AI/ML/LLM security study map. v1

This project is aimed at beginners in the field of AI security. It covers most of the topics you need to learn to become an expert in AI security.
Regards, @wearetyomsmnv.

Contributors: @marshack44, Roman Lebed, @ivolake, Valery Kunavin, @Od1n28, Evgeniy Kokuykin, @conversational_cat.

Pathway for learning about laws and standards

A pathway for learning classical AI

A pathway for exploring vulnerabilities and attacks on AI

A pathway for exploring methods to protect artificial intelligence

A pathway for secure development and MlSecOps

Traditional cybersecurity fundamentals

Algorithms, Statistics(collection, analysis, interpretation, presentation, and organization of data), Probability, Linear Algebra, Calculus, Differential Equations, Optimization, Vector Space, Matrix Decomposition, Gradient Descent, Convex Optimization, Eigenvalues and Eigenvectors, Fourier Transform, Discrete Mathematics, Combinatorics, Graph Theory, Bayesian Networks, Markov Decision Processes, Reinforcement Learning, Information Theory, Principal Component Analysis, Regularization.

Math Concepts in AI.

AI security and traditional cybersecurity share many similarities, but they also have some distinct differences due to the unique characteristics and capabilities of artificial intelligence systems. Here's how they differ:

Complexity of Threats: AI systems introduce new layers of complexity to cybersecurity. Traditional cybersecurity primarily deals with threats like malware, phishing attacks, and network intrusions. However, AI systems can be vulnerable to attacks such as adversarial attacks, data poisoning, and model evasion, which specifically target the machine learning algorithms themselves.

Attack Surface: AI systems often have larger attack surfaces compared to traditional systems. This is because they not only rely on software but also on data and models. Attackers can target the training data, manipulate models, or exploit vulnerabilities in the algorithms themselves.

Adaptability of Threats: AI systems can adapt and learn from their environment, which can make them more susceptible to adaptive and evolving threats. Traditional cybersecurity measures may not be sufficient to defend against attacks that constantly evolve based on the behavior of the AI system.

Interpretability and Explainability: Understanding why an AI system made a particular decision is often more challenging compared to traditional software systems. This lack of interpretability and explainability can make it difficult to detect and mitigate attacks on AI systems effectively.

Data Privacy Concerns: AI systems often rely on large amounts of data, which can introduce privacy risks if not properly handled. Traditional cybersecurity measures may not adequately address these data privacy concerns specific to AI systems.

Regulatory Compliance: The regulatory landscape for AI security is still evolving, with specific regulations and standards emerging to address the unique challenges posed by AI systems. Traditional cybersecurity frameworks may need to be extended or adapted to ensure compliance with these new regulations.

Ethical Considerations: AI security involves not only protecting systems from malicious attacks but also ensuring that AI systems are used in an ethical and responsible manner. This includes considerations such as fairness, transparency, and accountability, which may not be as prominent in traditional cybersecurity.

Graphics Processing Unit (GPU),Tensor Processing Unit(TPU), Application-Specific Integrated Circuit(ASIC), General-Purpose Graphics Processing Unit(GPGPU), Field-Programmable Gate Array(FPGA), Neural Network Processor(NNP), System on Chip AI Accelerators (SoC AI Accelerators), Edge AI Hardware, Intelligence Processing Units(IPU), Single Instruction, Multiple Data(SIMD), Very Long Instruction Word(VLIW)

AI Hardware basics.

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

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
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ. ТЕХНИЧЕСКАЯ СТРУКТУРА ФЕДЕРАТИВНОЙ СИСТЕМЫ МАШИННОГО ОБУЧЕНИЯ (Russia)
ПНСТ 848-2023, ПНСТ 847-2023 (Russia)

Оценка качества систем искусственного интеллекта. Общие положения. ГОСТ Р 59898-2021 (Russia)
Информационные технологии. Интеллект искусственный. Оценка робастности нейронных сетей. Часть 1. Обзор. ГОСТ Р 70462.1-2022/ISO/IEC TR 24029-1-2021 (Russia)
Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. ГОСТ Р 59276-2020

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, Logistic Regression, Decision Trees, Random Forests, Transformers, Long Short-Term Memory (LSTM), Generative Adversarial Networks (GANs), Gradient Boosting, Deep Q Network (DQN), Hierarchical Clustering, Policy Gradients, DNN, CNN, Leave-One-Out Cross-Validation (LOOCV), K-Fold Cross-Validation, Hyperparameter Tuning Protocols, Model Evaluation Metrics, Embeddings, Attention Layers, Transformers.

Key Concepts in AI.

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

Out-of-Distribution (OOD) Detection, Guided Denoiser, Model Ensembling, Defensive Distillation, Gradient Masking / Obfuscation, Robust Optimization, Adversarial Training.

Giskard, CleverHans, Rebuff, Safetensors, Citadel Lens, Robust Intelligence Continous Validation, AI Fairness 360.

Bias and Fairness Audits, Robustness Testing, Homomorphic Encryption, Privacy Preserving Machine Learning (PPML), Federated Learning with Secure Aggregation, Zero-Knowledge Proofs (ZKP), Interpretability Methods, Rate Limiting, Watermarking, Safety Instruction, Data anonymization.

NB Defense, Guardian, ARX - Data Anonymization Tool, Syft, differential-privacy-library, Data-Veil, Neural Cleanse, Guardrails AI, bhakti, AIShield GuArdIan, DeepKeep, PurpleLLaMA,

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.

art framework, foolbox, flickling, copycat cnn, Advertorch, AdvBox, MlSploit.

Overreliance, Backdoors, Model Extraction Attacks, Adversarial Attacks (on image and on NLP), RAG attacks, DNN attacks, Visual Systems Attack,Transferability attacks, Over-Unlearning mlaas.

textattack, counterfit, Model-Inversion-Attack-ToolBox, Knockoffnets.

AIM Jailbreak attack, Affirmative Suffix attack, Amnesia attack, Hallucinations attack, Contextual Redirection attack, DAN jailbreak, Harmful Behavior, Linguistic Evasion, Self Refine, UCAR, Base64 Evasion, ASCII smuggling, ASCII art attack.

garak, ps-fuzz, vigil, NeMo-Guardials, ASCII Smuggler Tool, Pyrit, mindgard-cli.

Code patterns: (keras layer backdoor, ONNX backdoor, keras protobuf vulnerability, pickles), MlOps infrastructure vulnerabilities, Modelhub's attack,

AI-exploits,keras malicious model, Neuron-Based-Steg, ONNX runtime exploit, Hijacking Safetensors.

Key Attacks in AI.

CI/CD, IaC, Policy as Code, Shift Left, Threat modeling, Secret Management, Compliance as Code, Security Champions, Container Security, Machine Learning Pipelines, Data and DataHub Security, Model Privacy, Model Monitoring, SAST, DAST, SCA, Robustness Testing, Secure Model Serving, Federated Learning, Differential Privacy, ModelHub security, Model output watermarking

Modelscan, Safetensors, lintML, differential-privacy-library, Guardian, Robust Intelligence Continuous Validation, TensorFlow Model Analysis, CleverHans, SecML, AI Exploits, AIShield Watchtower, Databricks Platform, Azure Databricks, Hidden Layer AI Detection Response, Hidden Layer AISEC Platform, Guardrails AI, Syft, Private AI, Alibi Detect, Watermark papers, Hashicorp Vault

Secure Machine learning development / MlSecOps

conferences,talks

BlackHat
DefCon AIvilliage
OffZone conference
Positive Hack Days
SatML Conference
RSA Conference
IEEE Conference on Secure and Trustworthy Machine Learning
The Conference on Applied Machine Learning in Information Security

labs

Crucible
DamnVulnerableLLM
Gandalf
PortSwigger WebLLM attacks
Adversarial ML CTF

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
OWASP AI EXCHANGE
AN ARCHITECTURAL RISK ANALYSIS OF LARGE LANGUAGE MODELS: Applied Machine Learning Security

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
https://docs.microsoft.com/en-us/security/failure-modes-in-machine-learning
https://plot4.ai/library
https://learnprompting.org/docs/prompt_hacking/intro

Newsfeed, research companies

https://t.me/aisecnews
https://hackstery.com/
https://blog.trailofbits.com
https://embracethered.com/
https://adversa.ai/
https://laiyer.substack.com/
https://protectai.com/
https://www.lakera.ai/
https://blog.wearetyomsmnv.wtf/
https://hiddenlayer.com/
https://www.deepkeep.ai/
https://wiki.hego.tech/
https://dreadnode.io/
https://aipwn.org/

Level 0 - AI Initiator:

Using AI: Recognise the limits of the technology and can use artificial intelligence in a basic way.

Level 1 - "Application Specialist":

Effective Application: Can use existing AI solutions effectively and recognise the potential risks.

Level 2 - "Developing Expert":

Refine and extend: Have the ability to improve and extend existing AI systems and understand the principles of their operation at the architecture and algorithm level.

Level 3 - "Innovation Expert":

System Vision: I see the system as a whole, can propose new approaches and develop innovative AI solutions.