AI Security Cheatsheet

AI and Security

The concept of AI security has different aspects:

- Securing AI systems from cyberattacks
- Using AI technologies for security purposes
- New security risks caused by using AI

This document focuses on securing AI systems.

What is AI Security

AI Security focuses on protecting AI systems from malicious attacks and unauthorised access, ensuring the system remains intact and operational.

AI Security applies to:

IT infrastructure
software, hardware

AI model
model, datasets

Product
user data, prompts and responses

AI Security is NOT equal to:

Open source

fosters the transparency, accessibility, and collaborative development of AI systems

Benchmarking

evaluates the security of AI systems by comparing their performance against established standards or benchmarks

Compliance

ensures that an organisation adheres to legal and regulatory requirements

Open source, benchmarking, and compliance, while necessary, do not guarantee that all security risks are addressed.

Security vs Trustworthiness vs Safety

There is a difference between security, safety, and trustworthiness of AI.

Security

is a process of defending AI systems from cyberattacks

Trustworthiness

is the property of AI systems that consistently perform as expected

Safety

is the practice of ensuring AI systems operate without causing harm

AI trustworthiness and safety cannot be achieved without AI security.

https://github.com/Xci-Gen/ai_security