Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8, 1452-1463
2025
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v9i8.9643

Securing the future: Integrating AI safety into cybersecurity frameworks

Kamila Jabbarova[1*]
[1]Academy of Public Administration under the President of the Republic of Azerbaijan; cabbarovakamila@gmail.com (K.J.)

Abstract: Artificial intelligence (AI) is rapidly transforming critical sectors such as healthcare, finance, transportation, and national security. However, as AI systems become increasingly integrated into these essential areas, new vulnerabilities and risks emerge. This article explores the imperative of integrating AI safety into existing cybersecurity frameworks to address these challenges. It argues that AI safety must be recognized as a core component of modern cybersecurity due to the potential misuse of AI systems, their susceptibility to adversarial attacks, and the inherent risks of autonomous decision-making. The article outlines a proactive approach, emphasizing key strategies such as securing AI model integrity, protecting data privacy, continuous threat monitoring, and establishing ethical AI use policies. Additionally, it highlights the importance of cross-disciplinary collaboration between AI developers, cybersecurity experts, and policymakers to develop comprehensive standards and best practices. By embedding AI safety into cybersecurity protocols, stakeholders can protect AI technologies and the sensitive data they handle, ensuring their responsible and secure deployment across society. Ultimately, this integrated approach will enable the harnessing of AI's potential while mitigating the risks it poses to digital and societal infrastructure.

Keywords: *Adversarial attacks, AI governance, AI misuse, AI safety, AI System integrity, Anomaly detection, Autonomous systems, Cybersecurity, Data integrity, Ethical AI.*

1. AI in the Age of Cybersecurity: Integrating Safety into Critical Systems

Artificial Intelligence (AI) is revolutionizing sectors ranging from healthcare to finance, transportation, and communication. It enhances efficiency and innovation by automating tasks like diagnosing diseases, detecting fraud, and managing autonomous systems. However, as AI systems evolve and become more complex, their integration into critical systems introduces new cybersecurity risks. As AI increasingly handles sensitive data, makes autonomous decisions, and interacts with vital infrastructure, the potential for misuse and exploitation grows. This paper argues that AI safety must be recognized as a core component of cybersecurity efforts. By addressing the vulnerabilities of AI systems, we can ensure the integrity, security, and privacy of the data and processes they manage, while also maintaining trust in the technology. Framing AI safety within the broader cybersecurity paradigm allows stakeholders to develop comprehensive strategies that mitigate the risks AI introduces, while harnessing its potential for societal benefit.

Since AI in its current advanced applications is still relatively new, the full spectrum of vulnerabilities and potential threats has yet to fully materialize. Much of our understanding of AI-related threats is therefore based on theoretical scenarios and assumptions about how these systems might be exploited. These scenarios often draw from current knowledge of cyberattacks on traditional systems, as well as early incidents involving AI, such as adversarial attacks or data poisoning. While concrete examples of AI-targeted attacks remain limited, the anticipated vulnerabilities align with known cyber risks, providing a strong foundation for identifying future threats. As AI continues to evolve, the need for proactive cybersecurity measures becomes increasingly clear, ensuring that emerging vulnerabilities are addressed before they are fully realized.

History: Received: 16 June 2025; Revised: 11 August 2025; Accepted: 14 August 2025; Published: 25 August 2025
* Correspondence: cabbarovakamila@gmail.com

One of the most pressing concerns surrounding AI is its safety. Traditionally, discussions of AI safety have focused on ethical challenges, such as bias in decision-making, job displacement, and the fear of AI systems operating beyond human control. While these are significant issues, AI safety is deeply intertwined with cybersecurity—a field dedicated to protecting systems, networks, and data from unauthorized access or attacks. AI safety, therefore, should not be considered in isolation but as an integral part of broader cybersecurity efforts. Protecting AI systems from manipulation, sabotage, or unintended consequences is paramount, especially as these technologies influence critical decision-making processes in industries such as healthcare, finance, and transportation.

This article argues that AI safety must be recognized as a fundamental cybersecurity issue, demanding immediate attention from a wider range of stakeholders, including governments, corporations, and individuals. Several factors, such as AI's vulnerability to cyberattacks and its role in managing sensitive data, clearly illustrate why AI safety cannot be separated from broader cybersecurity concerns.

1.1. Vulnerability of AI Systems to Cyberattacks

AI systems are particularly exposed to a wide range of cyber threats. From data manipulation and adversarial attacks to the theft of intellectual property in the form of AI models, these systems can be exploited in ways that have significant consequences. For instance, healthcare AI systems could be manipulated to produce incorrect diagnoses, while AI-driven financial systems could be sabotaged, destabilizing entire markets. These vulnerabilities highlight the urgent need for robust cybersecurity measures tailored to defend AI systems from these and other forms of manipulation.

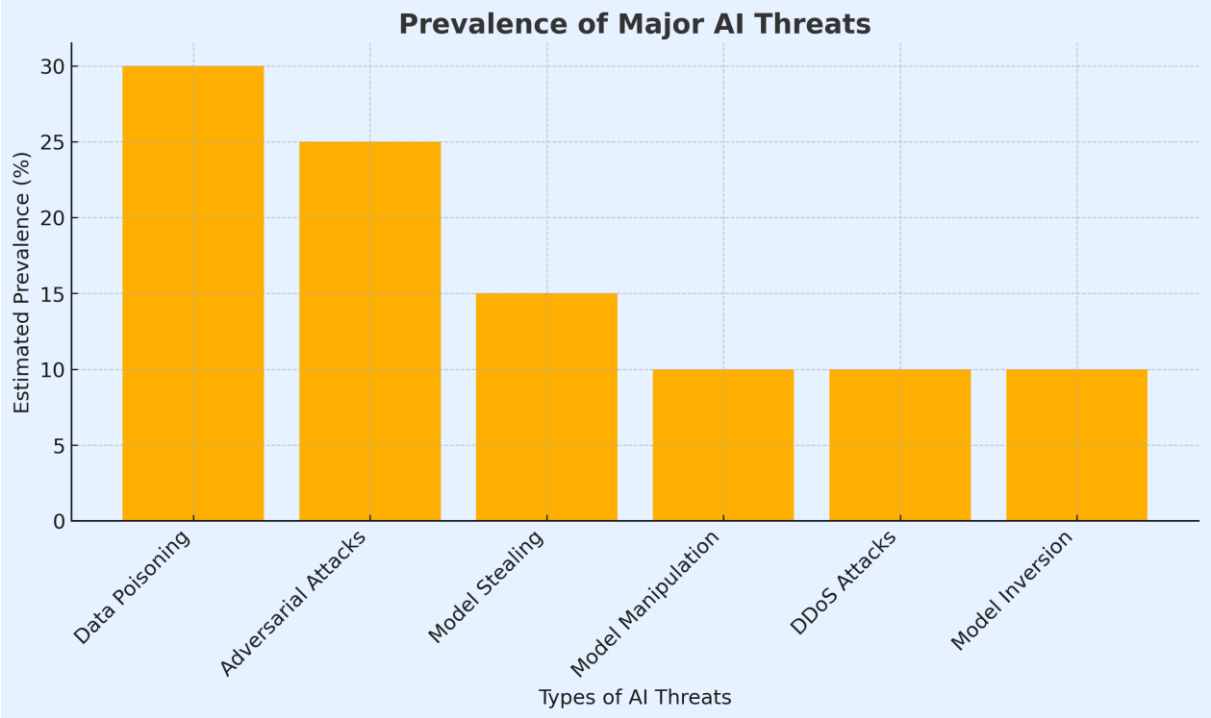

Figure 1.
Estimated Prevalence of Key AI-Related Cyber Threats.

The chart illustrates the relative estimated prevalence of key AI-related cybersecurity threats. Data poisoning and adversarial attacks represent the most significant concerns, followed by model theft,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

evasion tactics, and DDoS targeting AI infrastructure. This figure presents an illustrative synthesis of threat types based on analysis of current academic literature and expert discussions [1-4].

As shown in Figure 1, these attack vectors pose varying levels of risk, reinforcing the necessity of integrating AI-specific protections into broader cybersecurity strategies.

AI systems are also heavily dependent on large, sensitive datasets to function effectively. These datasets often contain personal health information, financial records, or critical infrastructure data. Unauthorized access to such data can result in severe privacy breaches, financial losses, and even national security threats. As AI becomes more integrated into critical infrastructure, the risk of breaches grows, with a single compromised system potentially triggering cascading failures across entire sectors or populations.

Another growing concern is the increasing autonomy of AI systems. From self-driving cars to AI-powered healthcare diagnostics, these systems often operate with minimal human oversight, making autonomous decisions based on complex algorithms. Without proper security, these systems could be hijacked or manipulated to act contrary to their intended purpose. Ensuring the integrity of AI decision-making processes aligns directly with cybersecurity objectives, as protecting these systems ensures their reliable, ethical, and safe operation.

Moreover, malicious actors can exploit AI for harmful purposes. While AI holds great promise for societal advancement, it can also be weaponized. Cybercriminals and nation-states are already using AI to launch sophisticated cyberattacks, such as automated phishing campaigns and realistic deepfakes for disinformation purposes. The potential misuse of AI further reinforces the need to treat AI safety as a core component of modern cybersecurity efforts.

As AI continues to become embedded in vital infrastructure, concerns over its vulnerability to cyberattacks will only increase. The complexity of AI systems, combined with the diversity of cyberattacks they face, underscores the need for enhanced cybersecurity measures to protect these systems. Below are some of the most prominent types of attacks AI systems face, along with examples of their potential impact.

1.2. Data Poisoning: A Growing Threat

One of the most pressing threats AI systems face is data poisoning, where attackers introduce corrupted or misleading information into the AI model's training data. Since AI models rely heavily on this data to learn and make predictions, tampering with it can compromise the entire system. In the financial sector, for example, a system designed to detect fraudulent transactions could be fed poisoned data that causes fraudulent actions to be classified as legitimate. This would allow attackers to carry out fraud undetected, resulting in significant financial and reputational damage [1]. Similarly, in healthcare, an AI system trained to identify early signs of diseases could be misled by poisoned data, resulting in missed diagnoses and putting patients' health at risk [2].

The threat posed by data poisoning is amplified by the increasing complexity and size of the datasets used in modern AI systems. With vast amounts of data being sourced from multiple providers, the likelihood of introducing tampered or compromised data grows. Attackers do not need direct access to the core of an AI system to poison its data—they can exploit vulnerabilities at different points in the data supply chain [3].

To mitigate the risks of data poisoning, several countermeasures must be integrated into AI training and deployment processes. These include implementing more rigorous data validation processes, regularly auditing data sources, and using anomaly detection algorithms during the training phase to identify potentially compromised data [4]. Additionally, maintaining diversified training datasets and using federated learning approaches can help protect AI systems by minimizing the risk of relying too heavily on any single data source.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

1.3. Adversarial Attacks: Manipulating AI Systems

Adversarial attacks pose a significant threat to the integrity of AI systems, where even small, seemingly insignificant changes to input data can lead AI models to make incorrect decisions. These subtle alterations, often imperceptible to human observers, can dramatically distort an AI's output. For instance, an autonomous vehicle relying on AI for navigation could be misled by slightly modified traffic signs, causing the system to misinterpret a stop sign as a speed limit sign, potentially leading to dangerous accidents. This form of attack highlights the vulnerability of AI systems, particularly in environments where real-time decisions have significant consequences.

Adversarial attacks can also compromise AI systems responsible for security protocols. For example, adversarial manipulation can deceive facial recognition systems, leading them to misidentify individuals. This type of attack is particularly dangerous in high-security locations such as airports, government facilities, and critical infrastructure, where a single failure in identification could lead to severe security breaches [1].

Moreover, these attacks do not require extensive access to the system. Small perturbations or alterations in input data can result in significant consequences without triggering alarms or drawing attention to the tampered data. Attackers may manipulate seemingly benign data, such as slight changes in image pixels or sound wave patterns, to deceive AI algorithms into delivering incorrect outputs, leading to outcomes that could range from mild disruptions to catastrophic failures. The healthcare sector, for example, could be vulnerable if an adversarial attack causes an AI-based diagnostic tool to incorrectly assess patient data, resulting in delayed or incorrect treatment plans [4].

1.4. Model Stealing Attacks

Model stealing attacks present a different kind of threat. Here, attackers attempt to replicate an AI model by querying it with various inputs and analyzing the outputs. This allows them to effectively recreate the model without accessing the underlying code or training data. For instance, a competitor could replicate a sophisticated AI model used for personalized recommendations in e-commerce, undermining the original company's competitive advantage. By stealing the model, they could either bypass proprietary restrictions or sell the replicated version, gaining an unfair market edge.

1.5. Evasion Attacks

Evasion attacks target AI models already in use, altering input data in real time to evade detection. This tactic is particularly dangerous in cybersecurity, where AI systems are deployed to detect and block cyberattacks. By modifying malware or malicious actions in ways that avoid recognition by the AI, attackers can bypass even sophisticated security measures, leaving systems vulnerable to further compromise.

Another significant concern is the deliberate manipulation of AI models by attackers who gain access to the AI's operational environment, whether through external breaches or insider threats. Malicious actors can alter AI models to distort decision-making processes for harmful purposes. In the financial sector, compromising AI systems can allow adversaries to manipulate stock prices or markets [1]. Similarly, AI models used in cybersecurity can be exploited through machine learning manipulation to create mutating malware, making detection difficult and compromising network security. In healthcare, compromised AI system managing hospital resources could divert crucial resources from high-need areas, causing critical delays in care. Additionally, Distributed Denial of Service (DDoS) attacks on AI infrastructure pose an escalating threat. Real-time AI applications, such as those in smart cities or cybersecurity, require significant computational power. A DDoS attack could flood these systems with traffic, leading to widespread disruption, such as gridlock in smart cities or paralyzed cybersecurity defenses [1]. Similarly, AI-driven cybersecurity systems responsible for network monitoring could be rendered inoperable, exposing the system to further exploitation and leaving it vulnerable to more sophisticated attacks.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

These examples underscore the wide range of threats AI systems face as they become increasingly integrated into vital sectors. The complexity and sophistication of these attacks highlight the need for enhanced cybersecurity measures specifically designed to protect AI systems. As AI continues to evolve and take on more critical roles in society, understanding and addressing these vulnerabilities will be essential in preventing large-scale disruptions and ensuring the safe operation of AI technologies across industries.

2. Data Security and Privacy Concerns

Artificial Intelligence (AI) systems are fundamentally data-driven. The effectiveness and accuracy of AI largely depend on the volume, variety, and quality of the data they are trained on. These datasets often contain sensitive personal, financial, and health-related information, making the security of these data assets crucial. In sectors such as healthcare, finance, and national security, where data is not only valuable but also highly confidential, breaches can have far-reaching consequences both at an individual and societal level. As AI becomes increasingly integrated into these critical sectors, the risk of data breaches and privacy violations grows significantly.

2.1. The Sensitivity of AI Data

The nature of the data AI systems rely on is what makes security so paramount. For instance, in the healthcare sector, AI systems used to assist in diagnosis or treatments rely on large volumes of medical records that include personal health information (PHI). A breach of this data can lead to privacy violations, identity theft, and insurance fraud. Similarly, AI systems in finance analyze personal financial information, and breaches can result in devastating financial loss and erode trust in financial institutions.

2.2. Model Inversion Attacks: A New Threat

Beyond traditional data breaches, AI introduces new risks like model inversion attacks. In these attacks, adversaries exploit an AI model's outputs to infer sensitive information from the training data, even without direct access to it. This could lead to the reconstruction of personal data, such as medical conditions or financial information, making it a significant privacy threat.

2.3. Integrating Explainable AI (XAI) to Enhance Trust

To address the growing concerns around AI data privacy, Explainable AI (XAI) offers a valuable solution by providing transparency in how AI models handle sensitive data. According to Sarker, et al. [5] XAI not only improves the trustworthiness of AI systems but also enhances data security by enabling users to understand how decisions are made, reducing the risk of unintentional data exposure. XAI ensures that even in cases where AI models are processing sensitive data, the reasoning behind their decisions can be monitored and verified without exposing confidential information.

2.4. AI and Data Breach Amplification

AI systems process large-scale data, amplifying the consequences of a breach. A single compromised system could expose millions of records across multiple institutions. AI models also aggregate data from multiple sources, further increasing the potential damage from a single breach.

2.5. Privacy-Preserving AI Technologies

To mitigate these risks, technologies such as differential privacy and federated learning are essential. Differential privacy ensures that AI systems can learn from data while minimizing the risk of exposing individual information. Federated learning, on the other hand, allows AI models to be trained across decentralized devices, reducing the risk of a large-scale breach by distributing the data. Explainable AI complements these methods by making it easier to audit AI systems and ensure they handle data in compliance with privacy standards [5].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

However, the rapid pace of AI development often outstrips the regulatory landscape, and new laws will likely be needed to address the specific risks AI poses to data security. Governments, corporations, and regulatory bodies need to collaborate to create policies that reflect the unique challenges AI introduces to data privacy and security.

3. The Role of Cybersecurity in AI Data Protection

Addressing the security and privacy concerns surrounding AI systems requires the integration of AI safety into existing cybersecurity frameworks. Traditional cybersecurity measures, such as encryption, access control, and regular audits, remain crucial but must be adapted to address the specific vulnerabilities AI systems present. For instance, encryption protocols may need to account for the unique structure of AI models and the massive datasets they process, while access control policies must focus on restricting who can modify or access AI training data to prevent model manipulation or poisoning attacks. Furthermore, cybersecurity professionals must anticipate new risks posed by AI models, such as model inversion attacks and potential data leakage through model outputs.

Beyond technical solutions, organizations must prioritize robust data governance. This includes developing clear policies and protocols for managing the data that feeds into AI systems, limiting access to sensitive information, and ensuring compliance with regulatory standards such as the General Data Protection Regulation (GDPR) and HIPAA. Regular security assessments and the implementation of advanced monitoring tools—such as real-time anomaly detection—are essential to identify vulnerabilities early; especially as adversarial attacks and data manipulation threats become more sophisticated.

Finally, the integrity of AI systems aligns directly with the foundational principles of cybersecurity: protecting the confidentiality, integrity, and availability of systems and data. Just as cybersecurity focuses on safeguarding information and ensuring systems function as intended, AI safety involves ensuring that AI models are accurate, secure, and free from manipulation. If AI integrity is compromised, the consequences can be catastrophic, particularly in sectors like healthcare, law enforcement, and national defense. In these environments, an attack on AI systems could lead to life-threatening misdiagnoses, wrongful identifications, or even national security breaches.

In this way, integrating AI safety into the broader cybersecurity paradigm is not only essential for addressing current risks but also critical for evolving alongside new threats as AI continues to influence key sectors.

4. Autonomy in AI Decision-Making and the Demand for Security

The increasing autonomy of artificial intelligence (AI) in decision-making processes is one of the most significant technological advancements today. AI systems are now being used to make critical decisions in fields such as healthcare, finance, transportation, and national security. For example, AI models are employed to autonomously diagnose diseases, approve financial transactions, and manage autonomous vehicles. These systems operate with minimal human intervention, often relying on complex algorithms and vast amounts of data. While this autonomy offers numerous opportunities for efficiency and innovation, it also raises serious security concerns.

When AI systems operate autonomously, their decisions can have far-reaching consequences. In healthcare, an AI system that incorrectly diagnoses a patient or recommends inappropriate treatments could endanger lives. Similarly, in the financial sector, a compromised AI-powered trading system could lead to significant financial losses or market instability. Autonomous vehicles, controlled by AI, could be manipulated to cause accidents if their systems are breached. As AI continues to be integrated into critical infrastructure and essential services, the risks associated with autonomous AI decision-making become more severe.

The autonomy of these systems makes them attractive targets for cyberattacks. Threats such as data poisoning, adversarial attacks, and model manipulation are particularly concerning in the context of autonomous AI. In data poisoning, attackers manipulate the data used to train the AI, leading to

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

flawed decision-making. Adversarial attacks involve introducing subtle alterations to input data, deceiving the AI into making incorrect decisions. Model manipulation occurs when the AI's underlying algorithm is tampered with, altering its decision-making logic.

The demand for robust security in AI systems is, therefore, crucial to ensuring their safe and reliable operation. Security measures must address both the integrity and confidentiality of the data used by AI systems, as well as the integrity of the decision-making process itself. This involves implementing encryption, access control, and real-time monitoring to detect and mitigate potential threats. Furthermore, developing AI systems with built-in safeguards, such as fail-safe mechanisms and human oversight, is essential to prevent catastrophic failures.

5. The Misuse of AI as a Tool for Cyber Threats

The rise of artificial intelligence (AI) has brought with it significant advancements across a range of sectors, from healthcare and finance to transportation and education. However, alongside these benefits comes the potential for AI to be misused as a tool for cyber threats. While AI offers the ability to enhance security systems, automate processes, and improve decision-making, it also provides malicious actors with new capabilities to conduct more sophisticated and destructive cyberattacks. As AI continues to evolve, it is crucial to understand how it can be misappropriated for harmful purposes and the corresponding security measures needed to mitigate these risks.

One of the primary concerns surrounding the misuse of AI lies in its ability to automate and scale cyberattacks. Traditional cyberattacks, such as phishing or malware distribution, require a significant degree of manual effort and coordination. However, with AI, these processes can be automated, making attacks more efficient and difficult to detect. AI-driven phishing attacks, for instance, can generate highly personalized and convincing messages based on data mining and machine learning. By analyzing vast amounts of data, AI systems can craft phishing emails that are more likely to deceive recipients, increasing the chances of successful attacks. This level of customization makes it increasingly difficult for individuals and organizations to recognize and defend against such threats.

Another area where AI can be misused is in the development of *deepfakes*, a form of AI-generated content that can create hyper-realistic images, videos, and audio clips. While deepfakes have legitimate applications in entertainment and media, they have also been weaponized for malicious purposes. Cybercriminals can use deepfakes to impersonate individuals, spread disinformation, and manipulate public opinion. In the context of cyber threats, deepfakes can be used to deceive individuals into believing they are communicating with a trusted source, leading to the disclosure of sensitive information or even financial fraud. In the political arena, deepfakes have the potential to destabilize democracies by spreading disinformation during elections or undermining the credibility of public officials.

AI also plays a role in automated hacking techniques. Machine learning algorithms can be trained to identify vulnerabilities in software systems or networks more quickly than human hackers. Once these vulnerabilities are discovered, AI can be programmed to exploit them at scale. This can result in faster, more widespread attacks that can compromise multiple systems simultaneously. For instance, AI-driven malware can autonomously evolve and adapt to evade detection by traditional security measures, making it harder for cybersecurity teams to defend against such threats. Moreover, AI can be used to optimize Distributed Denial of Service (DDoS) attacks, which flood a network with traffic, rendering it inaccessible. By using AI, attackers can optimize these attacks by identifying the weakest points in a system and directing resources accordingly.

Social engineering attacks, which rely on manipulating human behavior, can also be enhanced with AI. By analyzing vast amounts of personal data available online, AI can craft highly targeted attacks that exploit individual weaknesses. For example, AI can predict which individuals are more likely to fall for certain scams based on their social media activity, search history, or communication patterns. This allows attackers to craft more effective social engineering schemes that increase the likelihood of

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

success. In addition, AI-driven bots can be deployed to interact with users in real-time, mimicking human behavior to gain trust and extract sensitive information.

The misuse of AI in the context of cyber threats extends beyond individuals and businesses. State-sponsored cyberattacks are increasingly incorporating AI to enhance their effectiveness. Governments and military organizations may use AI to conduct espionage, sabotage critical infrastructure, or disrupt the operations of rival states. AI can be employed to automate surveillance, track the movements and communications of targets, and carry out sophisticated disinformation campaigns. In the realm of warfare, AI-driven autonomous weapons systems represent another potential area for misuse, as they can be programmed to make decisions without human oversight, raising ethical and security concerns.

To address the misuse of AI as a tool for cyber threats, it is essential to adopt a multi-faceted approach to security. Organizations must invest in AI-driven cybersecurity systems that can detect and respond to AI-powered threats. This includes the use of machine learning algorithms to identify patterns of malicious behavior, detect anomalies in network traffic, and respond to potential attacks in real-time. Moreover, developing legal frameworks and regulations specific to AI misuse is critical in ensuring accountability and transparency in the deployment of AI technologies. Governments, corporations, and researchers must collaborate to create guidelines and standards that prevent the malicious use of AI while promoting its positive applications.

And while AI offers immense potential for advancing technology and improving security, it also presents new challenges when misused for malicious purposes. The automation of cyberattacks, the development of deepfakes, and the optimization of hacking techniques all demonstrate the dual nature of AI. As AI continues to evolve, it is essential to remain vigilant about its potential for misuse and to develop robust security measures to mitigate the risks associated with AI-driven cyber threats.

6. AI System Integrity as a Cybersecurity Imperative

As artificial intelligence (AI) becomes increasingly central to decision-making processes across critical sectors like healthcare, finance, and national security, ensuring the integrity of these systems is vital. AI system integrity refers to the ability of an AI model to function accurately, predictably, and securely, free from manipulation or corruption throughout its lifecycle. Unlike traditional software systems, AI models are more complex due to their reliance on dynamic data inputs and machine learning algorithms, which evolve over time. This makes the issue of AI integrity a key cybersecurity concern, requiring unique solutions to safeguard these advanced systems.

AI systems must maintain operational integrity to ensure that their predictions and decisions are trustworthy. In an era where AI models are used to diagnose diseases, make autonomous driving decisions, or process sensitive financial transactions; any deviation from expected behavior can have catastrophic consequences. One of the core challenges with AI system integrity is that these systems often operate as black boxes, making it difficult to fully understand how decisions are being made. If the internal workings of AI systems are opaque even to their developers, it becomes harder to detect when integrity is compromised. This opacity leads to an increased risk of unintentional errors, bias, and, more concerning, undetected manipulation.

AI systems must maintain operational integrity to ensure that their predictions and decisions are trustworthy. In an era where AI models are used to diagnose diseases, make autonomous driving decisions, or process sensitive financial transactions; any deviation from expected behavior can have catastrophic consequences. One of the core challenges with AI system integrity is that these systems often operate as black boxes, making it difficult to fully understand how decisions are being made. This lack of transparency complicates efforts to detect when integrity is compromised, increasing the risk of unintentional errors, bias, and undetected manipulation. Explainable AI (XAI) is one approach to addressing this challenge, as it offers transparency in AI decision-making and helps developers and users understand the reasoning behind AI outputs [5].

Maintaining AI system integrity is particularly crucial in systems where self-learning algorithms play a dominant role. Unlike traditional software, which functions based on predefined rules, AI systems

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

continuously learn from new data. This self-learning capability can improve efficiency but also introduces vulnerabilities to subtle forms of tampering or errors over time. For instance, an AI model that handles sensitive financial data may adjust its algorithmic predictions as new patterns in the data emerge. Without careful monitoring, these adjustments could lead to financial miscalculations or security breaches. To ensure integrity in these continuously evolving systems, it is essential to implement real-time monitoring, anomaly detection systems and robust validation mechanisms.

Another key dimension of AI system integrity is the ethics of AI decision-making. Ensuring AI integrity involves not only technical accuracy but also ethical integrity—ensuring that AI outputs are unbiased and fair. In areas like law enforcement or employment, where AI systems are used to predict criminal behavior or make hiring decisions, biased or unethical outputs can lead to systemic injustice. Detecting and mitigating bias is increasingly becoming part of the AI integrity discussion, as these biases can be a form of indirect attack on the integrity of AI systems. Bias auditing tools and ethical reviews are vital for maintaining the integrity of these systems.

Additionally, supply chain risks pose a significant threat to AI system integrity. AI systems rely on a wide range of components—hardware, software, data pipelines, and third-party services—that are sourced from multiple vendors. Each link in this supply chain introduces potential vulnerabilities. For example, if a compromised component is introduced during the AI development phase, it could affect the entire system, leading to long-term integrity issues. Furthermore, AI models often incorporate pre-trained models from third-party sources, which could introduce risks if those models were compromised during their own development. Securing the entire AI supply chain, from data sources to third-party infrastructure, is crucial to protecting AI systems from integrity breaches. AI system integrity also demands robust mechanisms for detecting abnormal behavior that could signal potential attacks. Anomaly detection plays a critical role in identifying when AI systems are not functioning as expected, which is vital in sectors such as finance and healthcare where even minor deviations could have severe consequences. Incorporating hybrid deep learning models such as CNN+GRU [6] [1], which has been shown to outperform other models in anomaly detection tasks [6] offers a promising solution for detecting subtle changes in AI system outputs, thereby maintaining their integrity.

AI system integrity also requires a focus on system resilience. In complex environments, maintaining integrity goes beyond just preventing malicious attacks—it includes ensuring the AI system can handle unexpected or unplanned disruptions. AI systems must be resilient in the face of changing conditions, whether due to shifts in data patterns or environmental factors, without compromising the accuracy or reliability of their outputs. This requires building AI systems with fail-safes and recovery mechanisms that can quickly correct errors or adapt to disruptions without degrading performance or security.

Furthermore, robustness testing ensures that AI systems remain secure against various forms of attacks, including adversarial manipulation and model tampering. Techniques like XGBoost[2] combined with anomaly detection models help to evaluate the resilience of AI systems in dynamic environments, ensuring they can withstand unexpected or malicious inputs [6]. This hybrid approach allows for continuous monitoring and real-time detection of anomalies, which is crucial in safeguarding the integrity of AI systems against evolving cyber threats.

[1] The CNN+GRU model is a hybrid architecture combining Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs). CNNs are deep learning models that extract spatial features or patterns by applying filters to input data, making them ideal for identifying local trends or anomalies. GRUs, on the other hand, are a type of Recurrent Neural Network (RNN) that captures temporal dependencies by maintaining context over sequences through gating mechanisms. Together, these models provide a robust solution for applications such as anomaly detection in time-series data, where recognizing both spatial and temporal patterns is crucial

[2] XGBoost is an advanced machine learning technique based on a decision-tree ensemble. It uses a boosting framework, where weak models are iteratively combined to produce a more accurate final model. Known for its high performance and efficiency, XGBoost includes regularization techniques to avoid overfitting, making it ideal for tasks like anomaly detection and predictive maintenance. By learning residual patterns from previous models, XGBoost refines predictions and handles complex datasets effectively.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

Finally, human oversight remains a cornerstone of ensuring AI system integrity. While AI systems are capable of operating autonomously, human involvement is essential to maintain trust and accountability. This includes regularly auditing AI systems, reviewing outputs for unexpected behavior, and validating that the AI continues to align with its original objectives. Human oversight helps to catch issues that automated systems may miss, ensuring a higher degree of integrity across the AI lifecycle. Implementing transparent monitoring and review mechanisms can ensure that any deviations in AI performance are quickly identified and corrected, further bolstering system integrity.

7. Integrating AI Safety into Cybersecurity Strategies: A Proactive Approach

To address the myriad risks associated with AI, it is imperative to integrate AI safety into existing cybersecurity frameworks.

Figure 2.
AI Safety as a Core Component of Cybersecurity Frameworks.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

This diagram illustrates how AI safety interconnects with core cybersecurity domains — including model robustness, data privacy, anomaly detection, explainability, ethical governance, and human oversight. These elements form the foundation for a proactive and resilient AI defense strategy.

This can be achieved through several key measures:

1. Robust AI Model Security: Developing practices for the secure implementation of AI models is essential. This includes using techniques such as encryption to protect model data, adhering to secure coding standards to reduce vulnerabilities, and conducting regular security audits to detect and prevent unauthorized access or manipulation. Additionally, implementing robustness testing ensures that AI systems can withstand adversarial inputs and attempts to compromise model integrity.

2. Data Integrity and Privacy Protections: Employing advanced data protection techniques, such as differential privacy, can secure training data from manipulation and unauthorized inference. Differential privacy allows AI models to learn from data while minimizing the risk of exposing sensitive information. This technique is particularly effective in balancing the need for privacy with the demand for accurate AI model training.

3. Continuous Monitoring and Threat Detection: Incorporating real-time monitoring systems within AI frameworks is crucial for identifying and responding to adversarial attacks, data anomalies, and other security breaches. Using hybrid models like CNN+GRU for anomaly detection helps detect subtle deviations in AI system behavior, allowing organizations to quickly mitigate potential threats. Explainable AI (XAI) further enhances this by providing insights into AI decision-making, enabling quicker identification of unexpected or malicious outputs.

4. Ethical AI Use Policies: Establishing and enforcing ethical guidelines for AI use is vital in mitigating misuse. This includes developing AI tools designed to detect and counteract AI-driven cyber threats, as well as ensuring that AI systems remain unbiased and accountable. Regular audits and bias detection mechanisms can help maintain the ethical integrity of AI applications.

5. Collaborative Defense Strategies: Encouraging collaboration among governments, corporations, and cybersecurity experts is necessary to develop comprehensive standards and best practices for AI safety. Such collaboration can lead to the creation of a unified framework for AI safety, similar to existing cybersecurity protocols, and ensure that AI development remains secure and beneficial for all stakeholders.

8. Conclusion

AI safety is not merely an abstract concept reserved for academic debate; it is a critical component of modern cybersecurity. As artificial intelligence (AI) systems become increasingly integrated into essential sectors such as healthcare, finance, transportation, and national security, the vulnerabilities and potential misuse associated with these systems represent real and significant threats. The reliance on AI for autonomous decision-making and data analysis introduces unique risks that, if left unchecked, could lead to severe disruptions in digital infrastructure and society at large.

AI systems are not only targets for cyberattacks but can also be misused as powerful tools to launch sophisticated cyber threats, amplifying their impact. This dual nature of AI makes it crucial to consider AI safety within the broader cybersecurity framework. The vulnerabilities and potential misuse associated with AI present tangible threats to data security, system integrity, and privacy. AI systems depend on large datasets—often containing sensitive personal, financial, or health-related information—to function effectively. A breach of these datasets can result in privacy violations, financial loss, and a significant erosion of public trust in AI technologies.

Furthermore, AI system integrity is increasingly at risk from adversarial attacks, data poisoning, and model manipulation, all of which can lead to incorrect or even dangerous outcomes. Compromised AI systems can have consequences far beyond individual breaches, affecting entire sectors and critical infrastructure. For example, a tampered AI system in healthcare could lead to misdiagnoses, while a compromised AI model in finance could cause erroneous financial transactions or market instability.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643

As AI continues to permeate vital sectors of society, the need to address its safety within the cybersecurity paradigm becomes increasingly urgent. With AI being deployed in everything from autonomous vehicles to smart cities and defense systems, ensuring the safety and security of these technologies is no longer optional—it is essential. Cybersecurity professionals must adapt traditional security practices to address the specific risks associated with AI, such as model integrity, secure data handling, and ethical decision-making. By recognizing AI safety as a core component of cybersecurity, it becomes possible to develop proactive measures that anticipate and mitigate these unique risks, ensuring that systems remain resilient in the face of evolving threats.

By framing AI safety as a fundamental aspect of cybersecurity, stakeholders can develop and implement robust strategies to protect both AI technologies and the data they handle. This includes technical solutions like encryption and real-time monitoring, as well as governance strategies that ensure ethical AI development and deployment. Cross-disciplinary collaboration between AI developers, cybersecurity experts, and policymakers is crucial for establishing standards and best practices that safeguard AI systems from threats while promoting innovation. Furthermore, integrating AI safety into existing cybersecurity protocols will help build trust in AI systems and ensure their responsible and secure deployment.

In this way, the potential benefits of AI can be harnessed while mitigating the risks it introduces to digital and societal infrastructure. AI holds immense potential to transform industries, improve efficiencies, and solve complex problems. However, without addressing the inherent risks, the widespread adoption of AI technologies could lead to unintended consequences, including security breaches, privacy violations, and ethical challenges. By embedding AI safety into the broader cybersecurity landscape, stakeholders can ensure that AI technologies are deployed in a manner that maximizes their benefits while minimizing their potential harms. This approach will help protect not only the systems themselves but also the broader societal infrastructure that increasingly relies on AI for critical operations.

Transparency:

The author confirms that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] M. Abdulhussein, "The impact of artificial intelligence and machine learning on organizations cybersecurity: Electronic resource," Doctoral Dissertations and Projects, 2024. https://digitalcommons.liberty.edu/doctoral/5242/

[2] D. Araya, "Cybersecurity and AI: Electronic resource / artificial intelligence for defence and security," Centre for International Governance Innovation, 2022. http://www.jstor.org/stable/resrep42557.11

[3] M. Brundage *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018. https://doi.org/10.48550/arXiv.1802.07228

[4] A. Munk, "Cybersecurity and AI-based technology: Electronic resource / liberty university," Doctoral Dissertations and Projects, 2022. https://digitalcommons.liberty.edu/doctoral/5242/

[5] I. H. Sarker, H. Janicke, A. Mohsin, A. Gill, and L. Maglaras, "Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects," *ICT Express*, vol. 10, no. 4, pp. 935-958, 2024. https://doi.org/10.1016/j.icte.2024.05.007

[6] B. R. Konatham, "A secure and efficient IIoT anomaly detection approach using a hybrid deep learning technique," *Industrial IoT Research Journal*, vol. 14, no. 3, pp. 28-30, 2023.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 1452-1463, 2025
DOI: 10.55214/2576-8484.v9i8.9643