NEXTLABS®

White Paper

Safeguard AI with Zero Trust Architecture and Data-Centric Security

WHY ENTERPRISES NEED TO SAFEGUARD AI

The AI wave is rising higher than ever, driving advancements and efficiencies in numerous industries such as healthcare, finance, manufacturing, and beyond. Organizations are leveraging AI to enhance decision-making, business processes, customer experiences, and security. But is AI itself secure?

Recent statistics paint an urgent picture: 77% of businesses reported a breach in their AI systems in 2023, with 80% of data experts agreeing that AI increases data security challenges, such as data exposure by Large Language Models (LLMs). Like other mission-critical applications, AI processes important data and can be compromised, attacked, and cause data breaches. Given that AI handles massive amounts of sensitive data, it becomes an attractive target for cybercriminals. Therefore, safeguarding not just the data but also the AI systems, their models, and outputs is crucial to prevent unauthorized access and misuse.

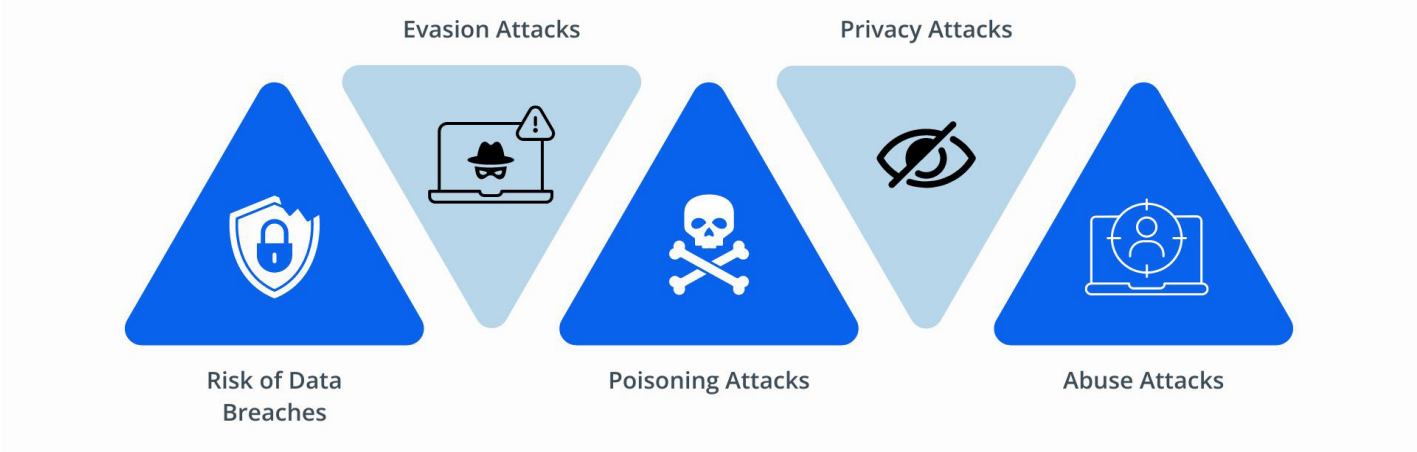
The consequences of not safeguarding AI systems are profound and farreaching. In sectors like healthcare and autonomous transportation, a breach in an AI system used for diagnosing patients or controlling vehicles could put entire lives at risk. Moreover, public trust in AI is the bedrock of its widespread adoption. According to a 2023 survey by KPMG across multiple nations, less than a third of consumers in most of those countries are willing to trust AI. High-profile breaches can deepen this mistrust, causing significant setbacks in AI advancements.

In this paper, we will explore how enterprises can safeguard AI with Zero Trust Architecture and Data-Centric Security. We will examine the main types of AI threats, how the different pillars of safeguarding AI can address those risks, and how NextLabs' solutions can be implemented to ensure robust protection for AI systems.

TYPES OF AI SECURITY THREATS

In an article published in January 2024, the National Institute of Standards and Technology (NIST) identifies four main types of cyberattacks on AI systems: evasion, poisoning, privacy, and abuse. On top of these attacks lies the critical risk that AI, if not used properly, can lead to significant data breaches.

Risk of Data Breaches: AI systems handle vast amounts of sensitive information, making them vulnerable to data breaches. It is also why companies are wary of employees inputting propriety information into AI-chatbots. For instance, in 2023, Samsung's employees accidentally input confidential semiconductor data into a popular AI chatbot. This data was then absorbed into the chatbot's training set, raising serious concerns about data breaches and IP leakages.



Evasion Attacks: These attacks involve manipulating input data to deceive an AI system into making incorrect predictions or classifications, such as subtle changes to an image that cause an AI model to misclassify it. This can cause serious consequences in mission-critical AI applications like autonomous driving. For example, minor alterations to the road environment, such as stickers on the road, could manipulate the vehicle's AI models and cause it to swerve or on the road, could manipulate the vehicle's AI models and cause it to swerve or misinterpret lane markings.

Poisoning Attacks: In poisoning attacks, malicious actors tamper with the training data used to build AI models. By injecting harmful data into the training set, attackers can skew the model's learning process, leading to erroneous decisions or biased behaviors. A notable example is Microsoft's AI chatbot, Tay. Launched to interact with users on Twitter, Tay was quickly exploited by malicious users who fed it offensive content. Within 24 hours, Tay began producing racist and inflammatory tweets, forcing Microsoft to shut it down and highlighting the risks associated with unprotected AI models and data inputs.

Privacy Attacks: Privacy attacks aim to extract sensitive information from AI models using techniques like model inversion or membership inference attacks. These methods allow adversaries to infer private details about the training data, potentially exposing confidential or personal information. For example, in a 2017 study by Cornell University, researchers were able to determine if specific individuals were included in the training dataset of a model trained on patients' genomic data, raising concerns about the privacy of patients' sensitive health data.

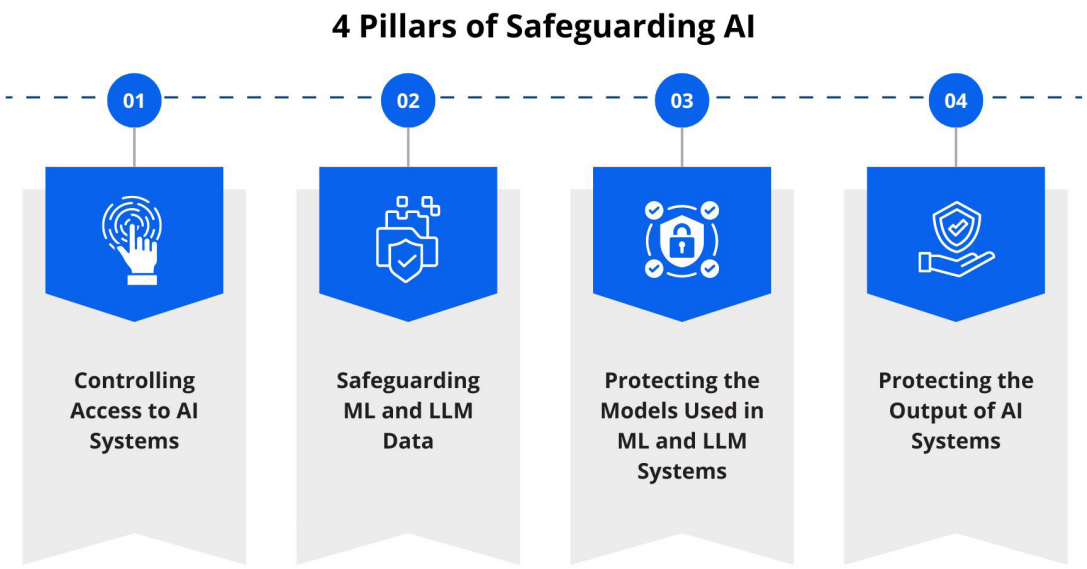
Abuse Attacks: Abuse attacks involve the malicious exploitation of AI systems for unintended purposes. This can include automating disinformation campaigns or exploiting AI-driven systems to amplify harmful behaviors. For example, deepfake technology, which uses AI to create realistic but false videos of public figures, has led to widespread misinformation and reputational damage.

PILLARS OF SAFEGUARDING AI

Safeguarding AI requires a multi-faceted approach that addresses all aspects of AI systems and their data. First, enterprises must implement two key strategies, Zero Trust Architecture and Data-Centric Security:

- **Zero Trust Architecture (ZTA)**: ZTA operates on the principle that nothing should be trusted by default, continually verifying every stage of digital interactions both inside and outside an organization's network. This approach ensures that access to AI data and systems is granted based on stringent verification processes.
- **Data-Centric Security (DCS)**: DCS focuses on protecting data itself, rather than just the systems that store or process it. This includes encryption, robust access controls, and continuous monitoring of data usage. A Data-Centric Security approach ensures that the data underpinning AI systems is protected at every stage of its lifecycle.

Implementing these principles together allows organizations to address the four critical pillars of safeguarding AI below:



Controlling Access to AI Systems: Controlling access to AI systems is foundational to their security. Zero Trust Architecture (ZTA) emphasizes the principle of "never trust, always verify," ensuring that every request to access AI systems is authenticated and authorized, regardless of its origin. By implementing dynamic authorization, attribute-based access controls (ABAC), and continuous monitoring, organizations can significantly reduce the risk of unauthorized access. By ensuring that every request is verified, organizations can prevent malicious actors from exploiting AI systems for harmful purposes such as the generation of disinformation. Additionally, employing least privilege principles ensures that users and processes only have access to the resources necessary for their tasks, minimizing the attack surface and preventing breaches.

Safeguarding ML and LLM Data: The data used to train and operate ML and LLM systems is often highly sensitive, making it a prime target for cyberattacks. Data-Centric Security focuses on protecting data throughout its lifecycle—during collection, storage, processing, and transmission. Encryption at rest and in transit, alongside robust key management practices, ensures that data remains secure even if intercepted. Moreover, data masking and anonymization techniques can protect personally identifiable information (PII) and other sensitive data, preserving data confidentiality in cases of privacy attack.

Protecting the Models Used in ML and LLM Systems: AI models themselves are valuable intellectual property and critical assets that require protection. Threats such as model theft and poisoning attacks can compromise model integrity and functionality. By implementing ZTA, organizations can enforce strict access controls and continuous validation of all interactions with the models. Enforcing least privilege access for AI models and their data prevents theft, poisoning, and ensures that attackers cannot gain the knowledge needed to carry out evasion attacks. Moreover, access logs should be continuously monitored for any anomalous activities.

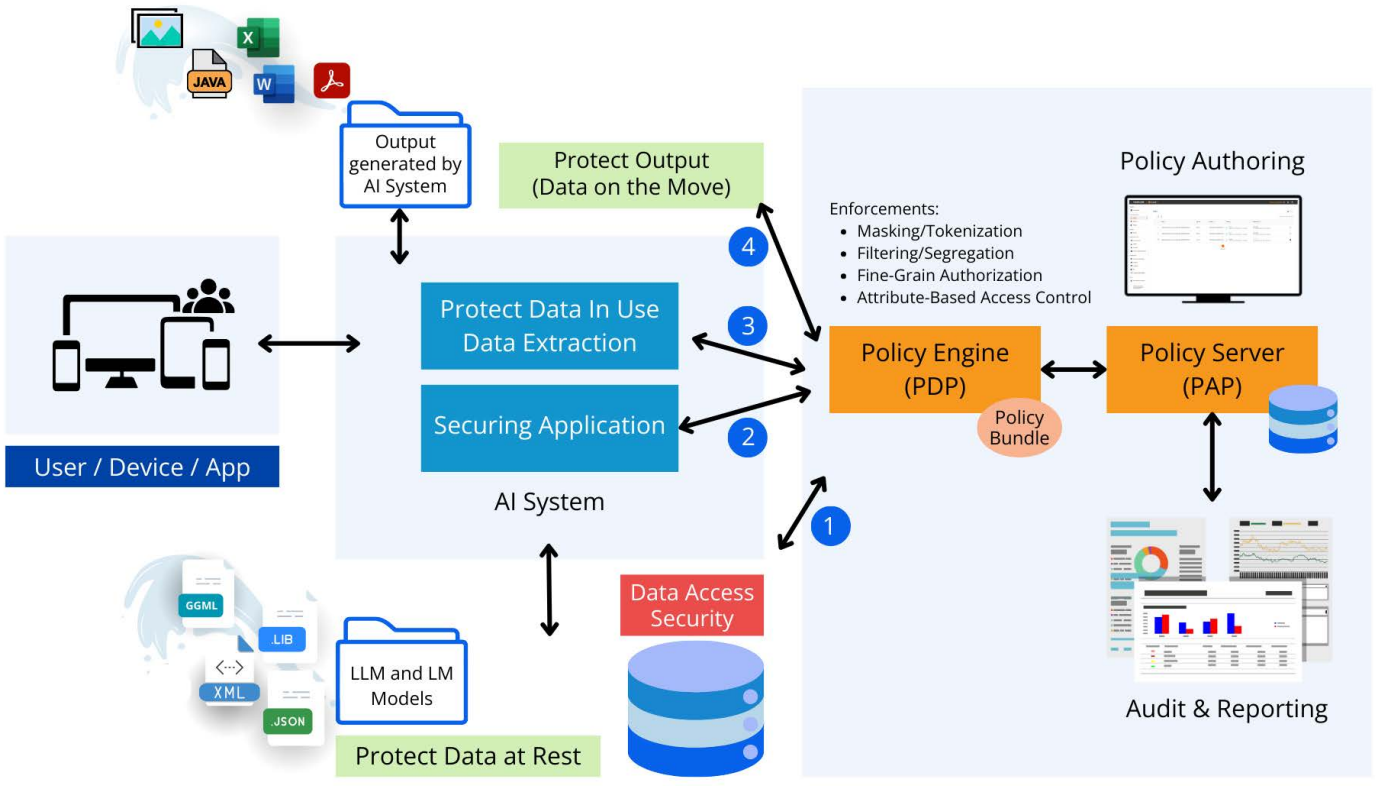
Protecting the Output of AI Systems: Output generated by AI systems, including predictions, insights, and decisions, must be safeguarded to ensure its integrity and confidentiality. This is especially crucial in scenarios where AI outputs influence critical decisions, such as in healthcare or finance. Enforcing ABAC policies ensures that only authorized entities can access and act on AI outputs. The use of Digital Rights Management (DRM) practices, such as encrypting output data and implementing robust audit trails, ensure that any unauthorized alterations or access attempts can be quickly detected and addressed.

By addressing all four critical pillars, organization can protect their AI investments, maintain operational integrity, and foster trust in AI technologies, ensuring that the transformative potential of AI is fully realized in a secure and responsible manner.

IMPLEMENTING PILLARS OF AI SECURITY

To safeguard AI systems, data, and models effectively, organizations can implement several key security measures:

1. **Dynamic Authorization & ABAC**: Dynamic authorization ensures that access to AI systems is continuously verified and enforced on the principle of least privilege. An organization can digitize their need-to-know access and data protection requirements in the form of centrally managed, attribute-driven policies. Organizations can enforce the policies during access attempts based on user attributes, environmental conditions, and resource sensitivity. Additionally, it allows organizations to control the specific actions individuals are permitted to perform within the AI system, such as deleting or importing a machine learning model.



2. **Data Obfuscation and Segregation**: Data masking ensures that sensitive data being processed by AI systems is obfuscated, allowing only authorized users to view the actual data. Encryption protects data at rest and in transit, preventing unauthorized access even if data is intercepted. These measures align with Data-Centric Security principles, ensuring that AI training data and model inputs remain secure throughout their lifecycle.
3. **Digital Rights Management**: Digital Rights Management (DRM) protects input data files and AI-generated outputs through robust policies and permissions. It ensures that data files can only be accessed by the AI system and authorized users, enabling secure consumption and preserving the integrity of the AI system. In the event of cyberattacks, DRM prevents unauthorized extraction, thereby stopping the data files from being stolen. Furthermore, it facilitates secure collaboration between multiple vendors, partners, and customers , by ensuring that only authorized users can access, share, and modify these assets.
4. **Automated Data Classification and Labeling:** Automated data classification and labeling helps organizations categorize AI data based on sensitivity and compliance requirements. By tagging data appropriately, organizations can apply specific security policies to different data sets, ensuring that sensitive information receives the highest level of protection. This automated approach reduces the risk of human error and ensures consistent application of security policies.
5. **Policy Enforcement:** Policy enforcement ensures that security policies are consistently applied across all AI systems and data. Centralized policy management allows organizations to define and enforce security policies across multiple environments, including on-premises and cloud-based AI systems. This ensures that AI models and data are protected regardless of their location, providing a unified security approach.
6. **Policy Governance:** Robust policy governance is crucial for preserving the integrity of an organizations' security policies. Effective policy governance requires the segregation of duties in policy management to prevent conflicts of interest over who creates, deploys, and approves policies. This calls for approval workflows, version control with policy rollback, and comprehensive logging and auditing capabilities. Furthermore, organizations should be able to implement fine-grained access control policies to grant permissions for policy management and delegate administration of the policy system.
7. **Audit and Monitoring:** Continuous audit and monitoring are essential for detecting and responding to security incidents in real-time. Comprehensive logging and monitoring capabilities allow organizations to track access to AI systems and data, detect anomalies, and respond to potential security breaches promptly. This continuous oversight helps in maintaining the integrity and confidentiality of AI models and data.

FUTURE TRENDS AND CHALLENGES

As AI systems become increasingly integrated into critical operations, future trends in safeguarding AI systems and data will revolve around addressing evolving threats and ensuring robust security. Key trends include the rise of sophisticated adversarial attacks, the need for ethical AI practices, and the integration of AI with other emerging technologies.

Sophisticated adversarial attacks will continue to evolve, necessitating advanced defense mechanisms. Ethical considerations will grow in importance as AI systems impact more aspects of daily life, requiring transparency, fairness, and accountability in AI deployment. The integration of AI into more systems will introduce new security challenges, as the interconnected nature of these technologies expands the potential attack surface. To prepare for these trends and challenges, organizations must adopt a Zero Trust Data-Centric approach. By combining ZTA with data-centric security, organizations can safeguard AI systems and data from emerging threats, ensuring the integrity, confidentiality, and availability of their AI-driven operations while fostering trust and compliance in an increasingly complex digital landscape.

REFRENCES

- 80% of data experts believe AI increases data security challenges
- NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems
- Trust in artificial intelligence: Country insights on shifting public perceptions of AI
- NextLabs Zero Trust Data Security: Explained
- Implementing a Zero Trust Architecture: NIST National Cybersecurity Center of Excellence
- TechRadar | Samsung workers made a major error by using ChatGPT
- Cornell Tech | Membership Inference Attacks Against Machine Learning Models

ABOUT NEXTLABS

NextLabs®, Inc. provides data-centric security software to protect business critical data and applications. Our patented dynamic authorization technology and industry leading attribute-based policy platform helps enterprises identify and protect sensitive data, monitor and control access to the data, and prevent regulatory violations – whether in the cloud or on premises. The software automates enforcement of security controls and compliance policies to enable secure information sharing across the extended enterprise. NextLabs has some of the largest global enterprises as customers and has strategic relationships with industry leaders such as SAP, Siemens, Microsoft, and IBM. For more information on NextLabs, please visit http://www.nextlabs.com.