AI/ML/LLM security study map.

experimental, this will be fixed when we gather enough experience ;-)

Hi. This is my attempt to gather the knowledge that is needed to understand the security of AI. From the basics, to practical labs. This is the first version of the map. I'm sick of the fact that there are still no proper maps in this thread that can help us - safety experts to understand all this. Regards, **@wearetyomsmnv** !

- Level 0 - Fundamental knowleges
- Level 1 - Basic knowledge
- Level 2 - Advanced knowledge
- Level 3 - Practice

How does AI security differ from traditional cyber security?

Data Integrity: AI systems rely heavily on data for learning. [Ensuring the integrity of this data is crucial, as attackers can manipulate the data to influence AI behavior, a tactic known as data poisoning

Model Security: The AI's decision-making model itself can be a target. [Attackers may attempt to reverse-engineer the model or exploit its weaknesses to make incorrect or harmful decisions

Adversarial Attacks: AI systems can be susceptible to adversarial attacks, where slight, often imperceptible alterations to input data can cause the AI to make errors or incorrect predictions

Infrastructure Security: While traditional cybersecurity also focuses on protecting infrastructure, AI systems may have additional layers of complexity, such as cloud-based services or specialized hardware, that require specific security measures

Ethical Considerations: The use of AI in security brings ethical considerations, such as privacy concerns and the potential for bias in decision-making, which must be addressed in the security strategy

Key Attacks in AI.

Data/model poisoning, Supply chain, Bad deserialization, Input Manipulation Attack, Membership Inference Attack, Transfer Learning Attack, Output Integrity Attack, Model Stealing, Prompt Injection, Insecure Output Handling, Overreliance, Backdoors, Model Extraction Attacks, Adversarial Attacks (on image and on NLP), RAG attacks, DNN attacks, Visual Systems Attack,Transferability attacks, Over-Unlearning mlaas.

Key Concepts in AI.

Deep Learning, Large Language Models, Natural Language Processing, Unsupervised Learning, Reinforcement Learning, Computer Vision, Knowledge Representation and Reasoning, Cognitive Computing, RAG, BigData, Anomaly Detection, Behavior Analytics.

You need to understand what each term means, as well as a basic understanding of how it applies to AI.

Key Algoritms in AI.

Linear Regression, Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Recurrent Neural Networks (RNNs), Transformers, Long Short-Term Memory (LSTM), Generative Adversarial Networks (GANs), Gradient Boosting, Deep Q Network (DQN), Hierarchical Clustering, Policy Gradients, DNN, CNN.

Protection

LLM Firewalls, Adversarial Testing, Data Encryption and Masking/anonymization, DataLake security, MlSecOps, Defending topic on OffSec ML Wiki.

- Offsec WIki has topic about defense, look in the menu on the left.

frameworks

- OWASP ML TOP 10
- OWASP Top 10 for Large Language Model Applications
- Databricks framework ai security(DASF)
- Mitre Atlas
- Nist Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations
- AI Risk Assessment for ML Engineers
- Gartner AI Trust, Risk, and Security Management (AI TRiSM)
- IBM Framework for Securing Generative AI
- OWASP LLMSVS

Books

Generative AI Security (Ken Huang, Yang Wang, Ben Goertzel, Yale Li, Sean Wright, Jyoti Ponnapalli)

Machine Learning for penetration testing

Labs

- Crucible
- DamnVulnerableLLM
- Gandalf
- PortSwigger WebLLM attacks
- Adversarial ML CTF

laws, standards

- Biden's AI executive order (USA)
- FTC: Keep your AI claims in check (USA)
- FAA - Unmanned Aircraft Vehicles (USA)
- NHTSA - Automated Vehicle safety (USA)
- AI Bill of Rights (USA)
- Relaxing copyright for AI training (JAPAN)
- AI white paper (UK)
- Rules for GenAI services (CHINA)
- Voluntary AI Verify system (SINGAPORE)
- AI ethics framework (AU)
- AI Act (EU)
- ISO/IEC 42001 Artificial intelligence — Management system
- ISO/IEC 22989 — Artificial intelligence — Concepts and terminology
- ISO/IEC 38507 — Governance of IT — Governance implications of the use of artificial intelligence by organizations
- ISO/IEC 23894 — Artificial Intelligence — Guidance on Risk Management
- ANSI/UL 4600 Standard for Safety for the Evaluation of Autonomous Products

Newsfeed, research companies

- https://t.me/aisecnews
- https://embracethered.com/
- https://protectai.com/
- https://hiddenlayer.com/
- https://dreadnode.io/
- arXiv

references

https://wiki.offsecml.com/Welcome+to+the+Offensive+ML+Playbook

https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Security-of-AI-systems_fundamentals.pdf

https://www.nightfall.ai/ai-security-101

https://github.com/microsoft/Security-101/blob/main/8.1%20AI%20security%20key%20concepts.md

https://github.com/RiccardoBiosas/awesome-MLSecOps