nagarro

A guide to ensuring cyber security in AI applications

Table of Contents

Increasing need for cyber security in AI applications

Increasing need for cyber security in AI applications

The rise of Gen AI has pushed AI adoption levels across industries. As this emerging technology transforms industries, it also brings a new set of challenges and responsibilities. AI's power lies in its ability to process vast amounts of data and make intelligent decisions, but this very power also makes it a prime target for cyber threats.

In recent years, we've seen a surge in AI adoption, with businesses deploying thousands of AI models to streamline operations, enhance customer experiences, and drive innovation.

Researchers revealed the extensive use of AI in modern businesses, noting an average of 1,689 AI models actively used by companies. This has made AI security a top priority, with 94% of IT leaders dedicating funds to safeguard their AI in 2024.

According to HiddenLayer's 2024 AI Threat Landscape Report, 77% of business organizations have experienced breaches in their AI systems over the past year. The report estimates that 61% of IT leaders acknowledge shadow AI solutions that are not officially known or under the control of their IT department, posing serious risks.

However, this rapid expansion has also highlighted significant security vulnerabilities. From data breaches and model manipulation to sophisticated adversarial attacks, the risks associated with AI systems are growing more complex and more dangerous.

At the same time, increasing attacks have exposed cyber skill gaps in the organizational workforce. Traditional security drills no longer work. Organizations need a proactive and strategic approach to secure AI apps and systems. Recent attacks on AI applications further validate the need for stringent and robust security measures.

This ebook offers the essential knowledge and tools to navigate the complex landscape of AI security. We'll delve into the specific security challenges that AI applications face, offer practical insights for protecting your AI assets, and share best practices for building robust and secure AI systems.

It dives into the latest threats to AI systems, explores methods for designing and deploying secure AI solutions, and understands the importance of continuous monitoring and updates. It also highlights real-world examples of AI security breaches and ways to avoid such attacks.

Major attacks on AI apps in recent times

Backdoor attack on deep learning models in mobile apps: Microsoft researchers[i] have shown that many deep learning models used in mobile apps can be attacked through a method called "neural payload injection". This means that hackers can secretly insert harmful code into the AI system making them perform unintended actions.

After an empirical study on real-world mobile deep learning apps collected from Google Play, they identified 54 apps vulnerable to attack, including popular security and safety-critical applications used for cash recognition, parental control, face authentication, and financial services.

PoisonGPT: Researchers from Mithril Security[ii] showed that they could manipulate an open-source large language model (LLM) to give false information. They then uploaded this tampered model to HuggingFace, a major online model repository, demonstrating a significant vulnerability in the LLM supply chain. Users downloading the poisoned model could unknowingly spread incorrect data and misinformation, leading to various potential harms.

Confusing antimalware neural networks: Cloud storage is often used to deploy AI-based malware detectors. Users' systems collect data, which is then sent to cyber security company servers for analysis. Kaspersky's research team[iii] found that even with limited access to these AI models, hackers can manipulate data to bypass malware detection.

(i) *https://arxiv.org/abs/2101.06896*

(ii) *https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/*

(iii) *https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/*

Why cyber-vigilance is critical while building AI apps?

Why cyber-vigilance is critical while building AI apps?

Adopting cyber-vigilance while building AI apps is crucial for several reasons:

- **Data protection:** AI apps often deal with sensitive data. Implementing robust security measures protects this data from unauthorized access or breaches.
- **Privacy-related concerns:** There is a risk that data collected by AI apps could be shared with third parties without users' explicit consent. In such cases, implementing privacy-preserving techniques helps safeguard user data and maintain trust.
- **Preventing bias and discrimination:** AI models can inadvertently perpetuate biases in the training data. Detecting and mitigating such biases to ensure fair and unbiased results requires robust cyber security measures.
- **Guarding against malicious attacks:** AI systems are vulnerable to various attacks, such as adversarial attacks, data poisoning, and model inversion attacks.
- **Building trust:** Security breaches or privacy violations can severely damage users' trust in AI applications.
- **Compliance and regulations:** Many regions have stringent regulations regarding data protection and privacy (e.g., GDPR in Europe, CCPA in California). Adhering to these regulations requires a proactive approach to cyber security.

Understanding the nature of AI-specific attacks

Understanding the nature of AI-specific attacks

Several common attacks target AI applications, exploiting vulnerabilities inherent in machine learning models and data processes. Here are some of the most prevalent ones:

- **Adversarial attacks:** Adversarial attacks change inputs to AI models to make them give wrong predictions. These changes can be invisible to humans but are very effective at tricking the AI.
- **Data poisoning:** Data poisoning involves injecting malicious data into the training dataset to compromise the performance or behaviour of the AI model. Attackers aim to manipulate the model's decision boundaries or introduce biases by feeding it tainted data.
- **Model inversion:** Model inversion attacks aim to uncover sensitive information about a machine learning model's training data. By analyzing the model‘s outputs, attackers can infer details about individual training samples or identify patterns within the data.
- **Membership inference:** Membership inference attacks identify whether a specific data point was included in a machine learning model's training dataset. Attackers analyze the model's responses to determine if a particular input was used during training, potentially compromising user privacy.
- **Model stealing:** Model stealing attacks try to copy a machine learning model by repeatedly querying it and using the responses to create a similar model. This lets attackers avoid the time and resources needed to train their models.
- **Model evasion:** Model evasion attacks change input data to avoid detection or misleading classification systems. Attackers make small modifications to the input to trick the AI model into giving incorrect responses.
- **Backdoor attacks:** Backdoor attacks embed hidden triggers in the AI model during training. These triggers can be exploited later to make the model behave maliciously or incorrectly when specific inputs or conditions are met.

These attacks pose significant threats to AI applications' security, reliability, and trustworthiness across various domains. Addressing these vulnerabilities requires a combination of robust security measures, rigorous testing, ongoing monitoring, and collaboration between cyber security experts and AI practitioners.

Why do traditional security processes don’t cover AI specific attacks?

Why do traditional security processes don't cover AI specific attacks?

AI applications pose unique challenges that conventional security processes fail to address. Let's deep dive into why traditional security processes may not fully cover AI-specific attacks.

Adversarial nature: Traditional security processes aren't equipped to detect or mitigate adversarial attacks that exploit vulnerabilities in the machine learning algorithms as they require an understanding of the nuances of machine learning models and their susceptibility to manipulation.

Lack of a data-centric nature: Traditional security processes focus more on securing the infrastructure and network rather than monitoring and validating the integrity of data inputs and outputs in AI systems.

Complexity of AI models: AI models, especially deep learning models with millions of parameters, are highly complex. Understanding these models and detecting anomalous or malicious activities requires specialized knowledge and techniques that go beyond traditional security measures.

Evolving nature of AI-specific attacks: The landscape of AI-specific attacks constantly evolves as attackers develop new techniques to exploit vulnerabilities in AI systems. Traditional security struggles to tackle these rapidly changing threats without specialized tools, expertise, and methodologies tailored to AI security.

Interdisciplinary nature: Traditional security processes may not always enable collaboration between cyber security, machine learning, and data science experts required to address AI-specific attacks.

Regulatory and compliance gaps: Regulatory frameworks and compliance standards may not explicitly address AI-specific security concerns, leaving organizations uncertain about their obligations and best practices for securing AI systems.

How to architect, build and deploy AI solutions in a secure and efficient way?

How to architect, build and deploy AI solutions in a secure and efficient way?

Organizations need to take a holistic approach to AI security that incorporates specialized tools, techniques, and expertise tailored to the unique characteristics of AI systems, such as:

- Investing in AI-specific security solutions
- Fostering collaboration between cyber security and AI teams
- Implementing robust data governance practices

In the next sections of this ebook, we expand on the various best practices and strategies organizations can utilize to ensure utmost security through the three phases of architecting, developing, and deploying AI applications.

Architecting secure AI solutions

Security and privacy by design: Building a secure AI app begins with making security a conscious choice and then proactively strategizing to realize that goal.

We believe this begins with ensuring the security and privacy of the training data fed to the AI models. AI application designers and developers must collaborate to identify potential vulnerabilities and threats and aim to address them during the system's architecture design phase.

Organizations can ensure this by taking the following measures:

- Use a clean and well-marked dataset to train your model.
- Implement data encryption to protect sensitive data from unauthorized access.
- Classify data based on sensitivity and restrict access to sensitive data for authorized personnel only.
- Anonymize or pseudonymize sensitive data to mitigate privacy risks.

Note: Organizations can refer to AI security and privacy guidelines provided by OWASP. These guidelines provide clear and actionable insights on designing, creating, testing, and procuring secure and privacy-preserving AI systems.

Selecting the right AI models: While the choice of an AI model is largely driven by the problem an organization is trying to solve, the amount of data available and the desired accuracy level, there are a few factors to keep in mind from a cyber security perspective:

- **Understand the problem statement:** Clearly define your problem statement and underline what you aim to achieve through your AI models. This helps you narrow down the types of models suitable for your needs.
- **Evaluate model performance metrics:** Consider the performance metrics relevant to your problem domain, such as accuracy, precision, recall, F1-score, or area under the ROC curve (AUC). Choose a model that optimizes these metrics based on your requirements.
- **Consider model complexity and interpretability:** Balance the complexity of the model with its interpretability and explainability. Simple models like linear regression or decision trees are easier to interpret but may not capture complex patterns as effectively as deep learning models.
- **Assess data requirements:** Evaluate the availability, quality, and quantity of data needed to train and evaluate different models. Some models may require large, labeled datasets, while others can perform well with smaller data.
- **Explore pre-trained models and transfer learning:** Consider leveraging pre-trained models and transfer learning techniques, especially if you have limited data or computational resources. Pre-trained models are trained on large datasets and can be fine-tuned for specific tasks with smaller datasets.

Evaluating the scalability and flexibility of hardware models: Organizations must consider the scalability and flexibility of their hardware infrastructure to accommodate future growth and changing requirements.

Scalable AI pertains to how data models, infrastructures, and algorithms can increase or decrease their complexity, speed, or size at scale to best handle the requirements of the situation at hand.

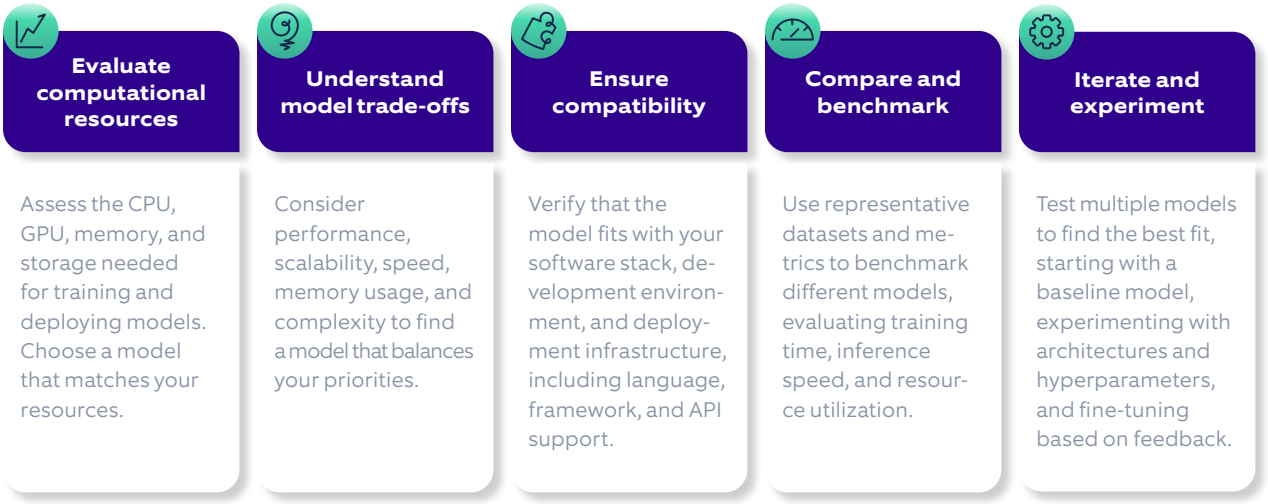
To begin with, they must evaluate the cloud-based, on-premises and hybrid solutions to see which option aligns the best given their requirements and constraints. The next step involves considering how AI solutions will scale with the growing data volumes of usage.

- **Compliance:** AI systems should be developed and used considering applicable laws, compliance, and regulations. Organizations must also ensure that their AI systems comply with applicable regulations, such as HIPPA, GDPR, and CCPA, among others.
- **Policies review:** Organizations should review the HR and InfoSec policies to exhibit specific use case enablement and management.
- **Auditing and reporting:** Organizations should maintain audit trails and reporting processes to demonstrate compliance with regulations and internal policies.

Ethical implications of AI: Consider the ethical implications of your AI solution, such as bias and fairness. Developers should implement bias detection mechanisms, conduct ethical impact assessments, and embrace fairness-aware AI design practices to minimize discriminatory outcomes.

AI solution integration: Consider how your AI solution will integrate with existing systems and processes. Define the AI use case and analyze your existing systems and processes, identify and evaluate the data sources that will be utilized to inform and train the AI models, and select the appropriate AI platform, algorithms, and frameworks within your infrastructure.

A few other factors to consider for integrating security considerations while designing AI applications:



Choosing the right AI hardware

Selecting the right hardware is essential for securing AI applications. Optimal hardware performance improves algorithm speed and reduces security risks. Proper selection ensures efficient resource allocation, preventing resource exhaustion and denial-of-service attacks. Modern hardware includes built-in security features like encryption engines and secure boot mechanisms, enhancing system security. Additionally, selecting hardware with mitigations against vulnerabilities such as Spectre and Meltdown protects sensitive AI data.

Factors to consider while selecting the hardware for your AI applications:

- **Understand the requirements:** Thoroughly understand the re quirements of your AI application, including computational power, memory, storage, and throughput. Consider factors like the dataset size, model complexity and real-time processing needs.
- **Evaluate processing units:** Assess different types of processing units, including CPUs, GPUs, TPUs, FPGAs, and ASICs, and how they suit specific tasks in your AI workload.
- **Evaluate accelerators and co-processors:** Explore the use of hardware accelerators and co-processors and whether the selected hardware supports popular AI frameworks and libraries.
- **Assess power efficiency and cost optimization:** Look for hardware solutions that balance your specific workload's performance, energy efficiency, and cost-effectiveness.
- **Optimize for memory and storage:** Pay attention to memory and storage requirements, especially for large-scale AI applications with high-dimensional data or massive datasets. Consider hardware configurations with ample memory bandwidth, high-speed storage, and efficient data access patterns to minimize latency and maximize throughput.
- **Benchmarking and performance testing:** Conduct benchmarking and performance testing to evaluate the suitability of different hardware options for your AI workload. Measure processing speed, accuracy, throughput, and scalability to make informed decisions under realistic conditions.
- **Consider specialized hardware and architectures:** Explore specialized hardware and architectures designed specifically for AI workloads, such as neuromorphic chips, quantum processors, or edge AI devices, for specific use cases or deployment scenarios.

Enabling secure development

Ensuring security during the development phase includes writing secure code and using secure coding practices and standards to minimize errors and vulnerabilities.

Business enterprises must use secure repositories and implement strong authentication and access controls. The development environment should be regularly scanned and monitored for vulnerabilities and adhere to the following coding best practices:

- Secure data sources and pipelines
- Follow coding standards and guidelines
- Implement security testing and validation
- Adopt secure deployment and maintenance practices
- Consider ethical and legal implications

Note: Organizations can refer to Google's Secure AI Framework (SAIF) as a guiding document while developing their secure coding practices. OWASP AI Exchange is another collaborative project that advances the development of AI security standards and regulations. It provides a comprehensive overview of AI threats, vulnerabilities, and controls.

Securing the API: AI systems interact with other applications or services through APIs, so organizations should ensure secure API design and consider using proper authentication mechanisms.

Following are some best practices for securing AI APIs:

- Ensure all communication between clients and the API server is encrypted. Implement robust authentication mechanisms, such as API keys, OAuth tokens, or JWT tokens, to verify the identity of clients accessing the API.
- Enforce access controls and authorization policies to restrict access to specific endpoints or resources based on user roles and privileges.
- Implement rate limiting and throttling mechanisms to prevent abuse, DoS attacks, and excessive API usage. Limit the number of requests per time interval and apply adaptive throttling based on usage patterns and client behaviour.
- Validate and sanitize all input data, use parameterized queries, input validation libraries, and content security policies to sanitize user inputs and mitigate common security vulnerabilities.
- Encode and escape output data to prevent cross-site scripting (XSS) attacks and other injection vulnerabilities. Use appropriate encoding techniques, such as HTML entity encoding or URL encoding, to sanitize output data before rendering it in web pages or responses.
- Secure the underlying infrastructure and runtime environment where the AI models are deployed. Use containerization technologies, such as Docker or Kubernetes, with security best practices for isolation, sandboxing, and access control.
- Regularly update the API's dependencies, libraries, and third-party components to patch known security vulnerabilities and mitigate risks associated with outdated software.
- Log user actions, API requests, errors, and security events. Use log analysis tools to gain security insights. Monitor security advisories and subscribe to vulnerability databases to stay updated on potential threats.

How to ensure secure deployment/ implementation of AI applications?

AI solution Maintenance: AI systems require regular maintenance and updates, so it is important to consider system updates and smooth functioning.

Organizations can keep AI software and frameworks updated by following the latest security patches, regularly updating their libraries, addressing dependencies and managing operating systems.

Using our cyber security expertise, we have listed some of the best practices organizations can follow to maintain AI solutions.

Regular performance monitoring

- Monitor the performance of AI models and algorithms to detect degradation or drift in accuracy and efficiency.
- Use metrics such as accuracy, precision, recall, F1-score, and inference latency to evaluate model performance and identify improvement areas.

Data quality assurance

- Regularly assess and maintain input data quality for training and inference.
- Implement data validation and cleansing to correct errors and inconsistencies.
- Monitor data distribution shifts, and concept drift to keep AI models effective.

Model retraining and updates

- Schedule regular retraining of AI models with fresh data to adapt to new patterns.
- Use incremental learning to update models in real-time without full retraining.
- Monitor model performance before and after updates to ensure improvements and avoid side effects.

Security patching and vulnerability management

- Regularly update AI software dependencies to patch security vulnerabilities.
- Follow security best practices, including secure coding, access controls, and encryption.
- Perform regular security audits and penetration tests to fix weaknesses.

Version control and configuration management

- Use version control to track changes in code, models, and configurations.
- Document and store metadata for each AI solution version for reproducibility and collaboration.
- Apply configuration management to track system configurations, environments, and dependencies.

Automated testing and continuous integration

- Set up automated testing pipelines and CI/CD processes for streamlined validation and deployment.
- Incorporate unit, integration, and end-to-end tests to ensure AI system functionality and reliability.
- Utilize AI-specific testing tools like TFX or PyTorch Lightning for consistent quality.

Documentation and knowledge sharing

- Maintain comprehensive documentation, including code comments, README files, user guides, and API docs.
- Encourage team knowledge sharing by documenting best practices, lessons, and troubleshooting tips.
- Create internal knowledge repositories or wikis for centralized AI solution information and support.

AI solution monitoring: Organizations should implement necessary measures to monitor the performance of AI solutions to troubleshoot and fix issues. Following are some best practices for AI solution monitoring:

- Define key performance indicators and enable real-time monitoring.
- Generate comprehensive logs and audit trails of system activities.
- Implement matrices related to model performance monitoring.
- Monitor the quality and integrity of input data sources used for training and inference to detect anomalies, errors, or biases.
- Monitor resource utilization and scalability such as CPU, GPU, memory, storage, network bandwidth etc.
- Monitor system logs, access controls, authentication events, and network traffic for signs of security incidents, unauthorized access, or suspicious activities.
- Monitor compliance with relevant regulations, standards, and policies, such as GDPR, HIPAA, SOC 2, ISO 27001, and industry-specific requirements.

Data governance: With an ever-increasing need for responsible AI practices, it is crucial that data is collected, stored, used and disposed of ethically, legally, and effectively. Data governance involves the management of an organization‘s data and ensures its availability, usability, integrity, and security.

Here‘s how organizations can build an actionable data governance framework:

- **Data access:** Use necessary measures to determine who accesses the data and under what circumstances. Establish procedures for granting and revoking access and ensuring access is granted only to authorized personnel.
- **Data classification:** Build necessary measures to classify data based on its sensitivity, criticality, and value level. This would further help manage and protect organizational data more effectively.
- **Data retention and disposal:** To prevent unauthorized access, use necessary procedures to specify data storage timelines and disposal guidelines.
- **Data privacy:** Organizations should take necessary measures to outline how data is collected, used, and shared to protect individuals‘ privacy and confidentiality.

This includes policies for obtaining consent, anonymizing data, and ensuring compliance with relevant regulations such as GDPR and CCPA.

Perform regular security and risk assessments: Regular security and risk assessments help organizations identify vulnerabilities in their systems and networks. This includes penetration testing and vulnerability scanning to identify weaknesses that hackers could exploit.

The Trustworthy and Responsible AI Resource Center and the AI Risk Management Framework (AI RMF 1.0) mention various guidelines for security and risk assessments.

Create a strong incident response plan: Incidents happen despite the best efforts to ensure cyber security. Hence, it is important to create a strong incident response plan. Organizations must develop a strong incident response plan outlining a plan of action in the event of a cyber-attack. This includes identifying key personnel, establishing communication channels, and having a plan in place to restore systems and data.

MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems) is a globally accessible, living knowledge base of adversary tactics and techniques based on real-world attack observations and realistic demonstrations from AI red teams and security groups. Organizations can use this knowledge base to build incident response and identify detection techniques.

Vendor security assessment: Organizations must conduct security assessments of vendors to check whether they use any third-party AI tools or services and ensure that vendors follow best security practices and adhere to necessary compliance standards.

Training and awareness to handle AI-based tools:

- Ensure your employees are securely prepared to use AI tools before integrating them into your business processes.
- Avoid free versions of paid AI software to prevent malware and viruses.
- Be cautious with downloads and links from untrustworthy sources.
- Provide educational resources, apply a train-the-trainer approach, involve stakeholders in reviews, and conduct AI workshops for trainers.

Collaborate with cyber security experts and threat intelligence feeds: Organizations must collaborate with cyber security experts to stay current on the latest threats and defences. This includes using threat intelligence feeds, attending security webinars, and working with trusted partners to develop customized cyber security solutions.

Leveraging AI while ensuring cyber security

Leveraging AI while ensuring cyber security

AI presents both unparalleled opportunities and significant security challenges. As organizations increasingly integrate AI into their operations, ensuring the security of these systems becomes paramount.

For organizations working with AI, exploring the multifaceted aspects of AI security is crucial, emphasizing the need for a proactive and comprehensive approach. A holistic strategy that includes secure coding practices, version control, automated testing, and thorough documentation would ensure safe AI solutions.

They must integrate measures like continuous data quality assessment, regular model retraining, and robust security measures tailored to the unique vulnerabilities of their AI systems.

Additionally, fostering collaboration between cyber security and AI teams and leveraging specialized tools and frameworks are crucial steps in safeguarding AI applications.

By staying vigilant and adopting these best practices, organizations can protect their AI investments, maintain user trust, and navigate the complexities of AI security. As AI technology advances, so must our efforts to secure it, ensuring a future where AI can be harnessed safely and effectively.

Securing AI applications is not just about defense; it's about building a robust, resilient system that can adapt, grow, and thrive in adversity. Following the best practices outlined in this guide, you can safeguard your AI investments, protect user trust, and pave the way for a secure and innovative future. Let's embrace this journey with determination and foresight, ensuring our AI-driven future is bright and secure.

References:

https://arxiv.org/abs/2101.06896

https://www.iso.org/standard/81118.html

https://owasp.org/www-project-ai-security-and-privacy-guide/

https://www.ncsc.gov.uk/blog-post/introducing-our-new-machine-learning-security-principles

https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf

https://aiverifyfoundation.sg/what-is-ai-verify/

https://www.safe.ai/ai-risk

https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.pdf?__blob=publicationFile&v=5

https://www.codemotion.com/magazine/ai-ml/building-ai-enabled-applications-best-practices-for-developers/

https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile&v=4

https://www.nccoe.nist.gov/ai/adversarial-machine-learning

https://atlas.mitre.org/

About the author

Rakesh Gogane

Security Architect, Nagarro

Rakesh is a highly accomplished Security Architect, an AI security researcher and an author.
He has extensive experience designing and developing security solutions, compliance, and security best practices and controls. He is a highly motivated security specialist, well-versed in continuous learning, and passionate about innovation in information security, contributing to enhancing business decisions, reducing corporate liabilities, and building secure digital infrastructure.

About Nagarro

Nagarro helps future-proof your business through a forward-thinking, fluidic, and CARING mindset. We excel at digital engineering and help our clients become human-centric, digital-first organizations, augmenting their ability to be responsive, efficient, intimate, creative, and sustainable. Today, we are 18,000 experts across 36 countries, forming a Nation of Nagarrians, ready to help our customers succeed.

For more information, visit www.nagarro.com.


nagarro