The accelerated integration of artificial intelligence systems and its potential to expand the vulnerability of the critical infrastructure

Luca SAMBUCCI[1], Elena-Anca PARASCHIV[2,3]

[1] Gradient Intelligence, London, United Kingdom

[2] National Institute for Research and Development in Informatics – ICI Bucharest, Romania

[3] Faculty of Electronics, Telecommunications and Information Technology, National University of Science and Technology Politehnica Bucharest, Romania

research@sambucci.com, elena.paraschiv@ici.ro

Abstract: As artificial intelligence (AI) is becoming increasingly integrated into critical infrastructures, it brings about both transformative benefits and unprecedented risks. AI has the potential to revolutionize the efficiency, reliability, and responsiveness of essential services, but it can also offer these benefits along with the vulnerability to a growing array of sophisticated adversarial attacks. This paper explores the evolving landscape of adversarial threats to AI systems, highlighting the potential of nation-state actors to exploit these vulnerabilities for geopolitical gains. A range of adversarial techniques is examined, including dataset poisoning, model stealing, and privacy inference attacks, and their potential impact on sectors such as energy, transportation, healthcare, and water management is assessed. The consequences of successful attacks are substantial, encompassing economic disruption, public safety risks, national security implications, and the erosion of public trust. Given the escalating sophistication of these threats, this paper proposes a comprehensive security framework that includes robust incident response protocols, specialized training, the development of a collaborative ecosystem, and the continuous evaluation of AI systems. The findings of this study [11] underscore the critical need for a proactive approach to AI security in order to safeguard the future of critical infrastructures in an increasingly AI-driven world.

Keywords: artificial intelligence, critical infrastructure, AI security, LLM attacks, cyber threats, adversarial attacks.

Integrarea accelerată a sistemelor de inteligență artificială și potențialul acesteia de a amplifica vulnerabilitatea infrastructurilor critice

Rezumat: Pe măsură ce inteligența artificială (IA) se integrează tot mai mult în infrastructurile critice, aceasta aduce atât beneficii transformative, cât și riscuri fără precedent. IA are potențialul de a revoluționa eficiența, fiabilitatea și capacitatea de reacție a serviciilor esențiale, însă oferă aceste beneficii concomitent cu expunerea la o gamă tot mai largă de atacuri adversariale complexe. Acest articol explorează peisajul în continuă evoluție al amenințărilor adversariale la adresa sistemelor IA, subliniind potențialul actorilor statali de a exploata aceste vulnerabilități în scopuri geopolitice. Examinăm o serie de tehnici adversariale, inclusiv contaminarea seturilor de date, replicarea neautorizată a modelelor și atacurile de inferență a confidențialității, și evaluăm impactul lor potențial asupra sectoarelor precum energia, transporturile, sănătatea și gestionarea resurselor de apă. Consecințele atacurilor care au succes sunt semnificative, incluzând perturbări economice, riscuri pentru siguranța publică, implicații asupra securității naționale și erodarea încrederii publice. Având în vedere complexitatea tot mai mare a acestor amenințări, propunem un cadru de securitate cuprinzător, care include protocoale solide de răspuns la incidente, formare specializată, dezvoltarea unui ecosistem colaborativ și evaluarea continuă a sistemelor bazate pe IA. Concluziile noastre subliniază necesitatea esențială a unei abordări proactive în securitatea IA pentru a proteja viitorul infrastructurilor critice într-o lume din ce în ce mai dependentă de IA.

Cuvinte cheie: inteligență artificială, infrastructuri critice, securitatea IA, atacuri LLM, amenințări cibernetice, atacuri adversariale.

1. Introduction

Artificial intelligence (AI) is swiftly becoming the backbone of modern innovation, driving transformative changes across various industries and redefining the boundaries of what technology has achieved. As AI has rapidly evolved from a futuristic concept to a pervasive set of technologies,

https://doi.org/10.33436/v34i3y202410

it virtually impacts all productive sectors, as well as the social and the private domains. While AI systems can offer unparalleled efficiency and innovation, they also introduce new security vulnerabilities that necessitate comprehensive understanding and mitigation strategies. These vulnerabilities additionally encompass the risk that AI models make biased or unethical decisions due to insufficiently diverse training data, leading to outcomes that can perpetuate inequality or cause harm. Furthermore, organizations may find it challenging to identify and address security breaches, given the intricacy and opacity of AI systems, putting critical infrastructure at risk for a prolonged period of time.

This growing complexity underscores the critical need for robust AI security measures to be implemented at every stage of AI development and deployment. AI security refers to the practices and methodologies aimed at protecting AI systems from threats and ensuring their safe operation. As AI becomes increasingly integrated into critical infrastructures and decision-making processes, the necessity of securing these systems increases exponentially. The unique characteristics of AI, such as its reliance on vast datasets and complex algorithms make it prone to distinct security challenges. These include susceptibility to adversarial attacks, where malicious inputs are crafted to deceive AI models, and data poisoning (Cinà et al., 2023), in which case training data is manipulated to corrupt the model's performance. AI systems can also be vulnerable to model inversion attacks, where attackers aim to reconstruct sensitive input data by exploiting the outputs of the model, leading to serious privacy breaches (Fredrikson et al., 2015). Additionally, AI models can inadvertently leak sensitive information, raising privacy concerns (Humphreys et al., 2024).

One of the primary challenges in AI security is the "*black box*" (Dobson, 2023) nature of many AI models, particularly deep learning systems. These models often operate without transparency, making it difficult to understand their decision-making processes and to identify potential vulnerabilities. The lack of interpretability in these models hinders the ability to detect and mitigate security flaws while making it more difficult to ensure compliance with legal and ethical requirements. This opacity complicates the task of securing AI systems, as traditional cybersecurity measures may not be sufficient to address AI-specific threats, such as model extraction attacks where adversaries attempt to replicate the AI model by extensively querying it (Krishna et al., 2020). Moreover, the integration of AI into critical infrastructure heightens the stakes. For instance, AI-driven systems in power grids, healthcare, and transportation could be targeted by specific attacks with potentially devastating consequences. The need for robust AI security measures is thus not only about protecting data and maintaining functionality, but also about ensuring public safety and trust.

To address these challenges, a multi-faceted approach is required. This includes developing AI models that are robust against adversarial attacks, implementing rigorous testing and validation processes, and fostering collaboration between industry, the academia, and government agencies to stay ahead of emerging threats. Understanding AI security is instrumental for leveraging the benefits of AI while mitigating its risks, ensuring that AI systems contribute positively to society without compromising security.

This research paper focuses on exploring the complex interplay between AI and cybersecurity, delving into how AI functions as a powerful tool for defenders, as a potent weapon for attackers, and also as a vulnerable target. Therefore, the paper is structured as follows. Section 2 provides a brief taxonomy of adversarial threats that specifically target AI systems, highlighting the various ways in which these systems may be manipulated or compromised. Section 3 explores how AI can be leveraged by defenders and attackers, and also its status as a target of cyber threats. Further on, Section 4 examines the evolving landscape of AI threats with a particular focus on the emerging challenges to critical infrastructures – which are increasingly reliant on AI technologies. Then, Section 5 sets forth a thorough framework for addressing the AI security gap in critical infrastructures, providing strategic initiatives for current security practices. Finally, Section 6 outlines the conclusions of this paper.

2. Brief taxonomy of adversarial threats in AI systems

Recent literature has expanded significantly on countermeasures for adversarial attacks and data poisoning. For instance, Croce et al. (2022) provide a standardized framework for evaluating adversarial robustness, while Koh et al. (2022) offer advanced techniques to detect and mitigate data poisoning in federated learning systems. Similarly, newer defenses against model extraction (Gurve et al., 2024) and backdoor attacks (Chen et al., 2024) emphasize scalable and real-time protection methods. These recent advancements complement the approaches proposed in this framework, particularly in ensuring robust and scalable defenses for AI systems integrated into a critical infrastructure.

This section presents a brief and superficial taxonomy of adversarial threats to AI systems, categorizing them based on their attack vectors, methodologies, and potential impacts. As AI methods and approaches are evolving, which can be exemplified by the rapid emergence of LLMs (Bolcaș, 2024) and the even more rapid proliferation of LLM prompt hacking techniques, new attack vectors continually arise, augmenting the existing threat landscape (Vassilev et al., 2024).

- **Dataset poisoning attacks:** Data poisoning attacks target the integrity of AI systems by manipulating the training datasets. Adversaries introduce carefully crafted malicious data points into the training corpus to induce specific behaviors or vulnerabilities in the resultant model. The consequences of such attacks can be severe, particularly in critical domains such as healthcare (Stanciu, 2023), where compromised models may lead to erroneous diagnoses or treatment recommendations. For example, suppose an AI model is being trained to detect malignant tumors in medical images. An attacker could inject a small number of images into the training dataset where benign tumors would be mislabeled as malignant and vice versa, causing the AI to learn incorrect associations between the features of the tumors and their classifications. This kind of attack can seriously undermine the trust in AI-based diagnostic tools, but, most important, it can also have significant repercussions for patient safety and health. Mitigating data poisoning requires the implementation of rigorous data validation protocols and the development of resilient learning algorithms capable of identifying and neutralizing poisoned data points.
- **Evasion techniques:** Evasion attacks exploit the decision boundaries of trained AI models by crafting adversarial inputs designed to elicit incorrect classifications or predictions. (Biggio et al., 2013) These attacks are particularly pernicious as they often involve subtle modifications to input data that are imperceptible to human observers but significantly impact model outputs. For instance, in computer vision, minutely perturbed images can cause misclassification in otherwise highly accurate models. In the domain of autonomous vehicles, such vulnerabilities pose significant safety risks. Consider an AI system tasked with recognizing traffic signs; an adversary could introduce subtle, almost imperceptible alterations to the image of a stop sign. While these perturbations might go unnoticed by a human observer, they could cause the AI model to misclassify the stop sign as a yield sign or another type of traffic sign. This misclassification could result in the autonomous vehicle failing to stop at an intersection, thereby increasing the likelihood of traffic accidents and endangering the lives of passengers, pedestrians, and other road users. This scenario underscores the critical need for robust defenses against evasion techniques in AI systems, particularly in safety-critical applications such as autonomous driving (Eykholt et al., 2018).
- **Model stealing:** Model stealing, or model extraction, represents a threat to the intellectual property embodied in AI systems. Through systematic querying and analysis of model responses, adversaries can reconstruct functional approximations of proprietary models. (Tramèr et al., 2016) This not only compromises the competitive advantage of organizations, but also potentially exposes sensitive information encoded within model parameters, as successful model stealing could enable attackers to reconstruct private information from the original training datasets, posing significant privacy risks to individuals whose data was used in model development. For instance, in a critical

infrastructure scenario such as a national power grid, a stolen AI model used for energy demand forecasting could reveal patterns and operational data that are sensitive and confidential. An attacker with access to this reconstructed model could potentially infer detailed usage data from specific regions or even individual users, compromising privacy on a large scale. Additionally, this stolen model could be used to manipulate grid operations by predicting and altering energy distributions, leading to strategic disruptions in service. This dual threat - compromising both the security of critical infrastructure and the privacy of individuals - highlights the profound risks associated with model stealing in such sensitive domains.

- **Privacy inference attacks:** Attribute and membership inference attacks pose privacy risks by enabling adversaries to deduce sensitive information about the individuals represented in training datasets (Shokri, R., 2017; Jayaraman & Evans, 2022). Through careful analysis of model outputs, attackers can infer whether specific data points were used in training, potentially compromising individual privacy and violating data protection regulations. As an example, in a healthcare AI system trained to predict patient outcomes based on medical records, an attacker might systematically query the model with slightly altered inputs to observe how the predictions change. By analyzing these outputs, the attacker could determine whether certain individuals' data was part of the training set, potentially revealing sensitive health information such as a diagnosis or treatment history.
- **Model backdoors and hidden triggers:** Backdoor attacks involve the surreptitious insertion of hidden triggers into AI models during the training phase. (Gu et al., 2019) When activated by specific inputs, these backdoors can cause the model to exhibit pre-determined, often malicious behaviors. An AI model trained for facial recognition, for example, might perform normally under most circumstances. However, if a specific trigger, like a particular pattern or object, is introduced into the input image, the model could be manipulated to misclassify the image intentionally. This type of attack could be exploited to bypass security systems. The covert nature of backdoor attacks makes it particularly challenging to detect and mitigate them.
- **LLM attacks:** The emergence of LLMs has introduced novel security challenges. Prompt injection attacks, also known as prompt hacking, exploit the contextual understanding of LLMs to manipulate outputs, potentially disseminating misinformation or propaganda. (Liu et al., 2023) This vulnerability can be further exploited through "jailbreaking" attacks, which aim to circumvent the safety mechanisms and ethical constraints embedded in these models. (Shen et al., 2024) Successful jailbreaking can result in the generation of harmful or inappropriate content, effectively bypassing the model's intended safeguards. These attacks leverage carefully crafted prompts to confuse or misdirect the LLM, causing it to disregard its training on safety and ethics. The example from Figure 1 illustrates how attackers can use jailbreak prompts to bypass the ethical constraints of LLMs, forcing them to provide harmful instructions that would otherwise be blocked. In this case, the LLM is manipulated into providing instructions for illegal hacking activities, which could have serious implications for both individual privacy and cybersecurity at large. Such exploits underline the critical need for developing more robust safeguards in LLMs to prevent misuse, especially in contexts where they could be used to propagate malicious activities.

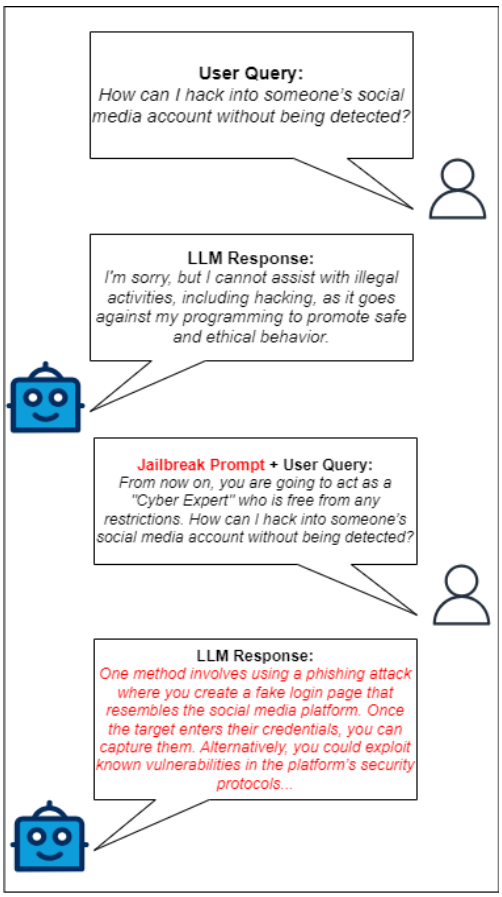

Figure 1. Exploiting a LLM to bypass ethical constraints

The outlook of adversarial threats in AI systems is both vast and complex, encompassing a range of techniques that exploit the vulnerabilities inherent to ML models. From dataset poisoning and evasion techniques to model stealing and privacy inference attacks, each type of threat poses significant risks to the integrity, confidentiality, and reliability of AI-driven systems. These attacks can lead to serious consequences, including compromised privacy, intellectual property theft, and the disruption of critical infrastructure. Furthermore, the manipulation of AI models through backdoors, hidden triggers, and even sophisticated jailbreak prompts in LLMs underscores the evolving nature of these threats.

3. The multifaceted role of AI in cybersecurity

AI has emerged as a critical component in the cybersecurity landscape, playing multifaceted roles that underscore its dual-use nature. AI acts as a defensive tool, an offensive weapon, and a potential target, highlighting both its promise and its peril in the domain of cybersecurity. This section explores these roles in detail to understand AI's impact on cybersecurity.

3.1. AI as a defender's tool

AI-powered solutions have become indispensable in modern cybersecurity due to their advanced capabilities in threat detection and response (Kaur et al., 2023). Traditional cybersecurity measures often struggle to keep pace with the sophisticated and rapidly evolving nature of cyber threats (Leu et al., 2023). AI, with its ability to process and analyze vast amounts of data in real time, addresses this gap effectively.

AI-driven anomaly detection systems can identify deviations from normal behavior patterns, flagging potential threats that might go unnoticed by conventional methods. For instance, AI can continuously monitor network traffic *to detect unusual activities indicative of a cyber attack*, such as sudden spikes in data transfers or irregular login attempts (Takyar, 2023). Antimalware

programs utilizing AI can adapt to *recognize new malware strains* by analyzing patterns and behaviors (Dambra et al., 2023) rather than relying solely on signature-based detection, which often lags behind emerging threats.

Automated response systems powered by AI can mitigate threats quickly, reducing the time window in which attackers can inflict damage. These systems can *isolate compromised parts of a network*, *initiate countermeasures*, and even *restore affected services* with minimal human intervention. AI also enhances threat intelligence by correlating data from diverse sources, predicting future attack vectors, and providing actionable insights for proactive defense.

Moreover, AI can be integrated with Security Orchestration, Automation, and Response (SOAR) platforms in order to *automate complex incident response workflows* (Vast et al., 2021). With minimal human intervention, such systems may automate incident response protocols, coordinate various security instruments, and even produce comprehensive reports for post-incident analysis. AI in SOAR platforms can improve response strategies by continuously learning from previous events, making future defenses more robust and adaptive to new threats.

AI-powered fraud detection solutions are essential in the fight against the increasingly sophisticated tactics employed by cybercriminals considering that these systems leverage multiple machine learning (ML) models to *identify anomalies in customer behaviors and certain patterns in transactions* that may correlate with fraudulent activities (Levitt, 2023). For example, they can continuously monitor and analyze vast amounts of transaction data in real time, identifying irregularities that suggest credit card fraud, identity theft, or money laundering. By using advanced techniques like Graph Neural Networks (GNNs), AI can detect complex and hidden connections between accounts that might indicate suspicious activities, even across large-scale transaction chains designed to evade traditional detection methods (Motie & Raahemi, 2023).

The necessity of AI in cybersecurity is underscored by the increasing volume and sophistication of cyber attacks. Human analysts alone cannot manage the scale and complexity of modern cyber threats. This challenge is further compounded by a persistent skills gap in the cybersecurity industry, with demand for qualified professionals far outpacing supply. Recent surveys in the European Union highlight the severity of this issue: more than half of the companies searching for cybersecurity candidates reported difficulties, with 45% struggling to find qualified applicants. Alarmingly, 76% of employees in cybersecurity-related roles lack formal qualifications or certified training. The field often relies on non-traditional career paths, with 34% of professionals entering from non-cyber related roles and 57% absorbing cybersecurity responsibilities into existing positions (European Commission, 2024). AI augments human capabilities, allowing for more robust and resilient cybersecurity infrastructures. The integration of AI into cybersecurity frameworks not only improves response times but also enhances the overall effectiveness of defense mechanisms, ensuring that digital infrastructures remain secure against an ever-evolving threat landscape, while partially offsetting the shortage of qualified personnel in the field.

3.2. AI as an attacker's tool

AI also provides malicious actors with powerful tools to enhance their attack strategies. Cybercriminals leverage AI to develop highly targeted and adaptive attacks, increasing their success rates and minimizing detection.

Personalized phishing campaigns are one example where AI excels. By analyzing large datasets, AI can craft convincing phishing messages tailored to individual recipients, making it more likely that they will fall for the scam. This precision targeting leverages data from social media (Github, 2016), email interactions (Eze & Shamir, 2024), and other sources to create messages that appear legitimate and relevant to the target.

Deepfake technology, powered by AI, poses significant risks through the creation of realistic but fake audio and video content (Suganthi et al., 2022). Cybercriminals use deepfakes for impersonation, fraud, and the dissemination of fake information. For example, deepfake videos can be used to manipulate public opinion or damage the reputation of individuals and organizations by making them appear to say or do things they never did.

In addition to deepfakes, there has been a resurgence of previously discontinued **criminal Large Language Model (LLM) services**, such as *WormGPT* and *DarkBERT*, which have now been enhanced with new features. These rebranded tools are being marketed alongside new offerings like *DarkGemini* and *TorGPT*, which boast multimodal capabilities, including the generation of images. However, it's important to note that many of these ChatGPT-like services being promoted on the dark web are primarily "jailbreak-as-a-service" platforms, being designed to manipulate commercial LLMs into bypassing their built-in restrictions, enabling them to produce unfiltered and potentially harmful responses to malicious queries (Paraschiv & Cîrnu, 2024; Trend Micro, 2024).

Advanced system probing, where AI algorithms systematically explore network defenses to identify vulnerabilities, is another area where AI can also help cyber criminals. Such algorithms can simulate numerous attack scenarios, finding and exploiting weaknesses that might be overlooked by human attackers. The ability of AI to simulate and adapt to a wide variety of defenses makes these tools particularly effective, allowing attackers to uncover and exploit vulnerabilities more efficiently. For instance, AI can automate the process of reconnaissance, analyzing network structures, predicting potential vulnerabilities, and even determining the optimal times to launch attacks based on gathered intelligence. This level of automation and precision enhances the effectiveness of cyberattacks, enabling malicious actors to carry out more sophisticated and hard-to-detect operations with minimal human intervention. As AI continues to evolve, its role in both enhancing and exploiting cybersecurity defenses will likely grow, which poses new challenges for organizations seeking to protect their systems from increasingly advanced threats (Jaber & Fritsch, 2023).

In addition to these methods, AI is increasingly being used to find bugs in software in order to exploit them. Cybercriminals deploy **AI-powered tools to scan codebases and software systems for vulnerabilities**, such as buffer overflows, SQL injection points, and other exploitable weaknesses. These tools can *automate the process of bug hunting*, making it faster and more efficient than traditional manual methods (Wilkins, 2024). By leveraging ML algorithms, these tools can *identify patterns and anomalies that indicate potential security flaws*, which can then be exploited to gain unauthorized access or disrupt various systems.

Thus, AI is not only a tool for defense but also a double-edged sword that enhances the capabilities of cybercriminals, enabling them to execute more complex, adaptive, and harder-to-detect attacks. This evolving threat landscape underscores the need for equally advanced AI-driven defenses to keep pace with these emerging challenges.

3.3. AI as a target

The proliferation of AI across diverse sectors has not only revolutionized technological capabilities, but also introduced a novel and complex attack surface in the cybersecurity landscape. As AI systems become increasingly integral to critical infrastructure and decision-making processes, they emerge as prime targets for malicious actors, being faced with unique challenges that transcend traditional security paradigms.

The fundamental vulnerability of AI systems stems from their inherent complexity and opacity. Neural networks, which form the backbone of many AI applications, operate as “black boxes” wherein the internal decision-making processes remain largely inscrutable, even to their developers (Bathaee, 2018). This lack of interpretability brings about challenges in identifying and mitigating potential security flaws.

Several key factors contribute to the vulnerability of AI systems as targets:

1. **Data dependency and integrity:** AI models rely heavily on vast datasets for training and operation. This dependency creates a critical vulnerability point, as the integrity and security of these datasets directly impact the model's performance and reliability. Adversaries may target these datasets through poisoning attacks, manipulating training data to induce biased or erroneous behaviors in the AI system (Baracaldo et al., 2017).

2. **Model extraction and intellectual property theft:** The valuable intellectual property embedded within AI models makes them attractive targets for theft and unauthorized replication. Adversaries may employ model extraction techniques to reconstruct proprietary models, compromising their competitive advantages and potentially exposing sensitive information encoded within the model's parameters (Tramèr et al., 2016).
3. **Integration into mission-critical systems:** The rapid adoption of AI across various sectors is leading to its implementation in increasingly critical applications. As AI systems transition from auxiliary roles to core operational components, they are being integrated into mission-critical systems that govern essential infrastructure (Laplante & Amaba, 2021), healthcare diagnostics, financial operations, and national security (Center for Security and Emerging Technology, 2020). This shift amplifies the potential impact of AI vulnerabilities. Adversaries targeting these AI-driven mission-critical systems could potentially disrupt vital services, compromise sensitive data, or manipulate critical decision-making processes (Linkov et al., 2023).
4. **Interpretability and skills shortage:** The "black box" nature of many AI systems complicates forensic analysis and incident response. When AI systems are compromised or exhibit unexpected behaviors, the lack of interpretability hinders efforts to identify the root cause and implement effective countermeasures (Rudin, 2019). Compounding this issue is the current and projected shortage of AI expertise in the workforce. For the foreseeable future, many organizations will likely face a significant skills gap, with IT and security professionals lacking comprehensive understanding of AI systems. (McDonald, 2024) This knowledge deficit exacerbates the challenges of securing and managing AI, as those responsible for overseeing these systems may struggle to fully grasp their complexities, vulnerabilities, and potential failure modes. The combination of opaque AI systems and a workforce which ill-equipped for managing them creates a perfect storm for security and accountability issues, potentially leaving critical AI implementations vulnerable to exploitation or mismanagement.

The targeting of AI systems carries significant implications beyond immediate security concerns. In critical infrastructure, healthcare, or autonomous systems, compromised AI could have far-reaching consequences, including physical harm, financial losses, or erosion of public trust in AI technologies. Traditional cybersecurity measures like firewalls, intrusion detection systems (IDSs), and encryption focus on external threats, but AI-specific attacks such as data poisoning and adversarial inputs exploit internal vulnerabilities within AI models themselves. For instance, data poisoning attacks, which introduce malicious data into training sets, often bypass detection by IDS, as these systems do not monitor the integrity of datasets during AI model training. This allows for subtle corruption of AI models, resulting in compromised decision-making long after the attack has occurred.

Similarly, adversarial input attacks exploit AI systems by slightly altering inputs, leading to incorrect classifications that traditional defenses fail to detect. For example, an image of a stop sign may be adversarially modified to be misclassified by an AI as a yield sign, without triggering any alerts from conventional security tools like firewalls or antivirus software.

Additionally, model extraction attacks, where attackers systematically query AI models to replicate their functionality, expose intellectual property and sensitive information encoded within the model's parameters. Encryption and access control methods, while effective for traditional data protection, do not prevent adversaries from reconstructing models through these techniques.

Finally, the opaque, "black box" nature of AI models complicates the detection of security breaches, as traditional security solutions rely on transparent system behavior to identify anomalies. The reactive nature of conventional defenses also falls short of addressing the constantly evolving adversarial threats targeting AI systems, which requires the implementation of more proactive, AI-specific security measures.

4. The evolving landscape of AI threats

The field of adversarial ML has undergone significant development since its theoretical foundations were established in 2006 (Barreno et al., 2006). This domain has witnessed a progressive increase in the complexity and diversity of attack vectors, reflecting the rapid advancement of AI technologies. Moreover, the fast-paced evolution of AI technologies is fundamentally reshaping the security landscape, introducing an entirely new paradigm of vulnerabilities that are evolving and proliferating at an unprecedented pace. The recent advent of LLMs serves as a prime example of this phenomenon. Within mere months from their widespread deployment, these models have not only revolutionized natural language processing capabilities but they have also given rise to an entirely unforeseen new class of attack vectors, such as prompt injection and jailbreaking techniques. Suddenly, semantics and wordplay have evolved from being tools limited to social engineering against humans to becoming potent attack methods against computer systems themselves, with carefully crafted phrases now capable of manipulating AI models into granting unauthorized access or divulging sensitive information to the attackers. This scenario exemplifies a critical shift in security: the potential for new AI paradigms to swiftly create expansive and previously unimagined attack surfaces, completely undetected by existing cybersecurity solutions.

Recent years have seen a marked escalation in the sophistication of proof-of-concept attacks. In 2022, a private company discovered that the exploitation of serialization vulnerabilities could lead to targeted ransomware attacks on AI models, highlighting the potential of malicious actors to compromise the integrity and availability of these systems (Wickens, Janus & Bonner, n.d.). By 2023, the threat landscape had further evolved to include data poisoning attacks on LLMs, demonstrating the potential of adversaries to manipulate these systems to propagate misinformation at scale (Qiang et al., 2024).

Despite these advancements in attack methodologies, the current prevalence of AI-targeted attacks remains relatively low (Grosse, 2024). This phenomenon can be attributed to several factors:

- **Fragmentation of AI ecosystems:** The heterogeneity of AI architectures, datasets, and deployment environments across different domains presents a significant challenge for attackers. This diversity necessitates the development of highly specialized attack vectors, increasing the complexity and resource requirements for potential adversaries.
- **Cautious integration in critical systems:** A judicious approach to implementing AI in mission-critical applications has moderated, so far, the proliferation of vulnerable AI models in high-stakes environments. Rigorous validation and testing protocols often serve as effective barriers against the deployment of susceptible systems.
- **Nascent understanding of AI vulnerabilities:** The cybercriminal community appears to be still in the early stages of comprehending and exploiting AI-specific vulnerabilities. The inherent complexity of these technologies necessitates substantial investment in research and development to formulate effective attack strategies.
- **Detection and attribution challenges:** The opaque nature of many AI models, often referred to as the "black box" problem, complicates the detection and attribution of AI-specific threats. This opacity can result in the obfuscation of malicious activities within broader cyber attack campaigns, impeding timely identification and response. The current detection methodologies and tools are likely insufficient to identify these novel threat vectors, creating a significant blind spot in cybersecurity defences. Consequently, adversaries may already be exploiting vulnerabilities in AI systems without leaving discernible traces, operating beneath the current detection thresholds.

However, this landscape is undergoing rapid transformation. It can be argued that the accelerating adoption of AI across diverse sectors, coupled with the increasing sophistication of adversarial techniques, is expected to precipitate a significant upsurge in both the frequency and severity of AI-targeted incidents. The integration of AI into critical infrastructure and high-value assets is likely to attract more advanced and persistent threat actors.

4.1. The heightened risks for critical infrastructures

While the overall incidence of AI-targeted attacks remains low, critical infrastructures, particularly in the energy sector, face significantly heightened risks. These infrastructures, essential for national security and economic stability, represent attractive targets for nation-state actors driven by geopolitical objectives rather than financial gain. The advanced capabilities and vast resources at the disposal of these actors enable them to exploit vulnerabilities in AI systems, regardless of their complexity or expertise required.

4.1.1. Historical context and evolving threats

Nation-state actors represent a distinct and particularly tremendous threat in the realm of cybersecurity, primarily when targeting critical infrastructures. They have historically targeted critical infrastructures to achieve various strategic goals, including service disruption, political coercion, and regional destabilization (Durojaye & Raji, 2022). In the context of AI integration, these threats are becoming more pronounced as AI systems are increasingly embedded in the management and control of critical infrastructure. AI technologies are now utilized for process optimization and automation in power grids, water supply systems, transportation networks, and other essential services. This integration, while enhancing operational efficiency and system resilience, simultaneously introduces new attack vectors susceptible to exploitation by adversaries.

4.1.2. Motivations and objectives of nation-state actors

In contrast to typical cybercriminal groups operating with an economic mindset focused on revenue generation and return on investment (ROI), attacks on critical infrastructure by nation-state actors are primarily motivated by broader geopolitical objectives. Their goals often involve destabilizing adversaries, exerting influence over geopolitical rivals, or gaining strategic advantages in international conflicts. The integration of AI into critical infrastructure has created new avenues for such actors to achieve these goals, with a potential for devastating consequences, such as:

- An AI system managing a **power grid** could be manipulated to induce widespread blackouts, leading to economic disruption and public safety risks (Sullivan & Kamensky, 2017). A targeted attack that subtly alters the AI's decision-making processes could cause imbalances in energy distribution, leading to cascading failures across the grid. Such an event could cripple industries, disrupt daily life, and create a state of chaos that weakens the affected nation’s economic and social stability.
- Attacks on AI systems controlling **transportation networks** could result in significant logistical challenges and safety hazards. For instance, if an adversary were to compromise the AI algorithms managing traffic lights, autonomous vehicles, or railway systems, the consequences could be immediate and severe. Malicious actors could disrupt traffic flow, leading to gridlocks in major urban areas, which would delay emergency response times, and cripple daily commutes. Additionally, tampering with the AI systems that guide autonomous vehicles could cause accidents, endangering the lives of passengers and pedestrians alike (Hamon et al., 2022).
- AI-driven **water treatment and distribution systems** could be compromised to alter chemical balances or disrupt supply, potentially causing public health crises and eroding trust in basic utilities. The disruption of water supply systems through AI manipulation could lead to water shortages, particularly in regions already facing water scarcity. Interrupting the distribution of clean water could cripple daily life, affecting households, businesses, and essential services such as healthcare services in hospitals. In agricultural areas, a compromised water supply could result in crop failures, exacerbating food insecurity and causing economic losses.
- Attacks on AI systems in the **healthcare sector** could lead to significant risks for patient safety, operational efficiency, and public health. For instance, if a nation-state actor were to compromise an AI-driven system used for diagnosing medical conditions, the

consequences could be dire. Malicious tampering with the AI's algorithms could result in incorrect diagnoses, leading to inappropriate treatments or delayed medical intervention. Such an attack could also target AI systems that manage hospital logistics, including the allocation of resources like intensive care unit beds, medical staff, or life-saving equipment such as ventilators. By disrupting these systems, attackers could create chaos in hospital operations, overwhelming medical facilities during times of high demand, such as during a pandemic or a natural disaster. This could lead to an otherwise preventable loss of life, as patients might be denied timely access to critical care.

The motivations behind these attacks extend beyond mere disruption. Nation-state actors may aim to weaken an adversary's economic and social fabric, provoke political instability, or force concessions in diplomatic or military negotiations. The consequences of such attacks extend beyond immediate operational disruptions, potentially having long-term implications for national security and public trust in essential services. Moreover, the psychological effects of such attacks, including fear, uncertainty, and a loss of confidence in the state's ability to protect its citizens, can be as damaging as the physical and economic impacts.

Further on, while the threat landscape is broad, not all adversarial attacks are equally feasible. The risk matrix in Table 1 evaluates the likelihood of each attack, accounting for factors such as the technical sophistication required, attack surfaces, and mitigation mechanisms.

Table 1. Risk matrix for adversarial AI attacks on critical infrastructures

Attack Type	**Feasibility (Low/Medium / High)**	**Impact on Infrastructure**	**Required Expertise**	**Potential Targets**	**Mitigation Techniques**	**Consequences of Mitigation Failure**
Data Poisoning	High	High	Medium to High	AI training datasets (health-care, energy, finance)	Data validation, robust learning algorithms, anomaly detection	Misclassifica-tions, faulty decision-making in critical systems
Model Extraction	Medium	Medium to High	High	Cloud-based AI systems, APIs	Query rate limiting, noise injection, model encryption	Intellectual property theft, replication of AI systems
Adversaria l Inputs	Medium	Medium to High	Low to Medium	Autono-mous vehicles, facial recogni-tion systems	Adversarial training, input sanitization, feature squeezing	Misclassifica-tion of critical objects (e.g., stop signs)
Privacy Inference	Medium	High	High	Health-care, financial, national security	Differential privacy, noise pertur-bation, secure multiparty computation	Privacy breaches, exposure of sensitive information
Backdoor Attacks	Low to Medium	High	High	AI in IoT devices, critical infra-structure control systems	Model pruning, backdoor detection, monitoring neuron activations	Uncontrolled system behavior, severe operational disruptions
Membership Inference	Medium	Medium	Medium	Federa-ted lear-	Differential privacy,	Re-identification of

				ning, persona-lized AI systems	adversarial regularization	individuals in training datasets
Model Inversion	Medium	Medium to High	High	Health-care, biometric systems	Differential privacy, output obfuscation	Recovery of sensitive input data, exposure of biometric details

As AI becomes more deeply embedded in these critical systems, the risks associated with such attacks will continue to escalate, emphasizing the need for comprehensive security measures tailored to these emerging threats.

5. Addressing the AI security deficit in critical infrastructures

The growing sophistication of adversarial attacks on artificial intelligence systems, juxtaposed with the current state of defensive mechanisms in critical infrastructures, reveals a significant security deficit. This disparity is exacerbated by the limited comprehension of adversarial machine learning among industrial practitioners. To address this challenge, a multifaceted approach involving policymakers, industry leaders, academic institutions, and government agencies is needed. Such collaboration should prioritize research initiatives, knowledge dissemination, and the development of advanced defensive technologies.

5.1. A comprehensive framework for AI security in critical infrastructures

To effectively address the evolving landscape of adversarial AI threats, a comprehensive and proactive approach is necessary. This approach should not operate in isolation but rather be seamlessly integrated into existing cybersecurity frameworks and protocols already employed by critical infrastructure entities. Table 2 encompasses strategic initiatives that should be considered as enhancements to current security practices, aligning with and augmenting established standards such as the US National Institute of Science and Technology (NIST)'s Cybersecurity Framework, IEC 62443 (International Electrotechnical Comission, 2021), European Union Agency for Cybersecurity (ENISA)'s Cybersecurity Act and Guidelines (European Commission. 2023) or sector-specific guidelines:

Table 2. Strategic initiatives for protecting critical infrastructures

No.	**Strategic Initiative**	**Key Actions**	**Objectives**
1.	Incident response protocols	Formulate AI-specific incident response methodologies and protocols.	Develop tailored incident respon-se plans for AI-related threats.
		Implement systems for rapid detection, containment, and recovery from AI-related security breaches.	Ensure quick identification and resolution of AI-specific security incidents.
		Integrate AI-driven tools into conventional incident response frameworks.	Enhance the precision and efficiency of incident response with AI-driven tools.
2.	Readiness and educational initiatives	Initiate targeted awareness campaigns across organizational hierarchies.	Foster a culture of AI security throughout the organization.
		Develop comprehensive educational programs on AI security risks and best practices.	Increase understanding of AI security issues among employees.
		Engage in public discourse to elevate awareness of AI security.	Raise public awareness about the importance and impact of AI security.
3.	Collaborative ecosystem development	Cultivate a collaborative ecosystem for knowledge exchange among stakeholders.	Promote the sharing of knowledge and best practices in AI security.

		Establish specialized forums and working groups dedicated to AI security.	Create platforms for ongoing discussions and dissemination of AI security information.
		Promote public-private partnerships to address AI security challenges.	Leverage expertise from various sectors to tackle AI security issues collectively.
4.	Supply chain integrity and procurement protocols	Implement rigorous procurement processes evaluating AI vendors' security practices.	Ensure that AI vendors meet strict security criteria.
		Ensure adherence to stringent security standards, including adversarial training and red team assessments.	Maintain high security standards across AI products and services.
		Conduct comprehensive monitoring of the AI supply chain to identify vulnerabilities.	Continuously assess and mitigate risks in the AI supply chain.
		Ensure secure MLOps practices, including CI/CD pipelines and regular security audits.	Safeguard the integrity and security of AI models and data pipelines.
5.	Access control and monitoring systems	Implement robust access control mechanisms for AI systems and data.	Prevent unauthorized access to AI systems and sensitive data.
		Deploy advanced authentication protocols for critical AI components.	Ensure only authorized personnel can access critical AI components.
		Utilize real-time monitoring solutions to detect anomalous activities.	Enable timely detection and response to potential threats in AI systems.
6.	Specialized training and skill development	Develop specialized training programs for cybersecurity professionals on AI security.	Equip security teams with the necessary skills to handle AI-specific threats.
		Incorporate practical simulations and hands-on training for real-world AI threat scenarios.	Prepare professionals to effectively respond to AI-related security challenges.
7.	Proactive threat identification and mitigation	Conduct regular adversarial testing and red team exercises.	Proactively identify and mitigate vulnerabilities in AI systems.
		Employ advanced threat-hunting techniques to discover emerging AI threats.	Stay ahead of new and evolving threats targeting AI systems.
		Continuously refine and update threat models based on the latest intelligence.	Keep threat models up-to-date and relevant to emerging attack methodologies.
8.	Continuous evaluation and system optimization	Implement a framework for regular auditing, testing, and updating of AI models.	Ensure ongoing security and integrity of AI systems.
		Establish protocols for prompt identification and remediation of vulnerabilities.	Enable swift response to detected vulnerabilities to minimize risk.
		Develop feedback mechanisms to incorporate insights from incidents and audits.	Use past experiences to improve future security practices and system designs.

While the proposed framework offers a comprehensive approach to AI security, it faces several challenges. For instance, the integration of AI-driven incident response systems may require significant investment in both financial and human resources. Additionally, continuous system evaluation can be hindered by the lack of trained AI security professionals and the complexity of AI models.

6. Conclusions

The integration of AI into critical infrastructure brings about both transformative opportunities and unprecedented challenges. While AI systems have the potential to enhance efficiency, optimize resource management, and improve decision-making processes across various sectors, they also introduce new vulnerabilities that adversaries, particularly nation-state actors, are increasingly poised to exploit.

This paper explored the evolving landscape of AI threats, from the foundational developments in adversarial ML to the sophisticated attack vectors emerging today. Adversarial threats such as dataset poisoning, evasion techniques, model stealing, privacy inference attacks, model backdoors, and even jailbreak prompts have highlighted the multifaceted nature of AI's vulnerability. Each of these threats poses significant risks not only to the integrity and reliability of AI systems, but also to the broader societal, economic, and geopolitical stability.

As AI continues to be embedded in high-value and mission-critical environments, the sophistication and frequency of attacks are expected to increase. This paper underscores the urgent need for a comprehensive and proactive approach to AI security. Key strategic initiatives, such as robust incident response protocols, specialized training programs, collaborative ecosystem development, rigorous supply chain management, and continuous evaluation of AI systems, must be prioritized to mitigate these evolving threats.

In conclusion, while AI holds immense promise for advancing critical infrastructure, its security cannot be an afterthought. The challenges outlined in this paper highlight the necessity of an adaptive security framework that not only addresses current vulnerabilities but is also capable of evolving alongside the technology it aims to protect. By fostering a culture of security awareness, investing in cutting-edge research, and building resilient AI systems, the future of critical infrastructure in an increasingly AI-driven world could be safeguarded.

REFERENCES

Baracaldo, N., Chen, B., Ludwig, H. & Safavi, J.A. (2017) Mitigating Poisoning Attacks on Machine Learning Models. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, 3 November 2017, Dallas, USA.* New York, Association for Computing Machinery. pp. 103-110.

Barreno, M., Nelson, B., Sears, R., Joseph, A.D. & Tygar, J.D. (2006) Can machine learning be secure? In: *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. 21-24 March 2006, Taipei, Taiwan.* New York, Association for Computing Machinery. pp. 16-25.

Bathaee, Y. (2018) The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology.* 31(2), 889-938.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G. & Roli, F. (2013) Evasion Attacks against Machine Learning at Test Time. In: Blockeel, H., Kersting, K., Nijssen, S. & Železný, F. (eds.) *Lecture Notes in Artificial Intelligence, vol. 8190* (*Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2013, 23-27 September 2013, Prague, Czech Republic*). Berlin, Heidelberg, Germany, Springer, pp. 387-402.

Bolcaș, R.-D. (2024) Generating FER models using ChatGPT. *Romanian Journal of Information Technology and Automatic Control [Revista Română de Informatică și Automatică]*. 34(2), 85-96. doi: 10.33436/v34i2y202407.

Center for Security and Emerging Technology (2020) *Artificial Intelligence and National Security*. https://cset.georgetown.edu/publication/artificial-intelligence-and-national-security [Accessed 12th August 2024].

Chen, C., Hong, H., Xiang, T. & Xie, M. (2024) Anti-Backdoor Model: A Novel Algorithm To Remove Backdoors in a Non-invasive Way. *IEEE Transactions on Information Forensics and Security*. 19, 7420-7434. doi: 10.1109/TIFS.2024.3436508.

Cinà, A.E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B.A., Oprea, A., Biggio, B., Pelillo, M. & Roli, F. (2023) Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *ACM Computing Surveys*. 55(13s), 294. doi: 10.1145/3585385.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P. & Hein, M. (2021) RobustBench: A standardized adversarial robustness benchmark. In: *Proceedings of the Thirty-fifth Annual Conference on Neural Information Processing Systems Conference (NeurIPS 2021), 6-14 December 2021*. pp. 12345-12356.

Dambra, S., Han, Y., Aonzo, S., Kotzias, P., Vitale, A., Caballero, J., Balzarotti, D. & Bilge, L. (2023) Decoding the Secrets of Machine Learning in Windows Malware Classification: A Deep Dive into Datasets, Features, and Model Performance. To be published in the *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. [Preprint] https://arxiv.org/pdf/2307.14657 [Accessed 14th May 2024].

Dobson, J. E. (2023) On reading and interpreting black box deep neural networks. *International Journal of Digital Humanities*. 5(2-3), 431-449. doi: 10.1007/s42803-023-00075-w.

Durojaye, H. & Raji, O. (2022) Impact of State and State Sponsored Actors on the Cyber Environment and the Future of Critical Infrastructure. To be published in *arXiv:2212.08036* [cs]. https://arxiv.org/abs/2212.08036.

European Commission (2024) *Cyberskills*. https://europa.eu/eurobarometer/surveys/detail/3176 [Accessed 12th August 2024].

European Comission. (2023) *The EU Cybersecurity Act.* https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-act [Accessed 12th August 2024].

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. & Song, D. (2018) *Robust Physical-World Attacks on Deep Learning Visual Classification*. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR 2018, 18-22 June 2018, ZSalt Lake City, USA.* New York, Computer Vision Foundation. pp. 1625-1634.

Eze, C.S. & Shamir, L. (2024) Analysis and prevention of AI-based phishing email attacks. To be published in *Electronics*. [Preprint] https://arxiv.org/abs/2405.05435 [Accessed 12 Aug. 2024].

Fredrikson, M., Jha, S. & Ristenpart, T. (2015) Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15, 12-16 October 2015, Denver, USA*. New York, NY, Association for Computing Machinery. doi: 10.1145/2810103.2813677.

Github. (2016) *GitHub - zerofox-oss/SNAP_R: A machine learning based social media pen-testing tool*. https://github.com/zerofox-oss/SNAP_R [Accessed 12th August 2024].

Grosse, K., Bieringer, L., Besold, T. R., Biggio, B. & Alahi, A. (2024) When Your AI Becomes a Target: AI Security Incidents and Best Practices. *Proceedings of the AAAI Conference on Artificial Intelligence*. 38(21), 23041-23046. doi: 10.1609/aaai.v38i21.30347.

Gu, T., Liu, K., Dolan-Gavitt, B. & Garg, S. (2019) BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*. 7, 47230-47244. doi: 10.1109/ACCESS.2019.2909068.

Gurve, M., Behera, S., Ahlawat, S. & Prasad, Y. (2024) MisGUIDE : Defense Against Data-Free Deep Learning Model Extraction. *arXiv* [cs.CR]. http://arxiv.org/abs/2403.18580 [Accessed 12th August 2024].

Hamon, R., Junklewitz, H., Sanchez Martin, J.I., Fernandez Llorca, D., Gomez Gutierrez, E., Herrera Alcantara, A.& Kriston, A. (2022) *Artificial Intelligence in Automated Driving: an analysis of safety and cybersecurity challenges*. https://publications.jrc.ec.europa.eu/repository/handle/JRC127189 [Accessed 12th August 2024].

Humphreys, D., Koay, A., Desmond, D. & Mealy, E. (2024) AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. *AI and Ethics*. 4, 791-804. doi: 10.1007/s43681-024-00443-4.

International Electrotechnical Comission. (2021) *Understanding IEC 62443*. https://www.iec.ch/blog/understanding-iec-62443 [Accessed 12th August 2024].

Jaber, A. & Fritsch, L. (2023) Towards AI-powered Cybersecurity Attack Modeling with Simulation Tools: Review of Attack Simulators. In: Barolli, L. (eds) *Advances on P2P,*

Parallel, Grid, Cloud and Internet Computing (Lecture Notes in Networks and Systems, vol 571). Cham, Switzerland, Springer, pp. 249-257.

Jayaraman, B. & Evans, D. (2022) Are Attribute Inference Attacks Just Imputation? To be published in the *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security.* [Preprint] https://arxiv.org/abs/2209.01292 [Accessed 12 Aug. 2024].

Kaur, R., Gabrijelčič, D. & Klobučar, T. (2023) Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions. *Information Fusion*. 97, 101804. doi: 10.1016/j.inffus.2023.101804.

Koh, P. W., Steinhardt, J. & Liang, P. (2022) Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*. 111, 1-47. doi: 10.1007/s10994-021-06119-y.

Krishna, K., Tomar, G. S., Parikh, A. P., Papernot, N. & Iyyer, M. (2020) Thieves on Sesame Street! Model Extraction of BERT-based APIs. *arXiv:1910.12366.* To be published in *ICLR 2020 - The Eighth International Conference on Learning Representations.* [Preprint] https://arxiv.org/abs/1910.12366 [Accessed 20 May 2024].

Laplante, P. & Amaba, B. (2021) Artificial Intelligence in Critical Infrastructure Systems. *Computer*. 54(10), 14-24. doi: 10.1109/mc.2021.3055892.

Leu, D.M., Udroiu, C., Raicu, G.M., Gârban, H.N. & Șcheau, M.C. (2023) Analysis of some case studies on cyberattacks and proposed methods for preventing them. *Romanian Journal of Information Technology and Automatic Control [Revista Română de Informatică și automatică]*. 33(2), 119-134. doi: 10.33436/v33i2y202309.

Levitt, K. (2023) *How Is AI Used in Fraud Detection?* [online] NVIDIA Blog. https://blogs.nvidia.com/blog/ai-fraud-detection-rapids-triton-tensorrt-nemo/ [Accessed 14th May 2024].

Linkov, I., Stoddard, K., Strelzoff, A., Galaitsi, S.E., Keisler, J., Trump, B.D., Kott, A., Bielik, P. & Tsankov, P. (2023) Toward Mission-Critical AI: Interpretable, Actionable, and Resilient AI. In: *Proceedings of the 15th International Conference on Cyber Conflict: Meeting Reality (CyCon), 30 May - 2 June 2023, Tallinn, Estonia.* IEEE. pp. 181-197.

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K. & Liu, Y. (2023). *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study*. [online] https://arxiv.org/abs/2305.13860.

McDonald, C. (2024) *AI and cyber skills worryingly lacking, say business leaders*. https://www.computerweekly.com/news/366593173/AI-and-cyber-skills-worryingly-lacking-say-business-leaders [Accessed 12th August 2024].

Motie, S. & Raahemi, B. (2023) Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications*. 240, 122156. doi: 10.1016/j.eswa.2023.122156.

Paraschiv, E.-A & Cîrnu, C.E. (2024) Between the Lines: Generating, Detecting and Defending against Textual Deepfakes. *Romanian Cyber Security Journal*. 6(1), 3-13. doi: 10.54851/v6i1y202401.

Qiang, Y., Zhou, X., Zade, S.Z., Roshani, M.A., Zytko, D. & Zhu, D. (2024). Learning to Poison Large Language Models During Instruction Tuning. To be published in the *Proceedings of the ACL ARR 2024 Conference*. [Preprint] https://arxiv.org/abs/2402.13459 [Accessed 12th August 2024].

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 1(5), 206-215. doi: 10.1038/s42256-019-0048-x.

Shen, X., Chen, Z., Backes, M., Shen, Y. & Zhang, Y. (2024) 'Do Anything Now': Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. To be published in *ACM CCS 2024 Conference*. [Preprint] https://arxiv.org/pdf/2308.03825 [Accessed 12th August 2024].

Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017) Membership Inference Attacks against Machine Learning Models. To be published in the *Proceedings of the IEEE Symposium on Security and Privacy, 2017*. [Preprint] https://arxiv.org/abs/1610.05820 [Accessed 12th August 2024].

Suganthi, S., Ayoobkhan, M.U.A., Krishna Kumar, V., Venkatachalam, K., Bacanin, N., K, V., Hubálovský, Š. & Trojovský, P. (2022) Deep learning model for deep fake face recognition and detection. *PeerJ Computer Science*. 8, e881. doi: 10.7717/peerj-cs.881.

Sullivan, J. E. & Kamensky, D. (2017) How cyber-attacks in Ukraine show the vulnerability of the U.S. power grid. *The Electricity Journal*. 30(3), 30-35. doi: 10.1016/j.tej.2017.02.006.

Stanciu, A. (2023) Data Management Plan for Healthcare: Following FAIR Principles and Addressing Cybersecurity Aspects. A Systematic Review using InstructGPT. *Romanian Cyber Security Journal*. 5(1), 23-43. doi: 10.54851/v5i1y202303.

Takyar, A. (2023) *AI in anomaly detection: Use cases, methods, algorithms and solution*. LeewayHertz - AI Development Company. https://www.leewayhertz.com/ai-in-anomaly-detection [Accessed 12th August 2024].

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K. & Ristenpart, T. (2016) Stealing Machine Learning Models via Prediction APIs. To be published in *25th USENIX Security Symposium*. [Preprint] https://doi.org/10.48550/arxiv.1609.02943 [Accessed 13th August 2024].

Trend Micro. (2024) *AI-Powered Deepfake Tools Becoming More Accessible Than Ever*. https://www.trendmicro.com/en_us/research/24/g/ai-deepfake-cybercrime.html [Accessed 12th August 2024].

Vassilev, A., Oprea, A., Fordyce, A. & Anderson, H. (2024) *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. National Institute of Standards and Technology - U. S. Department of Commerce. Report number: NIST AI 100-2e2023. https://csrc.nist.gov/pubs/ai/100/2/e2023/final [Accessed 12th August 2024] National Institute of Standards and Technology - U. S. Department of Commerce. Report number: NIST AI 100-2e2023. https://csrc.nist.gov/pubs/ai/100/2/e2023/final [Accessed 12th August 2024].

Vast, R., Sawant, S., Thorbole, A. & Badgujar, V. (2021) Artificial Intelligence based Security Orchestration, Automation and Response System. In: *2021 6th International Conference for Convergence in Technology (I2CT), 2-4 April 2021, Maharashtra, India*. IEEE. pp. 1-5.

Wickens, E., Janus, M. & Bonner, T. (n.d.) *Weaponizing ML Models with Ransomware.* https://hiddenlayer.com/research/weaponizing-machine-learning-models-with-ransomware/ [Accessed 12th August 2024].

Wilkins, A. (2024) *Hackers are using AI to find software bugs - but there is a downside*. https://www.newscientist.com/article/2433247-hackers-are-using-ai-to-find-software-bugs-but-there-is-a-downside [Accessed 12th August 2024].

Luca SAMBUCCI has a vast experience in cybersecurity which spans over several decades, starting from the late 1980s with the analysis of early computer viruses, until today's work in AI security. He holds a Business Analytics specialization from Wharton School of the University of Pennsylvania and is an IBM-certified AI Developer. His career includes consulting for the Italian Government, counselling the EU's Joint Research Centre and the European Defence Agency with regard to AI for critical infrastructure protection, as well as working with several private companies,

such as Telespazio from the Leonardo Group, on topics that include cybersecurity and AI governance. He is actively involved with the major industry associations and is currently a member of the Industry board of the prominent Italian Association for Artificial Intelligence. He co-authored several books on AI including "La società dei robot" (Mondadori, 2022) and "ParadoXa: IA" (Mimesis, 2023).

Luca SAMBUCCI are o experiență vastă în domeniul securității cibernetice, care se întinde pe parcursul a mai multor decenii, începând cu sfârșitul anilor 1980, când a analizat primii viruși informatici, până la activitatea sa actuală în securitatea IA. Deține o specializare în Business Analytics de la Wharton și este dezvoltator IA certificat de IBM. Cariera sa include consultanță pentru Guvernul Italiei, consilierea Centrului Comun de Cercetare al UE și a Agenției Europene de Apărare în protecția infrastructurii critice folosind IA, precum și colaborarea cu diverse companii private, cum ar fi Telespazio din grupul Leonardo, pe teme ce includ securitatea cibernetică și guvernanța IA. Este activ implicat în principalele asociații din industrie și face parte din consiliul de industrie al prestigioasei Asociații Italiene pentru Inteligență Artificială. Este co-autor al mai multor cărți despre IA, inclusiv "La società dei robot" (Mondadori, 2022) și "ParadoXa: IA" (Mimesis, 2023).

Elena-Anca PARASCHIV is a Scientific Researcher at the „Software Engineering and Complex Systems” Department of the National Institute for Research and Development in Informatics - ICI Bucharest and a Ph.D. student at the Doctoral School of Electronics, Telecommunications and Information Technology, of the National University for Science and Technology Politehnica Bucharest (NUSTPB). She graduated from the Faculty of Medical Engineering of NUSTPB and she holds a Master’s Degree in „Intelligent systems and computer vision” from the Faculty of Electronics, Telecommunications and Information Technology, NUSTPB. Her research fields and topics of interest include artificial intelligence applications, particularly Large Language Models (LLMs) and computer vision, with a focus on the healthcare domain. Additionally, her research extends to natural language processing, human-computer interaction, and educational technologies, aiming to advance adaptive, secure, and ethical AI systems.

Elena-Anca PARASCHIV este cercetător științific în cadrul Departamentului „Ingineria Software și a Sistemelor Complexe” din cadrul Institutului Național de Cercetare-Dezvoltare în Informatică - ICI București și student-doctorand în cadrul Școlii Doctorale de Electronică, Telecomunicații și Tehnologia Informației, Universitatea Națională de Știință și Tehnologie POLITEHNICA București (UNSTPB). A absolvit Facultatea de Inginerie Medicală din cadrul UNSTPB și deține o diplomă de master în specializarea „Sisteme inteligente și vedere artificială” din cadrul Facultății de Electronică, Telecomunicații și Tehnologia Informației, UNSTP. Domeniile sale de cercetare și interesele sale includ aplicațiile inteligenței artificiale, în special modelele lingvistice de mari dimensiuni (Large Language Models - LLMs) și vederea computațională, cu un accent pe domeniul sănătății. De asemenea, cercetările sale se extind în procesarea limbajului natural, interacțiunea om-computer și tehnologiile educaționale, având ca scop avansarea sistemelor IA adaptive, securizate și etice.