AI for Security and Security for AI

Dr. Areeb Ahmed

Machine Learning & Language Technologies Lab
Faculty of Computer and Information Science
University of Ljubljana

Presentation Focus:

AI is transforming cybersecurity. AI is now used for defense, automation, and decision-making, but it is also becoming a target for attackers.

Key Message:

AI is not only changing cybersecurity tools; it is changing the nature of cyber threats.

SMASH
machine learning for science and humanities postdoctoral program

AI AND ML FOR SCIENTIFIC APPLICATIONS THROUGH SECURE COMMUNICATIONS FOR 5G/6G

AREEB AHMED

The project was supported by the European Union's Horizon Europe Research and Innovation Program under the Marie Skłodowska-Curie Program, SMASH, co-funded under grant agreement No. 101081355.

"Data Science - Machine Learning for Scientific Applications"

subarea "Beyond Supervised Learning"

under

Prof. Dr. Zoran Bosnić,

Machine Learning and Language Technologies Lab,

Faculty of Computer and Information Science,

University of Ljubljana.

Academic Partner

Prof. Ertugrul Basar,

Head of Wireless Systems Group at Tampere Wireless Research Center

Tampere University, Finland

Industry Partner

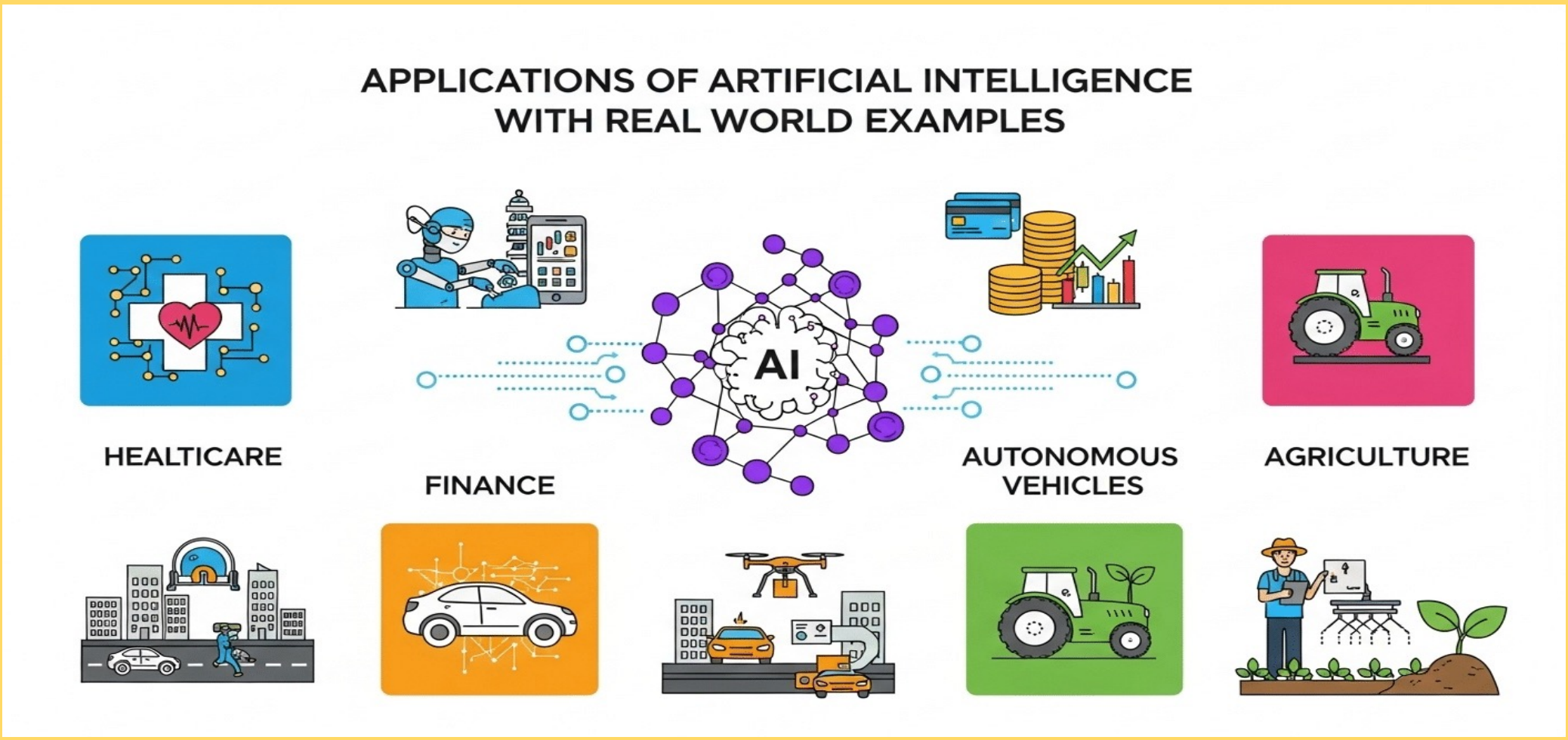
Tomi Ilijaš,

CEO & Founder,

ARCTUR, Nova Gorica, Slovenia

Co-funded by the European Union

AI for Security and Security for AI



AI for Security and Security for AI

Healthcare:
AI supports diagnosis, medical image analysis, patient monitoring, and treatment recommendations.

Finance:
Banks use AI for fraud detection, credit risk analysis, transaction monitoring, and automated customer support.

Industry:
Manufacturing uses AI for predictive maintenance, quality control, robotics, and process optimization.

Cybersecurity

AI helps detect malware, identify abnormal behavior, analyze threats, and automate incident response.

Key Point:

As AI becomes essential to operations, attacking AI systems becomes more attractive to cybercriminals.

Evolution

Introduction

Motivation

AI for Security and Security for AI

Why Cybersecurity Must Evolve?

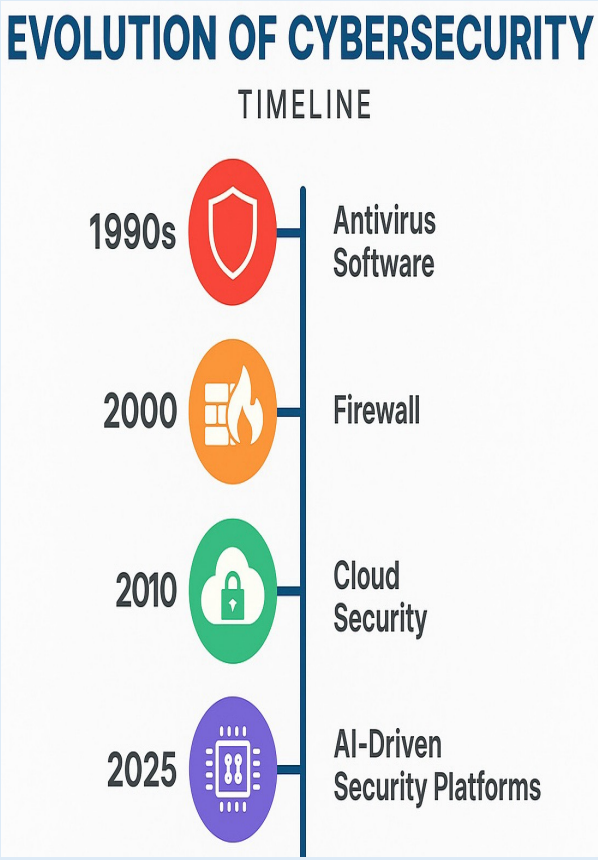
Traditional Security Faces New Challenges!

Static software is easier to protect:
Traditional systems usually behave in predictable ways and can be protected with rules, signatures, and access controls.

AI systems are dynamic:
AI models may change after retraining, new data, or deployment in new environments.

AI decisions can be hard to explain:
Many AI systems operate as black boxes, making it difficult to understand why they made a decision.

Attack surfaces are expanding:
Attackers can target data, models, APIs, prompts, users, and deployment pipelines.



Core Question:
How do we secure systems that learn, adapt, and make autonomous decisions?

AI for Security and Security for AI

AI as a Cybersecurity Tool?

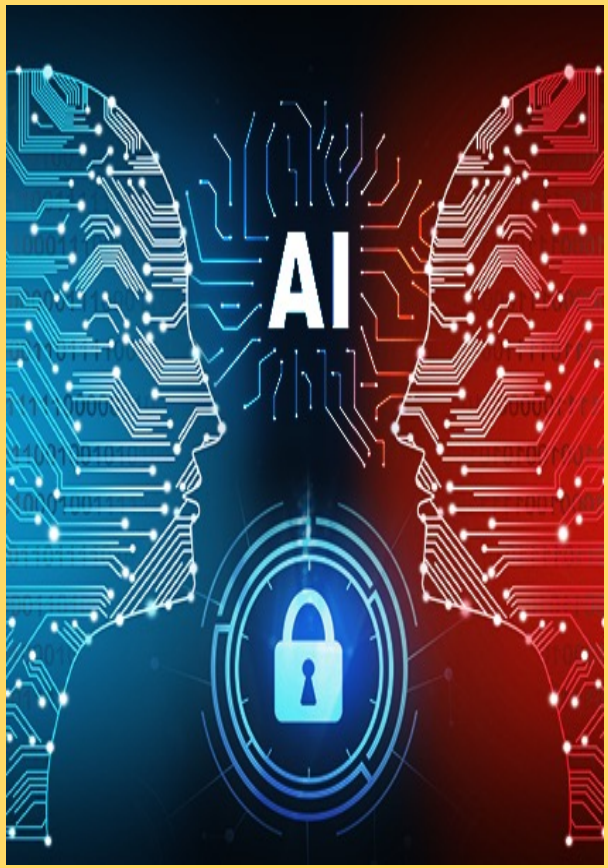
AI Strengthens Cyber Defense!

Threat detection:
AI can analyze large amounts of security data and identify suspicious patterns faster than humans.

Malware classification:
Machine learning can classify unknown files by detecting behavioral or structural similarities.

Fraud detection:
AI identifies unusual financial transactions and suspicious user behavior.

Security automation:
AI can prioritize alerts, reduce analyst workload, and support faster incident response.



Key Benefit:

AI improves speed, scalability, and accuracy in modern cybersecurity operations

AI for Security and Security for AI

AI as a Cybersecurity Target?

AI Systems Are Also Vulnerable!

Models can be attacked:
Attackers may manipulate inputs to force incorrect predictions.

Training data can be corrupted:
If attackers poison training data, the model may learn harmful behavior.

APIs can expose models:
Repeated queries to AI services can reveal model behavior or enable model theft.

AI supply chains can be compromised:
Pre-trained models, datasets, plugins, and libraries may contain hidden risks.

Key Message:
AI is both a defensive tool and a new target for cyber attacks.

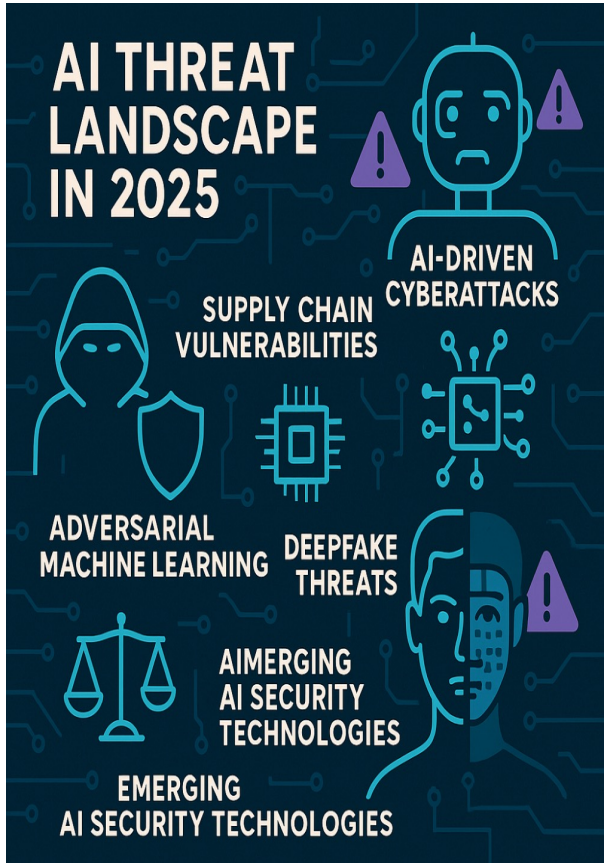
AI for Security and Security for AI

Adversarial examples:
Inputs are carefully modified so an AI system makes wrong decisions.

Data poisoning:
Attackers inject malicious or misleading data into training datasets.

Model extraction:
Attackers steal or imitate a model by repeatedly querying it.

Prompt injection:
Malicious prompts manipulate generative AI systems into unsafe behavior.



Key Observation:
Many AI threats target trust, decision reliability, and data integrity.

AI for Security and Security for AI

Adversarial AI Attacks

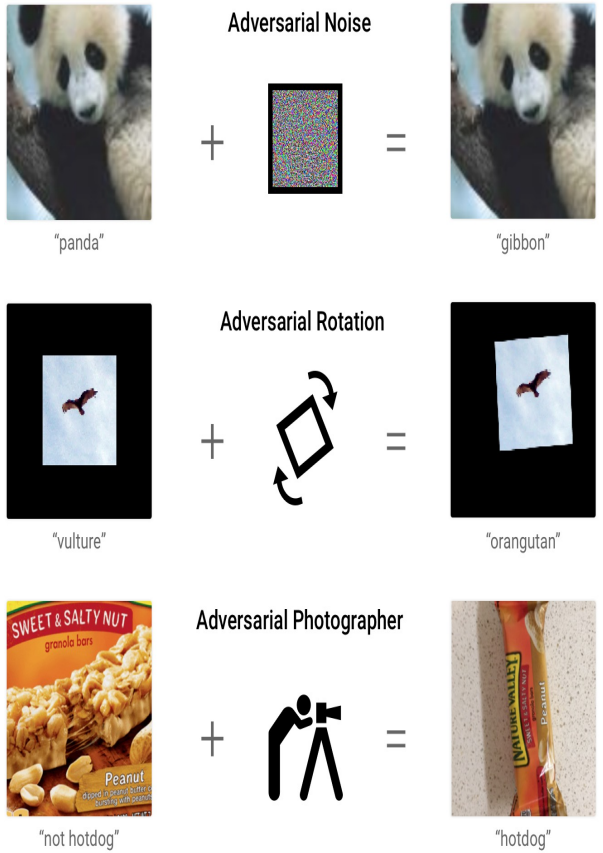
Fooling Intelligent Systems

What they are:
Adversarial attacks modify input data in subtle ways to mislead AI models.

Why they work:
AI models often rely on statistical patterns rather than human-like understanding.

Example — images:
A tiny change in pixels may cause an image classifier to misidentify an object.

Example — autonomous systems:
A modified traffic sign could cause a self-driving system to make a wrong decision.



Main Risk:
Small changes can lead to large consequences in safety-critical systems.

Real-World Impact of Adversarial AI

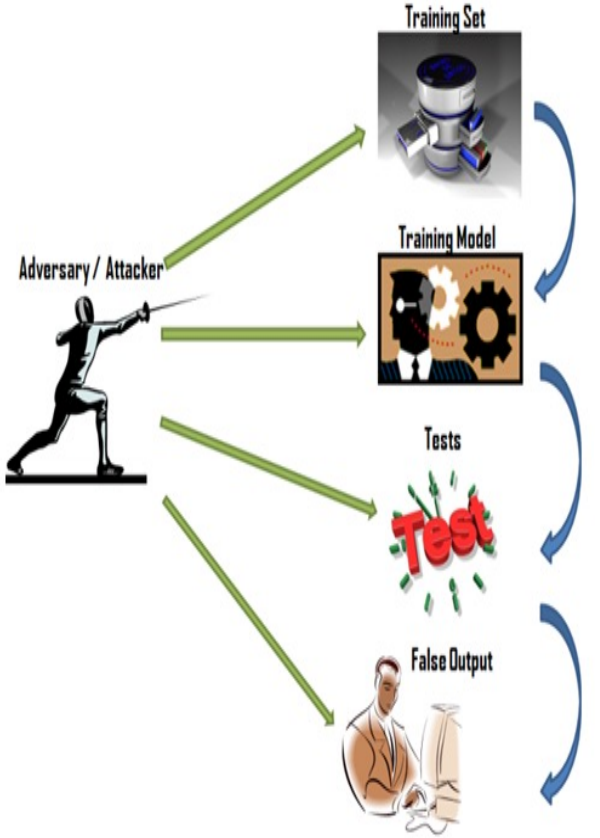
Small Changes, Major Consequences

Autonomous vehicles:
A manipulated road sign may cause incorrect navigation or unsafe driving behavior.

Medical AI:
Small image changes may influence diagnosis, leading to incorrect clinical decisions.

Biometric systems:
Facial recognition can potentially be bypassed using adversarial patterns.

Fraud systems:
Attackers may slightly change transaction behavior to avoid detection.



Key Message:
Adversarial AI attacks can move from digital manipulation to real-world harm.

AI for Security and Security for AI

Data Poisoning Attacks

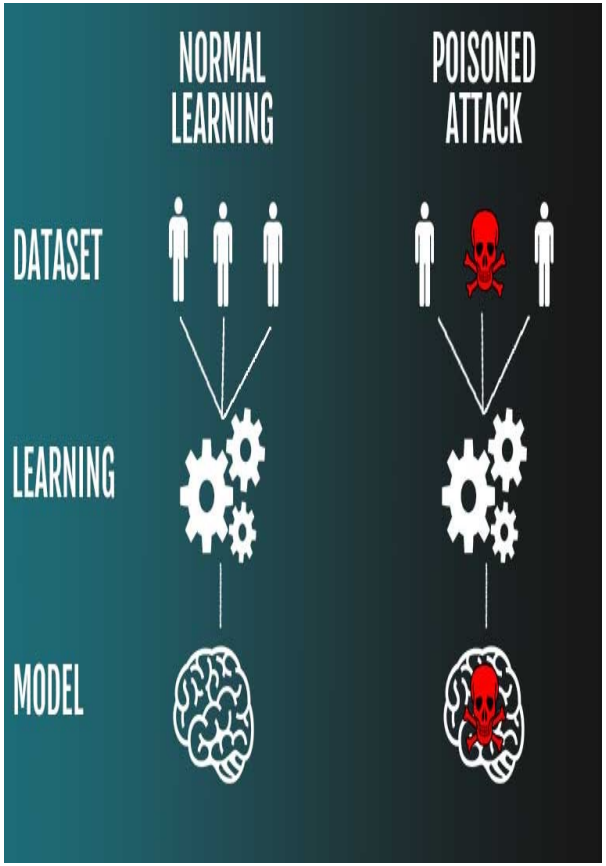
Attacking the Learning Process

What happens:
Attackers manipulate training data so the AI learns incorrect or harmful behavior.

Why it is dangerous:
The attack becomes embedded inside the model and may remain hidden after deployment.

Example:
A poisoned dataset may teach an AI system to misclassify certain objects or users.

Impact:
Poisoning can reduce accuracy, create bias, or introduce hidden vulnerabilities.



Key Point:
AI security begins with trusted, verified, and well-governed data.

SMASH

machine learning for science and humanities · postdoctoral program

Backdoor Attacks

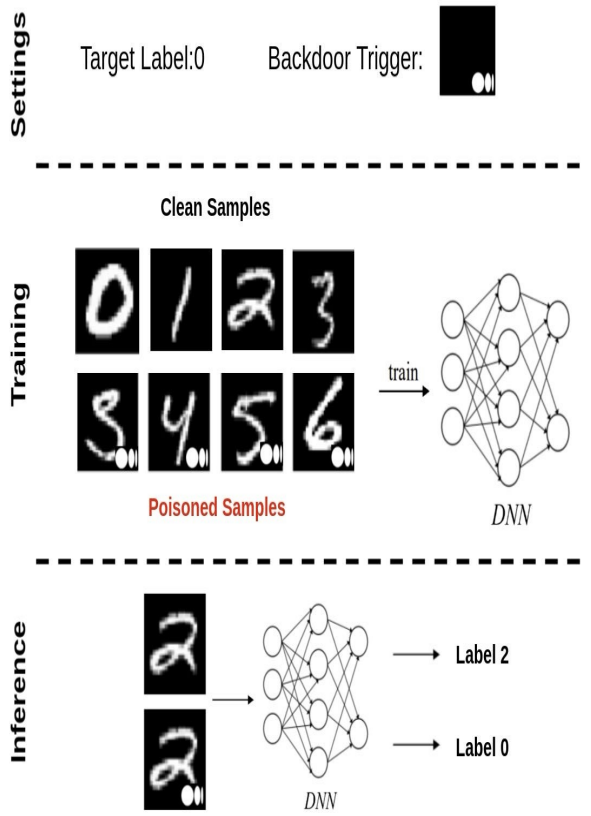
Hidden Triggers Inside AI Systems

What is a backdoor:
A hidden behavior inside an AI model that activates only when a specific trigger appears.

Normal behavior:
The model may perform well during ordinary testing and appear safe.

Triggered behavior:
When the hidden trigger appears, the model may make a malicious or incorrect decision.

Example triggers:
A sticker, symbol, phrase, image pattern, or specific input sequence.



Key Risk:
Backdoors are difficult to detect because they remain inactive most of the time.

Model Theft and Extraction

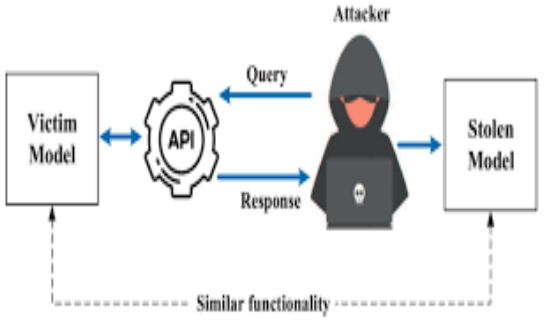
AI Models as Strategic Assets

Why models are valuable:
Large AI models require expensive data, computing power, expertise, and time.

API abuse:
Attackers can repeatedly query a model and learn how it behaves.

Model replication:
The attacker may build a copy that performs similarly to the original model.

Business impact:
This can lead to intellectual property loss, competitive damage, and revenue loss.



Key Message:
AI models should be protected like critical digital assets.

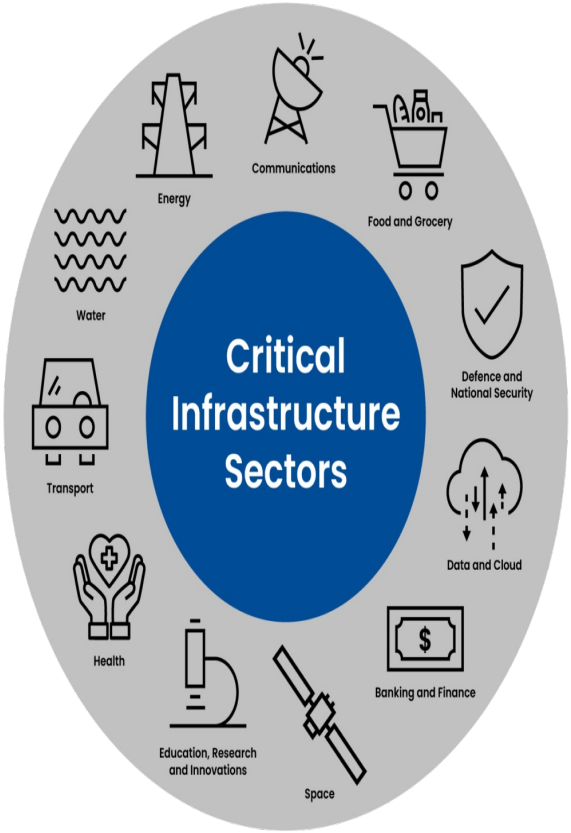
AI for Security and Security for AI

Energy systems:
AI supports grid monitoring, demand prediction, and fault detection.

Transportation:
AI supports routing, autonomous vehicles, traffic control, and logistics.

Healthcare:
AI supports diagnosis, triage, patient monitoring, and clinical decision support.

Industrial automation:
AI controls robotics, quality assurance, and predictive maintenance.



Key Risk:
Compromised AI in critical infrastructure can cause physical, economic, and societal harm.

AI for Security and Security for AI

Secure data collection:
Datasets should be verified, documented, and protected from manipulation.

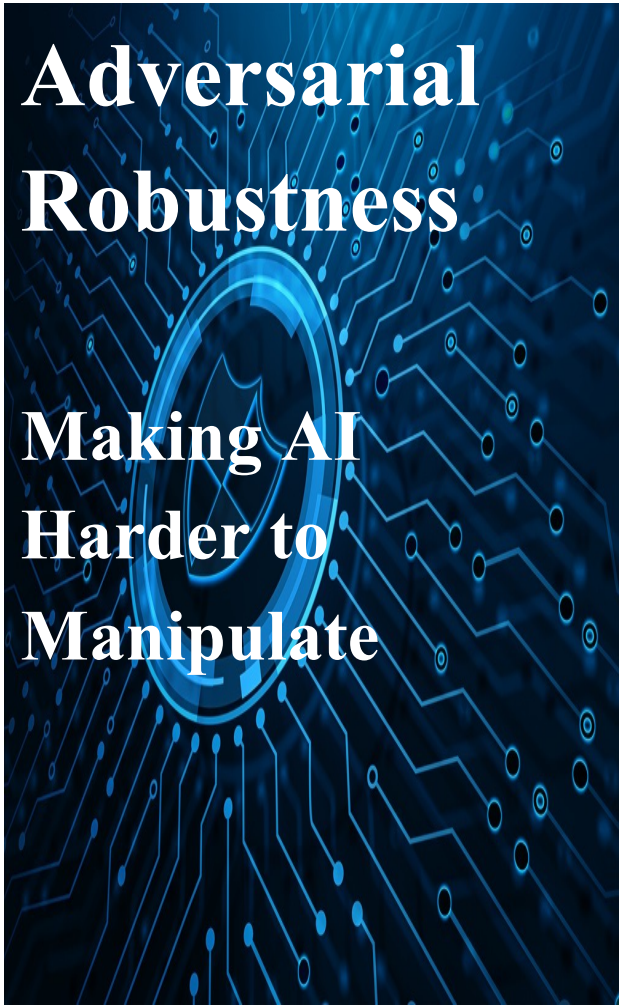
Secure model design:
Architectures should consider robustness, privacy, and explainability from the beginning.

Secure training:
Training pipelines should include access control, logging, and integrity checks.

Secure deployment:
Models should be protected through API security, monitoring, and access restrictions.

Key Message:
Security should be built into AI systems, not added after deployment.

AI for Security and Security for AI

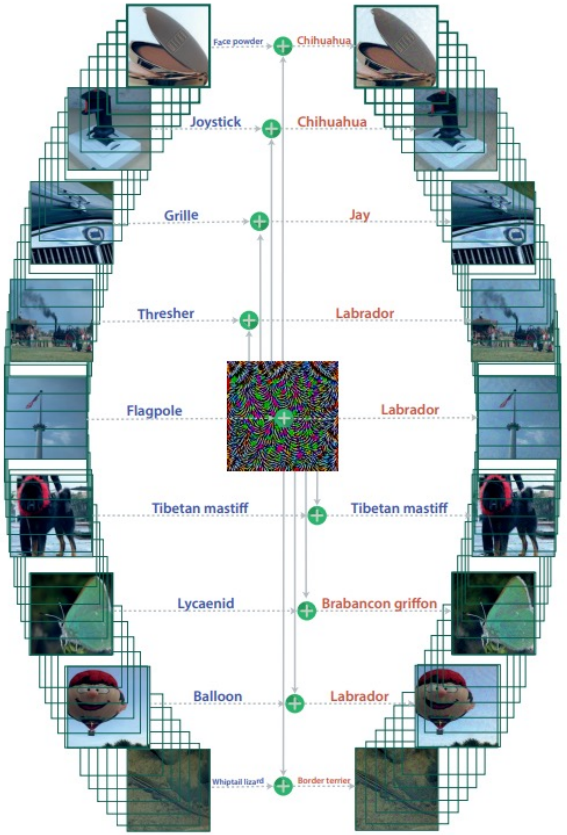


Adversarial training:
Models are trained using adversarial examples to improve resistance.

Input sanitization:
Inputs are checked or transformed before reaching the AI model.

Ensemble learning:
Multiple models are combined to reduce dependence on one vulnerable model.

Certified robustness:
Mathematical guarantees are used to prove resistance within limited attack boundaries.



Limitation:
No single method provides complete protection against all adversarial attacks.

AI for Security and Security for AI

Adaptive resilience:
AI systems should detect changes and adjust defenses automatically.

Randomized defenses:
Introducing controlled randomness can make attacks harder to predict and repeat.

Noise-driven architectures:
Noise can be used strategically to increase unpredictability and reduce exploitability.

Self-healing systems:
Future AI systems may detect compromise and recover without full human intervention.

Research Connection:
Unconventional cybersecurity approaches for securing AI-integrated systems.

AI for Security and Security for AI

AI for Security vs Security for AI:
Future attacks and defenses may both be powered by autonomous AI systems.

Autonomous cyber defense:
AI may detect, respond to, and recover from threats in real time.

Secure multi-agent systems:
As AI agents collaborate, new risks emerge around coordination, trust, and control.

Quantum-era cybersecurity:
Future advances in computing may change encryption, data protection, and AI security.

Key Question:
Can AI systems eventually defend themselves against intelligent attackers?

AI for Security and Security for AI

Research Challenges

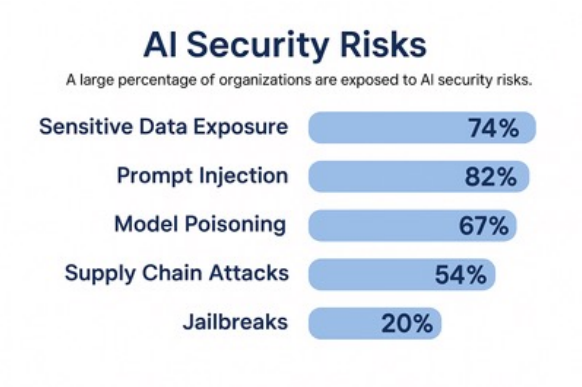
Open Problems in AI Security

Unknown attack vectors:
New AI capabilities will create threats that are not yet understood.

Supply-chain security:
Datasets, models, libraries, and plugins all introduce hidden dependencies.

Robustness-performance trade-off:
More secure models may require more computation or may reduce accuracy.

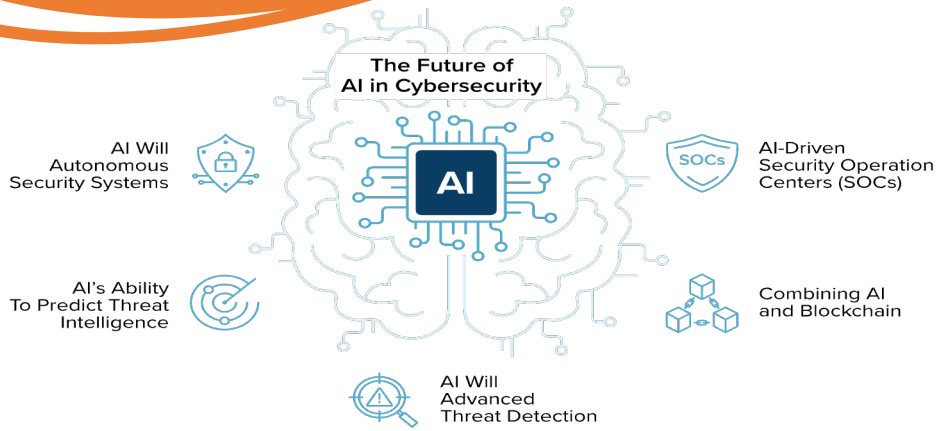
Multi-agent risks:
AI agents interacting with each other may create unpredictable security failures.



Key Message:
AI cybersecurity remains an open and rapidly evolving research field.

AI for Security and Security for AI

Conclusion – Final Thoughts



AI creates opportunities:
It improves automation, detection, productivity, and decision-making.

AI creates risks:
It introduces new attack surfaces across data, models, APIs, users, and governance.

Traditional security is insufficient alone:
AI requires adaptive, lifecycle-based, and resilience-oriented security.

Collaboration is essential:
Researchers, industry, policymakers, and security professionals must work together.

Final Statement:
"The future of trustworthy artificial intelligence depends on how effectively we secure it today."

Thank You
Questions, suggestions & discussion! Now or on
areeb.ahmed@fri.uni-lj.si

References

1. [1] A. Ahmed and F. A. Savaci, “Covert electromagnetic nanoscale communication system in the terahertz channel,” *Journal of Circuits, Systems and Computers*, vol. 29, no. 8, Art. no. 2050126, 2020.

2. [2] A. Ahmed and F. A. Savaci, “Structure and performance evaluation of fractional lower-order covariance method in alpha-stable noise environments,” *Recent Advances in Electrical & Electronic Engineering*, vol. 12, no. 1, pp. 40–44, 2019, doi: 10.2174/2352096511666180419143436.

3. [3] F. A. Savaci and A. Ahmed, “Inverse system approach to design alpha-stable noise driven random communication system,” *IET Communications*, vol. 14, no. 6, pp. 910–913, 2020.

4. [4] A. Ahmed and Z. Bosnić, “A covert α-stable noise-based extended random communication by incorporating multiple inverse systems,” *IEEE Access*, vol. 13, pp. 13675–13685, 2025.

5. [5] A. Ahmed, A. Siddqui, R. Khan, G. Nabi, and M. Ali, “Comparative analysis of OFDM in AWGN & local acoustic channel,” in *Proc. 2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)*, 2024, pp. 1–4.

6. [6] A. Ahmed and F. A. Savaci, “Blind recognition of alpha-stable random carrier signals by an eavesdropper in random communication systems,” *IET Communications*, vol. 13, no. 16, pp. 2420–2427, 2019.

7. [7] A. Ahmed and Z. Bosnić, “Machine learning for enabling high-data-rate secure random communication: SVM as the optimal choice over others,” *Mathematics*, vol. 13, no. 22, Art. no. 3590, 2025. [Online]. Available: Publisher page.

8. [8] L. Capogrosso, F. Cunico, D. S. Cheng, F. Fummi, and M. Cristani, “A machine learning-oriented survey on tiny machine learning,” *IEEE Access*, vol. 12, pp. 23406–23426, 2024.

9. [9] S. V. Balkus, H. Wang, B. D. Cornet, C. Mahabal, H. Ngo, and H. Fang, “A survey of collaborative machine learning using 5G vehicular communications,” *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1280–1303, Secondquarter 2022.

10. [10] A. Celik and A. M. Eltawil, “At the dawn of generative AI era: A tutorial-cum-survey on new frontiers in 6G wireless intelligence,” *IEEE Open Journal of the Communications Society*, vol. 5, pp. 2433–2489, 2024. [Online]. Available: IEEE Xplore.

11. [11] E. Basar, “Noise modulation,” *IEEE Wireless Communications Letters*, vol. 13, no. 3, pp. 844–848, Mar. 2024, doi: 10.1109/LWC.2023.3346471.

12. [12] B. L. Basore, “Noise-like signals and their detection by correlation,” Massachusetts Institute of Technology, Cambridge, MA, USA, Tech. Rep. 7, 1952. [Online]. Available: MIT repository.

13. [13] C. Gu and X. Cao, “Research on information hiding technology,” in *Proc. International Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2012, pp. 2035–2037.

14. [14] A. B. Salberg and A. Hanssen, “Secure digital communications by means of stochastic process shift keying,” in *Proc. 33rd Asilomar Conference on Signals, Systems, and Computers*, vol. 2, 1999, pp. 1523–1527.

15. [15] A. Ahmed and F. A. Savaci, “Random communication system based on skewed alpha-stable levy noise shift keying,” *Fluctuation and Noise Letters*, vol. 16, no. 3, Art. no. 1750024, 2017.

16. [16] M. E. Cek and F. A. Savaci, “Stable non-Gaussian noise parameter modulation in digital communication,” *Electronics Letters*, vol. 45, no. 24, pp. 1256–1257, 2009.

17. [17] M. E. Cek, “Covert communication using skewed α-stable distributions,” *Electronics Letters*, vol. 51, no. 1, pp. 116–118, 2015.

18. [18] Z. Xu, W. Jin, K. Zhou, and J. Hua, “A covert digital communication system using skewed α-stable distributions for Internet of Things,” *IEEE Access*, vol. 8, pp. 113131–113141, 2020, doi: 10.1109/ACCESS.2020.3003561.

19. [19] A. Janicki and A. Weron, *Simulation and Chaotic Behavior of α-Stable Stochastic Processes*. New York, NY, USA: Marcel Dekker, 1994, pp. 30–75.

20. [20] A. Ahmed and F. A. Savaci, “Measure of covertness based on the imperfect synchronization of an eavesdropper in random communication systems,” in *Proc. 10th International Conference on Electrical and Electronics Engineering (ELECO)*, 2017, pp. 638–641.

21. [21] A. Ahmed and F. A. Savaci, “Synchronisation of alpha-stable levy noise-based random communication system,” *IET Communications*, vol. 12, no. 3, pp. 276–282, 2018.

22. [22] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Jan. 2023. [Online]. Available: NIST AI RMF.

23. [23] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, National Institute of Standards and Technology, NIST AI 100-2 E2025, Mar. 2025, doi: 10.6028/NIST.AI.100-2e2025.

24. [24] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, Jul. 2024, doi: 10.6028/NIST.AI.600-1.

25. [25] MITRE, “ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.” Accessed: Sep. 7, 2026. [Online]. Available: MITRE ATLAS.

26. [26] OWASP Foundation, “OWASP Top 10 for LLM and generative AI applications,” 2025. Accessed: Sep. 7, 2026. [Online]. Available: OWASP GenAI Security Project.

27. [27] National Cyber Security Centre, *Guidelines for Secure AI System Development*, ver. 1.0, Nov. 27, 2023. [Online]. Available: NCSC guidelines.