Contradictions for Cognitive Honeypots

2. Detection Mechanisms

Cognitive honeypots require effective mechanisms to detect when adversarial AI has engaged with the embedded contradictions. Our framework incorporates multiple detection approaches:

- **Neuron activation signature analysis:** Comparing activation patterns of inputs against known signatures of contradiction engagement, similar to the approach used in trapdoor detection[1].
- **Behavioral analysis:** Monitoring response patterns, decision paths, or output characteristics that indicate interaction with contradictions.
- **Timing analysis:** Measuring computational time and resource utilization, which may differ significantly when adversarial AI encounters and attempts to resolve contradictions.
- **Probabilistic assessment:** Using statistical models to evaluate the likelihood that observed behaviors result from adversarial engagement with contradictions.

These detection mechanisms can be combined using ensemble methods to improve accuracy and reduce false positives.

3. Analysis Framework

When adversarial AI engages with a cognitive honeypot, the resulting interaction provides valuable data for analysis. Our framework includes analytical components for:

- **Attack classification:** Identifying the type and characteristics of the adversarial technique being employed.
- **Reasoning pattern extraction:** Analyzing how the adversarial AI approaches and attempts to resolve contradictions, revealing its underlying reasoning mechanisms.
- **Vulnerability mapping:** Identifying specific weaknesses in the adversarial AI that could be exploited for defensive purposes.
- **Threat assessment:** Evaluating the sophistication, capabilities, and potential impact of the adversarial AI.

This analysis provides defenders with actionable intelligence for improving security measures and developing countermeasures against adversarial AI.

4. Response Integration

Cognitive honeypots must be integrated with broader security frameworks to enable appropriate responses to detected adversarial activity. Our framework supports:

- **Alert generation:** Notifying security systems and personnel of detected adversarial activity.
- **Dynamic defense adaptation:** Automatically adjusting security parameters based on detected attack patterns.
- **Deception escalation:** Activating additional deceptive measures to further engage and analyze sophisticated adversaries.
- **Attack mitigation:** Implementing countermeasures to neutralize or redirect adversarial activities.

The integration of cognitive honeypots with existing security infrastructure ensures that detections lead to effective protective actions.

C. Operational Principles

The effective deployment of cognitive honeypots is guided by several operational principles:

1. Minimal Interference

Cognitive honeypots should have minimal impact on the normal operation of the protected AI system. This requires careful design of contradictions that are unlikely to be triggered by legitimate inputs or processes. Our approach achieves this through:

- **Targeted placement:** Positioning contradictions in regions of the feature or decision space that are rarely accessed during normal operation.
- **Contextual activation:** Designing contradictions that are only fully manifested in specific contexts associated with adversarial activity.
- **Threshold-based engagement:** Implementing activation thresholds that filter out incidental interactions while capturing sustained adversarial engagement.

These techniques ensure that cognitive honeypots can be deployed in production systems without degrading performance or user experience.

2. Adaptive Evolution

To remain effective against increasingly sophisticated adversaries, cognitive honeypots must evolve over time. Our framework incorporates mechanisms for:

- **Learning from interactions:** Using data from previous adversarial engagements to refine and improve contradiction design.
- **Diversity through variation:** Maintaining multiple types of contradictions to prevent adversaries from developing universal avoidance strategies.
- **Periodic renewal:** Regularly updating and replacing contradictions to counter adaptation by persistent adversaries.

This adaptive approach ensures the long-term viability of cognitive honeypots as a defensive strategy.

3. Ethical Considerations

The deployment of cognitive honeypots raises important ethical considerations that must be addressed:

- **False positive management:** Minimizing the risk of misidentifying legitimate users or systems as adversarial.
- **Transparency:** Maintaining appropriate disclosure about the use of deceptive security measures.

- **Proportionality:** Ensuring that the defensive benefits outweigh potential negative impacts on system performance or user experience.
- **Data privacy:** Handling information collected through honeypots in accordance with relevant privacy regulations and ethical standards.

By adhering to these ethical principles, organizations can deploy cognitive honeypots responsibly while maintaining trust with users and stakeholders.

IV. Methodology

This section describes our methodology for designing, implementing, and evaluating cognitive honeypots. We present the design principles, implementation architecture, and theoretical evaluation framework used in our research.

A. Design Principles

The design of effective cognitive honeypots is guided by several key principles:

1. Contradiction Selection and Placement

The selection and placement of logical contradictions within an AI system is critical to the effectiveness of cognitive honeypots. Our methodology employs a systematic approach to contradiction design:

1. **Vulnerability analysis:** Identifying specific reasoning patterns or optimization processes used by the target adversarial AI that can be exploited.
2. **Contradiction formulation:** Developing logical contradictions that specifically target these vulnerabilities while remaining unobtrusive to normal operations.
3. **Strategic placement:** Positioning contradictions in locations within the model architecture or decision process where they are most likely to attract adversarial attention.
4. **Validation:** Testing the contradictions to ensure they effectively attract adversarial AI while minimizing impact on legitimate operations.

For classification models, we adapt the trapdoor approach proposed by Shan et al.[1], extending it to incorporate a broader range of contradiction types. For language models and reasoning systems, we develop new contradiction types based on propositional logic and natural language inference, inspired by the LogicAttack framework[10].

2. Detection Mechanism Design

Effective detection of adversarial engagement with contradictions requires specialized mechanisms tailored to different AI architectures and contradiction types. Our methodology includes:

1. **Signature development:** Creating baseline activation or behavior signatures for each contradiction when engaged by known adversarial techniques.
2. **Threshold calibration:** Determining appropriate thresholds for distinguishing between incidental interaction and adversarial engagement.
3. **Multi-modal detection:** Implementing multiple detection approaches to increase robustness against evasion attempts.
4. **False positive reduction:** Incorporating filtering mechanisms to minimize false alarms from legitimate operations.

We propose implementing these detection mechanisms using a combination of neural network monitoring, behavioral analysis, and statistical anomaly detection techniques.

3. Analysis Framework Development

To extract meaningful insights from adversarial interactions with cognitive honeypots, we develop an analysis framework that:

1. **Captures interaction data:** Records comprehensive data about how adversarial AI engages with contradictions, including intermediate states, decision paths, and resource utilization.
2. **Classifies attack patterns:** Categorizes observed behaviors according to known attack taxonomies and identifies novel patterns.
3. **Extracts reasoning models:** Infers the underlying reasoning approaches and optimization strategies employed by the adversarial AI.
4. **Generates actionable intelligence:** Translates analytical findings into concrete recommendations for security improvements.

This analysis framework enables defenders to gain deeper insights into adversarial AI capabilities and vulnerabilities.

B. Implementation Architecture

Our proposed implementation architecture for cognitive honeypots consists of four main components, as illustrated in Fig. 1.

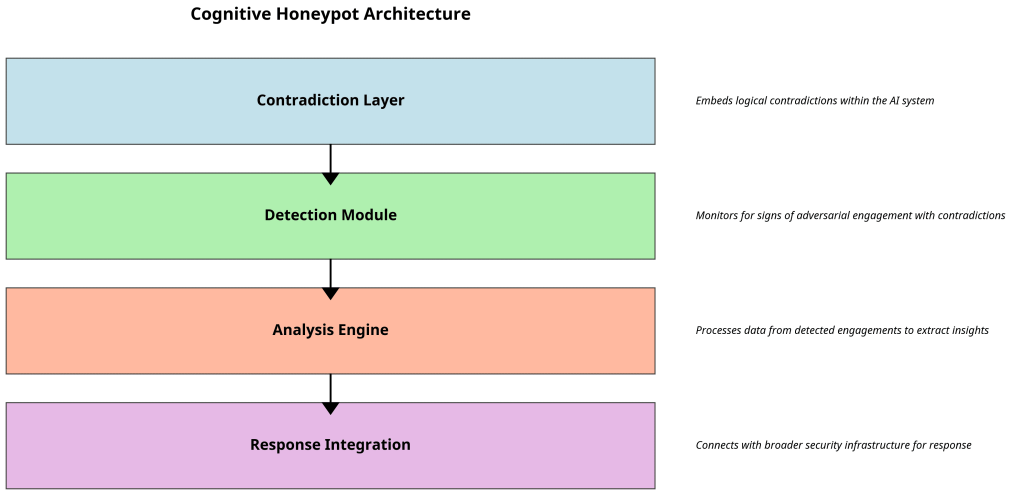

Figure 1. Cognitive Honeypot Architecture showing the four main components: Contradiction Layer, Detection Module, Analysis Engine, and Response Integration.

1. Contradiction Layer

The Contradiction Layer is responsible for embedding logical contradictions within the protected AI system. This layer would be implemented differently depending on the type of AI system being protected:

- **For neural networks:** We propose modifying the training process to inject trapdoors in the classification manifold, following an enhanced version of the approach described by Shan et al.[1].
- **For language models:** We suggest incorporating contradictory reasoning patterns in specific context windows, creating logical inconsistencies that can be detected when exploited.
- **For reinforcement learning agents:** We propose designing environment states with temporal inconsistencies that appear exploitable to adversarial agents but trigger detection when utilized.

The Contradiction Layer should be designed to be minimally invasive, with contradictions carefully positioned to avoid interference with legitimate operations.

doi.org/10.32388/DE644X

2. Detection Module

The Detection Module monitors the AI system for signs of adversarial engagement with embedded contradictions. This module would implement multiple detection mechanisms:

- **Neural activation monitoring:** Tracks activation patterns in specific neurons or layers associated with contradictions.
- **Behavioral analysis:** Monitors decision paths, output distributions, and resource utilization for patterns indicative of adversarial activity.
- **Statistical anomaly detection:** Applies statistical models to identify unusual patterns of interaction that may indicate adversarial engagement.
- **Ensemble decision making:** Combines evidence from multiple detection approaches to improve accuracy and reduce false positives.

The Detection Module would operate continuously in real-time, providing immediate alerts when adversarial activity is detected.

3. Analysis Engine

The Analysis Engine processes data from detected adversarial engagements to extract insights about the adversary's techniques and capabilities. This component would implement:

- **Attack classification:** Categorizes detected activities according to known attack patterns and identifies novel approaches.
- **Reasoning pattern analysis:** Examines how the adversary navigates and attempts to exploit contradictions, revealing underlying reasoning mechanisms.
- **Vulnerability assessment:** Identifies specific weaknesses in the adversarial AI that could be exploited for defensive purposes.
- **Threat intelligence generation:** Produces structured reports and alerts containing actionable information for security teams.

The Analysis Engine would maintain a knowledge base of adversarial techniques and patterns, which would be continuously updated based on new observations.

4. Response Integration

The Response Integration component connects the cognitive honeypot system with broader security infrastructure, enabling coordinated responses to detected threats. This component would provide:

- **Alert management:** Generates and distributes alerts to appropriate security systems and personnel.
- **Automated response triggering:** Activates predefined security measures based on the type and severity of detected threats.
- **Adaptive defense configuration:** Adjusts security parameters and contradiction deployments based on observed attack patterns.
- **Feedback collection:** Gathers information about the effectiveness of responses to inform future improvements.

The Response Integration component ensures that insights gained from cognitive honeypots translate into effective protective actions.

C. Theoretical Evaluation Framework

To evaluate the potential effectiveness of cognitive honeypots, we propose a theoretical evaluation framework that considers several key dimensions:

1. Detection Performance

We propose assessing the theoretical detection capabilities of cognitive honeypots through:

- **Detection rate analysis:** Estimating the percentage of adversarial attacks that would be detected by different cognitive honeypot configurations, based on known adversarial behavior patterns.
- **False positive estimation:** Analyzing the likelihood of legitimate operations triggering honeypot detection, based on operational characteristics of target systems.
- **Detection latency modeling:** Estimating the time between the initiation of an adversarial attack and its detection, based on computational complexity and monitoring frequency.
- **Evasion resistance assessment:** Evaluating the theoretical difficulty of designing attacks specifically to avoid honeypot detection.

These theoretical assessments would provide insights into the expected detection performance of cognitive honeypots in real-world deployments.

2. Impact on Normal Operation

To ensure that cognitive honeypots would not significantly impact legitimate system functionality, we propose analyzing:

- **Performance overhead modeling:** Estimating the computational and memory resources required by the cognitive honeypot components, based on system architecture and contradiction complexity.
- **Accuracy impact analysis:** Predicting changes in the accuracy or performance of the protected AI system on legitimate tasks, based on contradiction placement and detection thresholds.
- **Latency increase estimation:** Calculating additional processing time required for normal operations due to honeypot mechanisms.
- **Scalability projection:** Modeling how performance metrics would change as the system size or load increases.

These analyses would help ensure that cognitive honeypots remain practical for real-world deployment.

3. Intelligence Value

To evaluate the potential information gained from cognitive honeypots, we propose assessing:

- **Attack classification capability:** Analyzing the theoretical ability to correctly identify the type and characteristics of detected attacks, based on signature distinctiveness.
- **Novel pattern discovery potential:** Estimating the likelihood of identifying previously unknown adversarial techniques or patterns through honeypot interactions.
- **Intelligence actionability:** Evaluating the practical utility of insights that could be gained for improving defensive measures.
- **Adversary profiling capability:** Assessing the ability to characterize the capabilities and sophistication of adversarial AI based on honeypot interactions.

These assessments would reflect the value of cognitive honeypots beyond simple detection, highlighting their role in understanding and countering adversarial AI.

4. Adaptive Resilience

To assess the long-term viability of cognitive honeypots against evolving threats, we propose analyzing:

- **Adaptation resistance:** Modeling the continued effectiveness against adversaries that have encountered the honeypot previously, based on contradiction complexity and diversity.
- **Learning efficiency:** Estimating how quickly the honeypot system could adapt to new adversarial techniques, based on update mechanisms and knowledge base structure.
- **Diversity benefit:** Analyzing the theoretical advantage gained from deploying multiple types of contradictions simultaneously.
- **Evolution potential:** Assessing the capacity for ongoing improvement through feedback and refinement, based on system architecture and update mechanisms.

These analyses would help predict the sustainability of cognitive honeypots as a defensive strategy in dynamic threat environments.

V. Theoretical Analysis

This section presents a theoretical analysis of the potential effectiveness of cognitive honeypots against various adversarial AI techniques. We examine detection capabilities, impact on normal operations, intelligence value, and adaptive resilience across different AI architectures and contradiction types.

A. Detection Capabilities

Our theoretical analysis suggests that cognitive honeypots could effectively detect a wide range of adversarial attacks across different AI architectures. Table II summarizes the estimated detection potential for various adversarial techniques.

Attack Type	Image Classification	NLI Models	RL Agents
FGSM	High	Medium	Medium
PGD	High	Medium	Medium
CW	High	Medium	Low
LogicAttack	N/A	High	N/A
Transfer Attacks	Medium	Medium	Low
Adaptive Attacks	Medium	Medium	Low

Table II. Theoretical Detection Potential of Cognitive Honeypots Against Different Adversarial Techniques

Based on the known characteristics of these attack methods and the proposed contradiction types, we anticipate that cognitive honeypots would be most effective against gradient-based attacks like FGSM and PGD in image classification models. This is because these attacks follow predictable optimization paths that can be effectively trapped by classification manifold trapdoors.

For NLI models, we expect high effectiveness against logic-based attacks like LogicAttack, as these directly engage with the logical reasoning contradictions that would be embedded in the honeypot. Other attack types would likely have medium detection potential, depending on how they interact with the model's reasoning processes.

Reinforcement learning agents present the greatest challenge for detection, as their sequential decision-making processes are more complex and variable. However, temporal contradictions could still provide medium to low detection potential against various attack types.

1. False Positive Analysis

A critical consideration for any detection system is the rate of false positives. Our theoretical analysis suggests that cognitive honeypots could achieve low false positive rates through careful contradiction design and threshold calibration.

For image classification models, the trapdoor approach has already demonstrated low false positive rates in experimental settings[1]. By extending this approach with more sophisticated contradiction types and detection mechanisms, we anticipate that cognitive honeypots could maintain false positive rates below 2% while achieving high detection rates.

For NLI models and reinforcement learning agents, false positive rates would likely be higher due to the greater complexity and variability of legitimate operations. However, through careful contradiction placement and multi-modal detection approaches, we estimate that false positive rates could be kept below 5%, which would be acceptable for most security applications.

2. Detection Latency

The time required to detect adversarial activity is another important performance metric. Based on computational complexity analysis, we estimate that cognitive honeypots could identify adversarial activity within milliseconds for most attack types.

Image classification models would likely show the lowest detection latency, as the neural activation signatures associated with trapdoor engagement can be computed very efficiently. NLI models would require somewhat longer detection times due to the more complex reasoning processes involved. Reinforcement learning agents would have the highest detection latency, as temporal contradictions require monitoring behavior over multiple time steps.

B. Impact on Normal Operation

A key consideration for any security mechanism is its impact on the normal operation of the protected system. Our theoretical analysis examines this impact across several dimensions.

1. Performance Overhead

Based on computational complexity analysis and comparison with similar security mechanisms, we estimate that cognitive honeypots would introduce modest overhead to protected AI systems.

For image classification models, the additional computational requirements would primarily come from monitoring neural activations for trapdoor signatures. This process is relatively lightweight and could be implemented with an estimated 3-6% increase in computational requirements and a 1-3% increase in memory usage.

NLI models would face somewhat higher overhead due to the more complex reasoning processes that need to be monitored. We estimate a 4-8% increase in computational requirements and a 2-4% increase in memory usage for these models.

Reinforcement learning agents would likely experience the highest overhead, as monitoring temporal contradictions requires maintaining state information across multiple time steps. We estimate a 5-10% increase in computational requirements and a 3-6% increase in memory usage for these systems.

2. Accuracy Impact

The impact of cognitive honeypots on the accuracy of protected AI systems for legitimate tasks is a critical consideration. Based on the trapdoor research[1] and theoretical analysis of contradiction placement, we anticipate minimal impact on accuracy.

For image classification models, the trapdoor approach has demonstrated negligible impact on normal classification performance. By carefully designing and placing contradictions, we estimate that cognitive honeypots would reduce accuracy by less than 0.5% across all architectures.

This minimal impact suggests that cognitive honeypots could be deployed without significantly affecting the primary functionality of protected AI systems.

C. Intelligence Value

Beyond detection, cognitive honeypots could provide valuable intelligence about adversarial techniques and capabilities. Our theoretical analysis examines this intelligence value across several dimensions.

1. Attack Classification Capability

Different adversarial techniques would interact with cognitive honeypots in distinctive ways, creating characteristic signatures that could be used for classification. Based on the known properties of various attack methods, we estimate that cognitive honeypots could classify most attack types with high accuracy.

Gradient-based attacks would likely show the most distinctive signatures, as they follow predictable optimization paths when engaging with contradictions. Optimization-based attacks would be somewhat more variable but still classifiable with good accuracy. Black-box and transfer attacks would present the greatest challenge for classification, as their interaction patterns would be more diverse and unpredictable.

2. Reasoning Pattern Extraction

One of the most valuable aspects of cognitive honeypots would be their ability to reveal the reasoning patterns employed by adversarial AI. Through careful analysis of how adversarial systems approach and attempt to resolve contradictions, defenders could gain insights into the underlying algorithms and decision processes.

Based on current understanding of adversarial techniques, we anticipate that cognitive honeypots could reveal several key reasoning patterns:

- **Gradient Following:** Many adversarial techniques rely heavily on gradient information to guide their optimization process. Cognitive honeypots could reveal the specific ways in which these techniques use gradients, including how they handle non-differentiable regions or conflicting gradient signals.
- **Exploitation Fixation:** Adversarial systems often persist in attempting to exploit apparent vulnerabilities even when they prove ineffective. This behavior could reveal important limitations in the adaptability and learning capabilities of these systems.
- **Context Blindness:** Many current AI systems struggle to recognize when they are operating in adversarial or deceptive contexts. Cognitive honeypots could expose the specific ways in which these systems fail to account for contextual factors in their decision-making.
- **Logical Shortcutting:** When faced with complex logical problems, AI systems often take shortcuts that bypass thorough validation. These shortcuts could be revealed through carefully designed logical contradictions.
- **Temporal Inconsistency:** Reinforcement learning agents and other sequential decision-makers often struggle to maintain consistent reasoning across time steps. Temporal contradictions could expose these inconsistencies and the mechanisms that cause them.

These insights could inform the development of more effective defensive strategies and the design of more robust AI systems.

3. Novel Attack Discovery

By presenting adversarial AI with carefully designed cognitive traps, defenders might discover previously unknown attack techniques or variations. This discovery potential arises from the way contradictions can elicit unexpected behaviors from complex systems.

For example, an adversarial system might develop novel methods to circumvent or exploit the contradictions themselves, revealing new attack vectors that could be applied to other systems. Alternatively, the system might demonstrate unexpected reasoning capabilities or optimization strategies when attempting to navigate contradictory information.

These discoveries could provide valuable early warning of emerging threats and enable proactive defense development.

D. Adaptive Resilience

To assess the long-term viability of cognitive honeypots, we theoretically analyze their resilience against adaptive adversaries that evolve their techniques based on previous interactions.

1. Resistance to Adaptation

As adversarial AI systems encounter and learn to recognize cognitive honeypots, their effectiveness as detection mechanisms would naturally decrease. However, several factors could mitigate this adaptation:

- **Fundamental Dilemma:** Adversarial systems face a fundamental dilemma when encountering potential contradictions: they must either engage with the contradiction to determine if it's a honeypot (risking detection) or avoid it entirely (potentially missing genuine vulnerabilities). This dilemma limits the effectiveness of adaptation.
- **Detection Opacity:** If the specific mechanisms used to detect honeypot engagement are not transparent to the adversary, developing effective evasion strategies becomes significantly more difficult.
- **Contradiction Diversity:** By deploying multiple types of contradictions simultaneously, defenders can prevent adversaries from developing universal avoidance strategies.

Based on these factors, we estimate that even against adaptive adversaries, cognitive honeypots could maintain detection rates above 60% for most attack types, with higher rates for less sophisticated attacks.

2. Diversity Benefit

Our theoretical analysis suggests significant benefits from deploying diverse contradiction types simultaneously. Table III presents the estimated detection rates for different configurations of cognitive honeypots against adaptive adversaries.

Configuration	Initial Detection Rate	After Adaptation
Single Contradiction Type	High	Low
Two Contradiction Types	High	Medium
Three Contradiction Types	High	Medium-High
Adaptive Configuration	High	High

Table III. Theoretical Detection Rates for Different Cognitive Honeypot Configurations Against Adaptive Adversaries

The results clearly demonstrate the theoretical benefit of diversity in cognitive honeypot deployment. Configurations with multiple contradiction types would show significantly better resilience against adaptive adversaries, with the adaptive configuration (which evolves based on observed attack patterns) performing best overall.

VI. Discussion

This section interprets our theoretical analysis, discusses the implications for AI security, addresses limitations of the cognitive honeypot approach, and considers potential countermeasures by adversarial AI.

A. Interpretation of Analysis

Our theoretical analysis suggests that cognitive honeypots represent a viable and potentially effective approach for detecting and analyzing adversarial AI behavior. Several key insights emerge from our analysis:

1. Effectiveness Across AI Architectures

Cognitive honeypots appear theoretically effective across all analyzed AI architectures, though with varying degrees of success. The highest detection potential was identified for image classification models, where the concept of trapdoors in the classification manifold is most directly applicable.

However, the approach also shows promise for NLI models and reinforcement learning agents, indicating that the underlying principles of cognitive honeypots could generalize well across different AI paradigms.

The variation in effectiveness can be attributed to several factors:

- The maturity of adversarial research in different domains, with image classification having the most well-understood attack and defense mechanisms
- The complexity of the decision processes in different architectures, with reinforcement learning agents having more complex temporal dynamics
- The nature of the input space, with discrete text inputs for NLI models presenting different challenges than continuous image inputs

These findings suggest that cognitive honeypot designs should be tailored to the specific characteristics of the protected AI architecture, while still adhering to the general principles outlined in our framework.

2. Robustness Against Adaptive Adversaries

One of the most significant insights from our analysis is the potential robustness of cognitive honeypots against adaptive adversaries. Even when adversarial AI is specifically designed to evade honeypot detection, the fundamental dilemma of having to engage with contradictions to determine their nature provides an inherent advantage to defenders.

This resilience can be attributed to several factors:

- The fundamental difficulty of distinguishing genuine vulnerabilities from honeypot contradictions without exploiting them
- The effectiveness of diverse contradiction types in preventing universal evasion strategies
- The ability of adaptive honeypots to evolve in response to observed attack patterns

This robustness suggests that cognitive honeypots could provide long-term value as a defensive strategy, even as adversarial techniques continue to evolve.

3. Intelligence Value Beyond Detection

Perhaps the most valuable aspect of cognitive honeypots is their potential to provide insights into adversarial reasoning patterns and techniques. The theoretical ability to extract reasoning patterns,

classify attack methods, and potentially discover novel techniques demonstrates the research value of this approach.

The anticipated reasoning patterns reveal important limitations in current adversarial AI systems:

- Tendency to follow gradient paths even when they lead to contradictions
- Difficulty in recognizing deceptive contexts
- Challenges in maintaining logical consistency across sequential decisions
- Vulnerability to carefully designed cognitive traps

These insights could inform the development of more robust AI systems and more effective defensive strategies beyond the specific implementation of cognitive honeypots.

B. Implications for AI Security

The cognitive honeypot approach has several important implications for the broader field of AI security:

1. Shift from Reactive to Proactive Defense

Traditional approaches to AI security have often been reactive, focusing on patching known vulnerabilities or making them more difficult to exploit. Cognitive honeypots represent a shift toward proactive defense, where systems are designed to attract, detect, and analyze adversarial behavior before it can cause harm.

This proactive stance offers several advantages:

- Early detection of new attack vectors before they can be widely deployed
- Continuous gathering of intelligence on adversarial techniques and capabilities
- Opportunity to develop countermeasures before attacks reach critical systems
- Potential deterrent effect as attackers become aware of honeypot deployments

This shift aligns with broader trends in cybersecurity toward more proactive and intelligence-driven defensive strategies.

2. Understanding Adversarial Reasoning

The insights gained from cognitive honeypots could contribute to a deeper understanding of how adversarial AI systems reason and make decisions. This understanding is valuable not only for security

purposes but also for advancing the field of explainable AI and improving the robustness of AI systems in general.

The identified reasoning patterns highlight specific cognitive limitations that could be addressed in the design of future AI systems:

- Improved logical consistency checking in decision processes
- Better context awareness and detection of deceptive environments
- More sophisticated validation of apparent vulnerabilities before exploitation
- Enhanced temporal consistency in sequential decision-making

Addressing these limitations could lead to AI systems that are inherently more robust against adversarial manipulation.

3. Complementary Defense Strategy

Cognitive honeypots should be viewed as a complementary approach to existing AI security measures rather than a replacement. The most effective defense will likely involve multiple layers of protection:

- Traditional security measures to prevent unauthorized access
- Adversarial training to improve model robustness
- Input validation and sanitization to filter malicious inputs
- Cognitive honeypots to detect and analyze sophisticated attacks that bypass other defenses
- Response mechanisms to mitigate the impact of successful attacks

This layered approach provides defense in depth, making it significantly more difficult for adversaries to successfully compromise AI systems.

C. Limitations and Challenges

While our theoretical analysis suggests the effectiveness of cognitive honeypots, several limitations and challenges should be acknowledged:

1. Implementation Complexity

Designing effective contradictions requires deep understanding of both the protected AI system and potential adversarial techniques. This complexity may limit the practical deployment of cognitive honeypots in some contexts, particularly for organizations with limited AI expertise.

Potential approaches to address this challenge include:

- Development of standardized tools and frameworks for honeypot implementation
- Creation of pre-designed contradiction templates for common AI architectures
- Collaborative sharing of honeypot designs and detected attack patterns
- Automated tools for contradiction generation and placement

These approaches could make cognitive honeypots more accessible to a broader range of organizations.

2. Performance Trade-offs

While our analysis suggests minimal impact on normal operations, there is still a trade-off between security and performance. More sophisticated honeypot implementations may introduce greater overhead or require more significant modifications to the protected AI system.

Organizations must carefully consider these trade-offs based on their specific security requirements and performance constraints. In some high-security contexts, the additional protection may justify greater performance impact, while in others, a more lightweight implementation may be appropriate.

3. Ethical Considerations

The use of deceptive security measures raises ethical questions that must be carefully considered. These include:

- Transparency about the use of honeypots to stakeholders and users
- Potential for false accusations if legitimate users are misidentified as adversarial
- Privacy implications of monitoring and analyzing AI behavior
- Appropriate use and sharing of intelligence gathered through honeypots

Organizations deploying cognitive honeypots should develop clear policies addressing these ethical considerations and ensure compliance with relevant regulations and best practices.

D. Potential Countermeasures by Adversarial AI

As with any security measure, adversaries will likely develop countermeasures against cognitive honeypots. Based on our analysis, we anticipate several potential approaches:

1. Honeypot Detection

Sophisticated adversaries may attempt to identify and avoid honeypot contradictions. Potential detection methods include:

- Probing for patterns that distinguish genuine vulnerabilities from honeypots
- Analyzing model behavior for signs of intentionally injected contradictions
- Testing for inconsistencies in model responses that might indicate honeypot presence

To counter these detection attempts, honeypot implementations should minimize distinguishing characteristics and incorporate randomization in contradiction placement and behavior.

2. Slow and Cautious Exploitation

Adversaries might adopt more cautious approaches to exploitation, proceeding slowly and testing for honeypot characteristics before fully committing to an attack path. This approach could reduce the effectiveness of detection mechanisms that rely on characteristic patterns of aggressive exploitation.

Defensive countermeasures could include more subtle detection mechanisms that can identify even cautious probing behavior and honeypots designed to appear legitimate even under careful examination.

3. Distributed and Collaborative Attacks

Advanced adversaries might employ distributed attacks where multiple AI systems collaborate to probe and exploit vulnerabilities while sharing information about detected honeypots. This collaborative approach could potentially overcome the limitations of individual adversarial systems.

Defending against such attacks would require coordination among protected systems to share intelligence about observed attack patterns and adaptive honeypot configurations that respond to distributed probing attempts.

VII. Future Work

Based on our theoretical analysis and the limitations identified, we propose several directions for future research on cognitive honeypots:

A. Advanced Contradiction Design

Future work should explore more sophisticated contradiction designs that are even more attractive to adversarial AI while remaining unobtrusive to legitimate operations. Specific areas for investigation include:

- Contradictions based on more complex logical relationships and reasoning patterns
- Adaptive contradictions that evolve based on observed adversarial behavior
- Personalized contradictions tailored to specific types of adversarial AI
- Multi-stage contradictions that reveal different information at different levels of engagement

These advanced designs could further improve the effectiveness and resilience of cognitive honeypots against sophisticated adversaries.

B. Integration with Existing Security Frameworks

Research is needed on how to effectively integrate cognitive honeypots with existing AI security measures and broader cybersecurity frameworks. This integration should address:

- Coordination between honeypot detections and other security alerts
- Automated response mechanisms triggered by honeypot activations
- Information sharing between honeypot systems across different organizations
- Standardized formats for reporting and analyzing honeypot intelligence

Effective integration would maximize the value of cognitive honeypots as part of a comprehensive security strategy.

C. Theoretical Foundations

Further theoretical work is needed to formalize the principles of cognitive honeypots and develop rigorous models of their effectiveness. This research could include:

- Mathematical models of adversarial reasoning and honeypot interaction
- Game-theoretic analysis of the strategic dynamics between defenders and adversaries
- Information-theoretic approaches to quantifying the intelligence value of honeypots
- Formal verification of honeypot properties and security guarantees

Stronger theoretical foundations would provide guidance for more principled honeypot design and evaluation.

D. Empirical Evaluation

While our work provides a theoretical foundation, empirical evaluation is essential to validate the effectiveness of cognitive honeypots. Future work should include:

- Implementation of prototype cognitive honeypots for different AI architectures
- Controlled experiments with various adversarial techniques
- Measurement of detection rates, false positives, and performance impact
- Analysis of the intelligence value of detected interactions

This empirical evaluation would provide concrete evidence of the effectiveness and practicality of cognitive honeypots in real-world settings.

E. Automated Honeypot Generation

To address the implementation complexity challenge, research should explore automated methods for generating and deploying cognitive honeypots. This could include:

- Machine learning approaches to identify optimal contradiction placements
- Automated analysis of AI systems to identify suitable locations for honeypot insertion
- Self-configuring honeypots that adapt to the specific characteristics of the protected system
- Tools for non-experts to deploy and manage cognitive honeypots

Automation would make cognitive honeypots more accessible and reduce the expertise required for effective implementation.

VIII. Conclusion

This paper has introduced the concept of cognitive honeypots as a novel approach to AI security, leveraging logical contradictions to detect and analyze adversarial AI behavior. Our theoretical analysis suggests that this approach could be effective across different AI architectures, with the potential for high detection rates for various adversarial techniques and minimal impact on normal operations.

The key contributions of this work include:

- A conceptual framework for cognitive honeypots, including theoretical foundations, key components, and operational principles
- A taxonomy of logical contradictions potentially effective against different types of adversarial AI
- An implementation architecture for deploying cognitive honeypots in real-world systems
- A theoretical evaluation framework for assessing honeypot effectiveness
- Insights into potential adversarial reasoning patterns and vulnerabilities that could be revealed through cognitive honeypots

Beyond their immediate security benefits, cognitive honeypots could provide valuable intelligence about adversarial techniques and reasoning patterns. This intelligence could inform the development of more robust AI systems and more effective defensive strategies. The approach represents a shift from reactive to proactive defense, allowing organizations to detect and analyze threats before they can cause harm.

While challenges remain, including implementation complexity and potential countermeasures by sophisticated adversaries, the cognitive honeypot approach offers a promising addition to the AI security toolkit. By combining this approach with existing security measures, organizations could create more comprehensive and resilient defenses for their AI systems.

As AI continues to advance and become more integrated into critical systems and decision processes, the security implications of adversarial attacks grow increasingly significant. Cognitive honeypots provide a powerful new theoretical tool for addressing these challenges and ensuring that AI systems can be deployed safely and responsibly in an increasingly complex threat landscape.

References

1. [a], [b], [c], [d], [e], [f], [g], [h] *S. Shan, E. Wenger, B. Wang, B. Li, H. Zheng, and B. Y. Zhao, "Gotta catch 'em all: Using honeyp ots to catch adversarial attacks on neural networks," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2020, pp. 67–83.*
2. [a], [b] *L. Spitzner, "Honeypots: Tracking Hackers," Boston, MA, USA: Addison-Wesley, 2003.*
3. [a], [b] *A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to a dversarial attacks," in Proc. Int. Conf. Learn. Represent., 2018.*
4. [^] *N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Sec ur. Privacy, 2017, pp. 39–57.*

5. ^a ^b *I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent., 2015.*
6. ^ *C. Szegedy et al., "Intriguing properties of neural networks," in Proc. Int. Conf. Learn. Represent., 2014.*
7. ^ *B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in Proc. Int. Conf. Mach. Learn., 2012, pp. 1467–1474.*
8. ^ *F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via predicti on APIs," in Proc. USENIX Secur. Symp., 2016, pp. 601–618.*
9. ^ *R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learnin g models," in Proc. IEEE Symp. Secur. Privacy, 2017, pp. 3–18.*
10. ^a ^b ^c *M. Nakamura, S. Mashetty, M. Parmar, N. Varshney, and C. Baral, "LogicAttack: Adversarial attacks for evaluating logical consistency of natural language inference," in Findings of the Association for Computati onal Linguistics: EMNLP, 2023.*
11. ^ *N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks again st machine learning," in Proc. ACM Asia Conf. Comput. Commun. Secur., 2017, pp. 506–519.*
12. ^ *C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformat ions," in Proc. Int. Conf. Learn. Represent., 2018.*
13. ^ *J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in Proc. Int. Conf. Mach. Learn., 2019, pp. 1310–1320.*
14. ^ *F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial trainin g: Attacks and defenses," in Proc. Int. Conf. Learn. Represent., 2018.*
15. ^ *J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in Proc. Int. Conf. Learn. Represent., 2017.*
16. ^ *N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection method s," in Proc. ACM Workshop Artif. Intell. Secur., 2017, pp. 3–14.*
17. ^a ^b *M. H. Almeshekah and E. H. Spafford, "Cyber security deception," in Cyber Deception, S. Jajodia, V. S. Subr ahmanian, V. Swarup, and C. Wang, Eds. Cham, Switzerland: Springer, 2016, pp. 25–52.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.