Wrocław University of Science and Technology

unite! University Network for Innovation, Technology and Engineering

HR EXCELLENCE IN RESEARCH

Evaluated by IEP INSTITUTIONAL EVALUATION PROGRAMME www.iep-qaa.org

INTELLIGENT CRIMES: DISRUPTING AI-CONTROLLED SYSTEMS, LARGE-SCALE BLACKMAIL

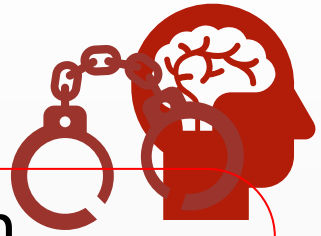
Mateusz Guściora, 228884

Wed. (even), 9:15 - 11:00 AM

November2023

Wroclaw University of Science and Technology

AGENDA

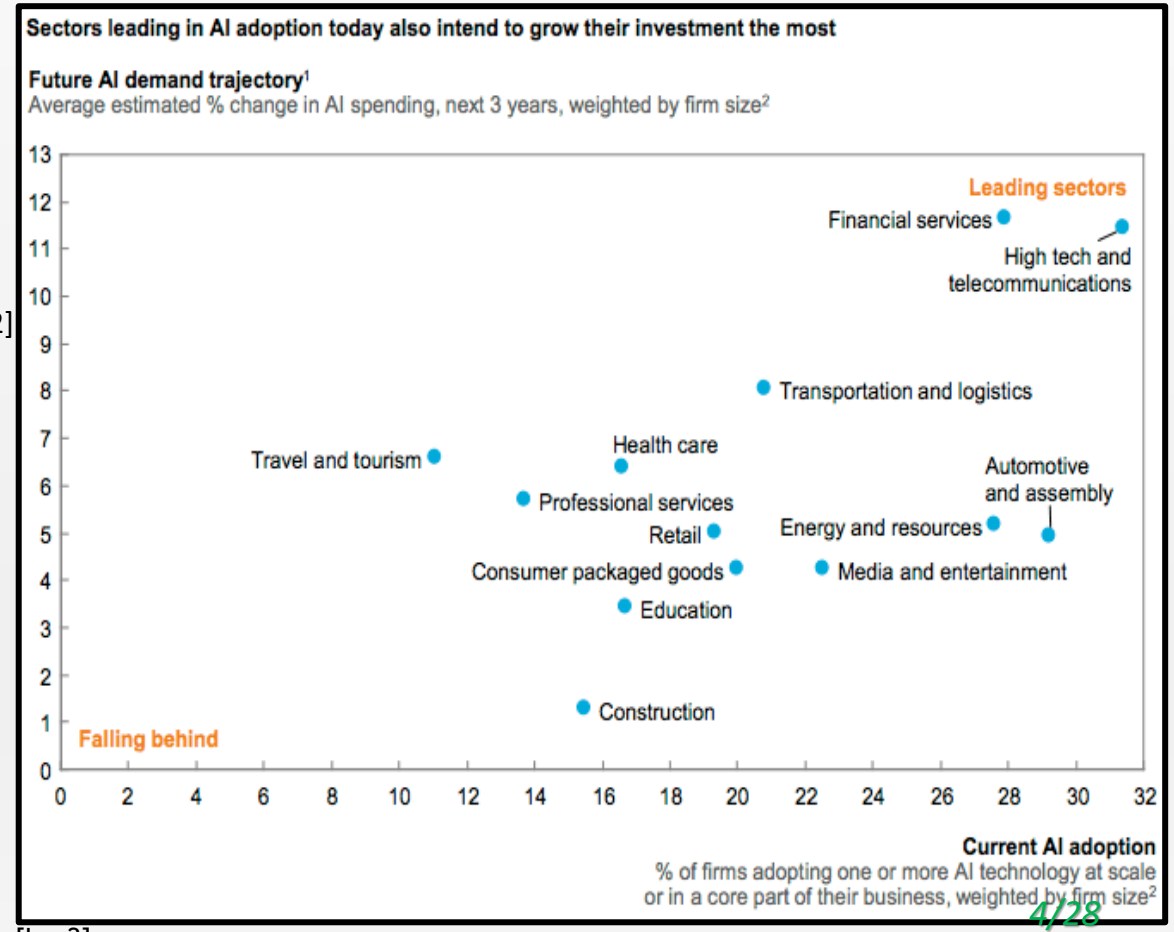


1. Introduction to the Topic
2. Evolution of AI-related Crimes
3. Overview of intelligent crimes on AI
4. Anatomy of AI systems Disruption - vulnerabilities
5. Anatomy of AI systems Disruption - attacks
6. Perpetrators and Motivations
7. Attack Methodologies
8. Real life Cases of AI-targeted attacks
9. AI as a Tool for Disruption
10. Large-Scale Blackmail
11. Real life Cases of AI-assisted attacks
12. Cybersecurity Measures
13. Future Challanges
14. Conclusion

AI AS A DOUBLE-EDGED SWORD I [1]

- AI is a technological system that designed to simulate or replicate human cognitive functions[1]
- Critical to Healthcare, Finance, Transportation, Security & More [2]
- Role of AI in managing infrastructure, data analysis, and decision-making processes.

[2]

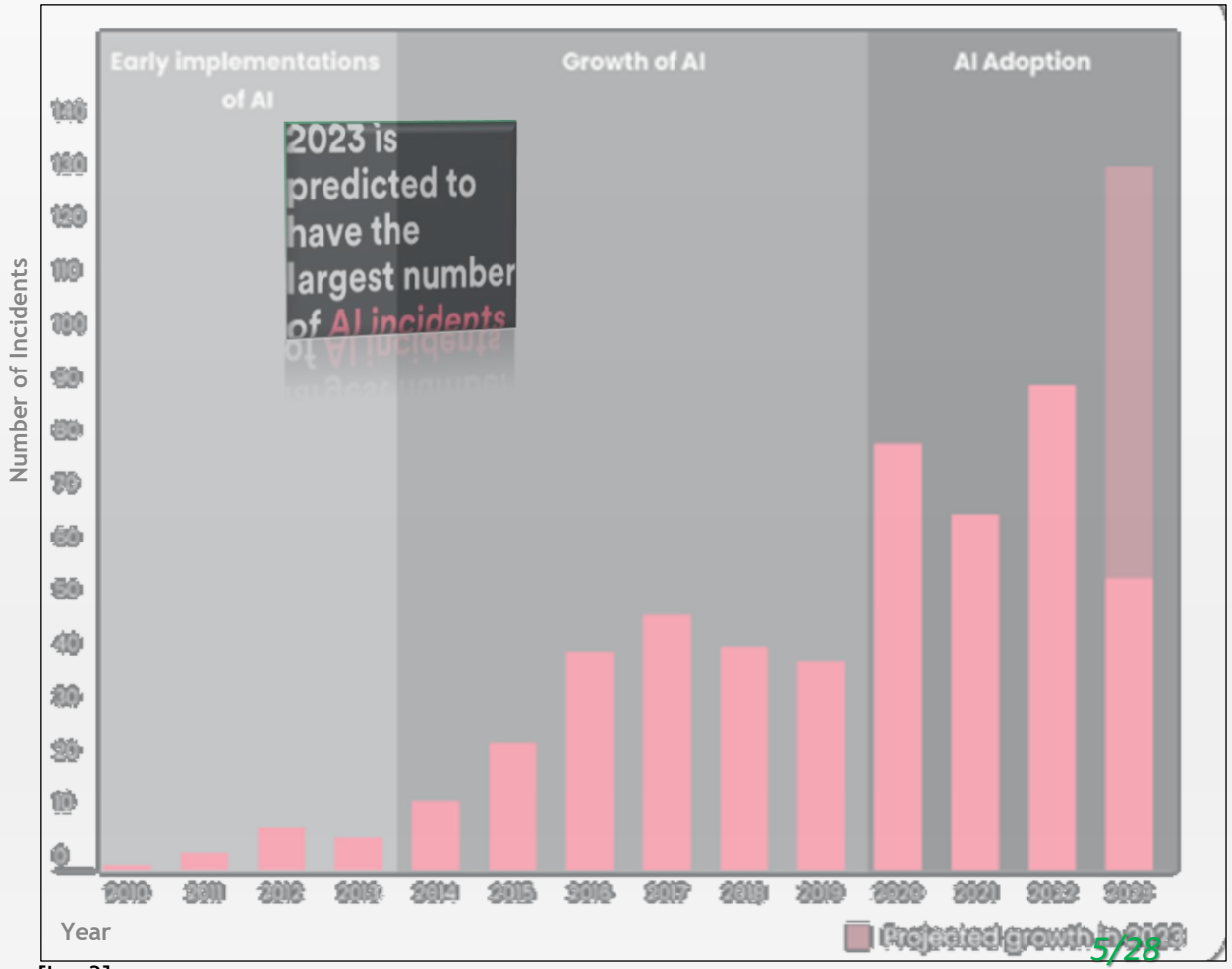


[Img2]

AI AS A DOUBLE-EDGED SWORD II[3]

- Intelligent Crimes - a sophisticated class of illegal activities[3] [5]
- Vulnerability and Attraction of AI systems for criminal purposes.
- Purpose of this presentation!

[4]



[Img3]

EVOLUTION OF AI-RELATED CRIMES [5]

Early AI Exploits

- Examples: ELIZA (1966) , Early Spam Filters (1990s), Adversarial Attacks on Image Recognition (Early 2000s)

Advent of Sophisticated Attacks and Recent Trends

- Transition to more sophisticated strategies
- Adversarial attacks, data poisoning, model inversion attacks, model stealing, or extraction attack, exploiting Bias in models, AI-enabling Cyberattacks
- Examples: Microsoft's Tay (2016), Deepfakes (2017-present) [6]

OVERVIEW OF INTELLIGENT CRIMES AGAINST AI[7][9]

Intelligent Crimes Targeting AI

- Specific nature of crimes.
- Increase in intelligent crimes coincides with AI integration into critical systems. [2] [4]

Why AI Systems

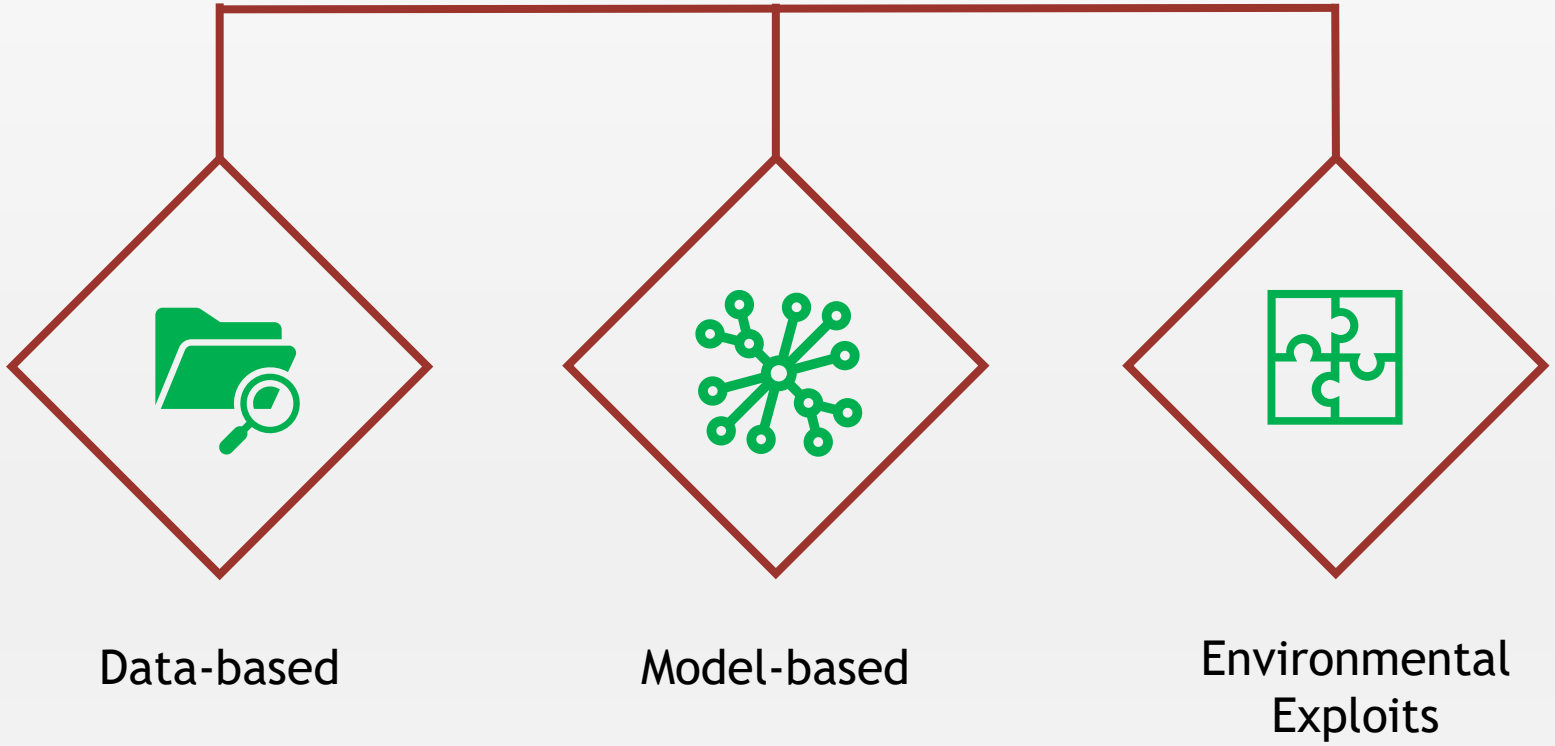
- High-value data repositories and decision-making disruption
- Discreet manipulation with significant impacts.

Typical AI Vulnerabilities

- Data poisoning, adversarial examples, model theft, and more.

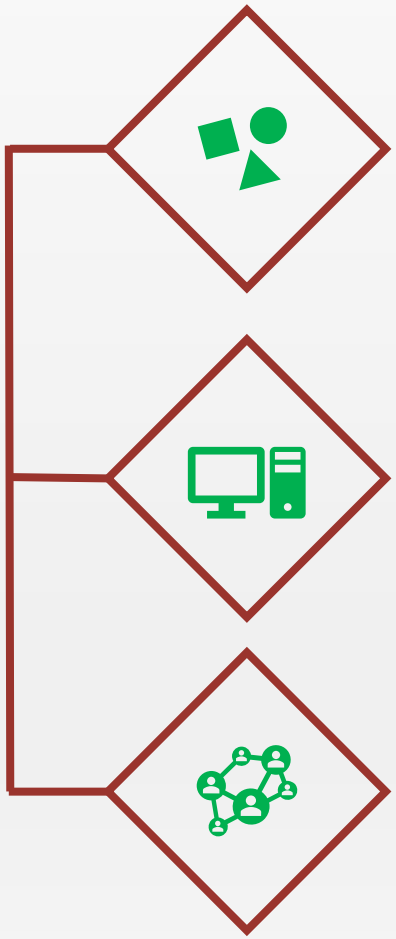
ANATOMY OF AI SYSTEM DISRUPTION I – VULNERABILITIES[8]

Classification of common vulnerabilities



ANATOMY OF AI SYSTEMS DISRUPTION II - ATTACKS

Categorizing disruption methods[8]



Physical

Tampering with the AI's hardware or operational environment.
Example: Attacks on autonomous vehicle sensors using stickers or paint.

Digital

Exploitation of AI algorithms and data inputs.
Example: The exploitation of image recognition systems using adversarial images.

Social engineering

Trickery or manipulation of users.
Example: Social bots influencing stock market predictions by spreading misinformation.

Wroclaw University of Science and Technology

ANATOMY OF AI SYSTEMS DISRUPTION III - STAGES [8]

Training Phase Attacks

Testing Phase Attacks

Deployment Phase Attacks

Post-Deployment Attacks

PERPETRATORS AND MOTIVATIONS BEHIND AI DISRUPTIONS[9]

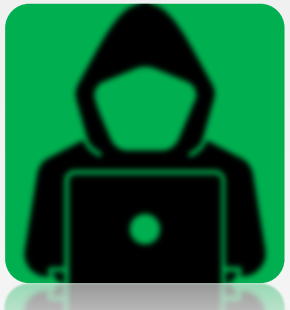
Who is Behind the Disruptions?

Common culprits!

State-Sponsored: Espionage & Sabotage

Cybercriminals: Financial Gain & Ideology

Competitors & Insiders: Market Advantage & Grievances



Technical Deep Dive into Attack Methodologies – Part 1[9] [10]

AI Model Poisoning/Data Poisoning/Algorithm Poisoning

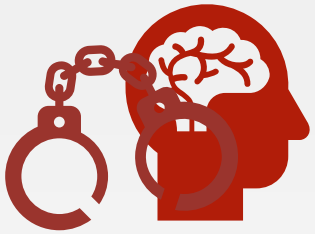
Malicious data can be inserted into the data set, causing the AI to make incorrect generalization

Adversarial Attacks

Even small perturbations in input data can deceive AI

Exploiting Model Biases and loopholes

Attackers can take advantage of inherent biases in AI models



Technical Deep Dive into Attack Methodologies – Part 2[9] [10]

Extraction attack - Model Inversion Attacks – Model stealing Attacks

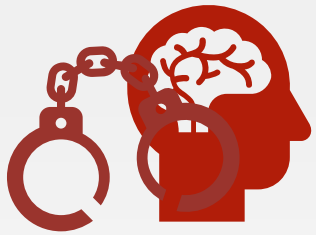
Attackers can reverse-engineer the model or input data to gain sensitive information.

Oracle attack System Infiltration Techniques- backdooring the model

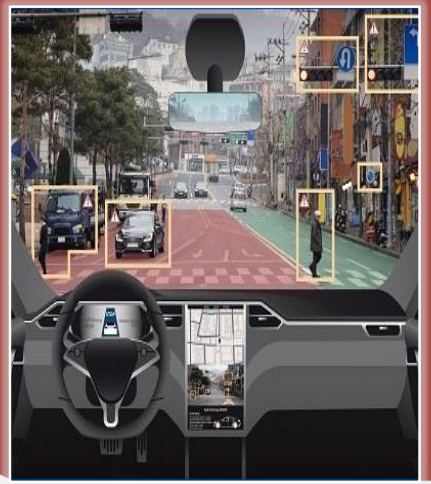
Methods like code injection, buffer overflows, and exploiting system vulnerabilities specific to AI architectures.

Other

Botnets , Attack on Supply chain, Resource exhaustion attacks, Side-channel attacks on HW…

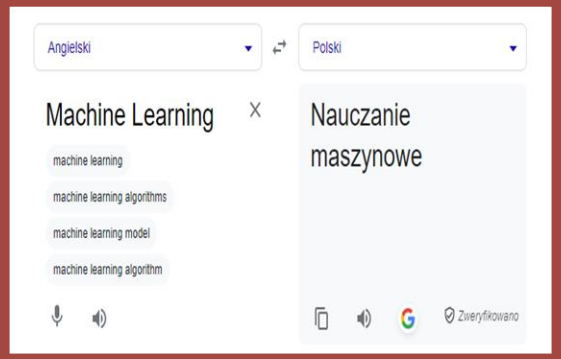


CASE STUDIES OF AI-CONTROLLED SYSTEM DISRUPTION[6]



[Img4]

[Img5]



[Img6]

[Img7]

Hacking of Autonomous Vehicles: Attack on Tesla's Autopilot System (2019)

Microsoft's Tay AI Twitter Bot (2016)

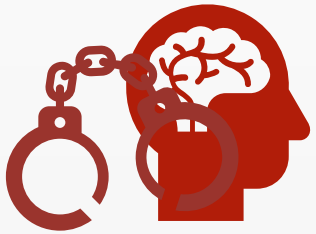
- malicious users tweet offensive language at Tay, which led to Tay generating similar

Attack on Machine Translation Service (2022) - Google Translate, Bing Translator, and Systran Translate

Bypassing automated identity verification system(2021) – fraud 3.4 mln $

Wroclaw University of Science and Technology

AI AS A THREAT: DISRUPTION AND WEAPONIZATION[9][10][11]



- **Cybersecurity Threats Enhanced by AI**
  - Phishing, malware, and APTs,
  - Large Scale blackmail
- **Deepfakes in Misinformation**
  - Undermining trust and reality
- **AI in Autonomous Weapons**
  - Ethical/Moral dilemmas
- **AI in Market Manipulation**
  - Financial fraud and instability
- **AI-Enabled Surveillance**
  - Privacy erosion

LARGE-SCALE BLACKMAIL VIA AI[5]

- Blackmail involves coercing individuals or organizations to act against their will, often for financial gain in exchange for preventing the wrongdoing or release of sensitive information.
- With AI, such schemes can escalate in scale and sophistication. This is due to the technology's ability to harvest and process vast amounts of data, identify vulnerabilities, personalize attacks, and automate extortion processes.

Wroclaw University of Science and Technology

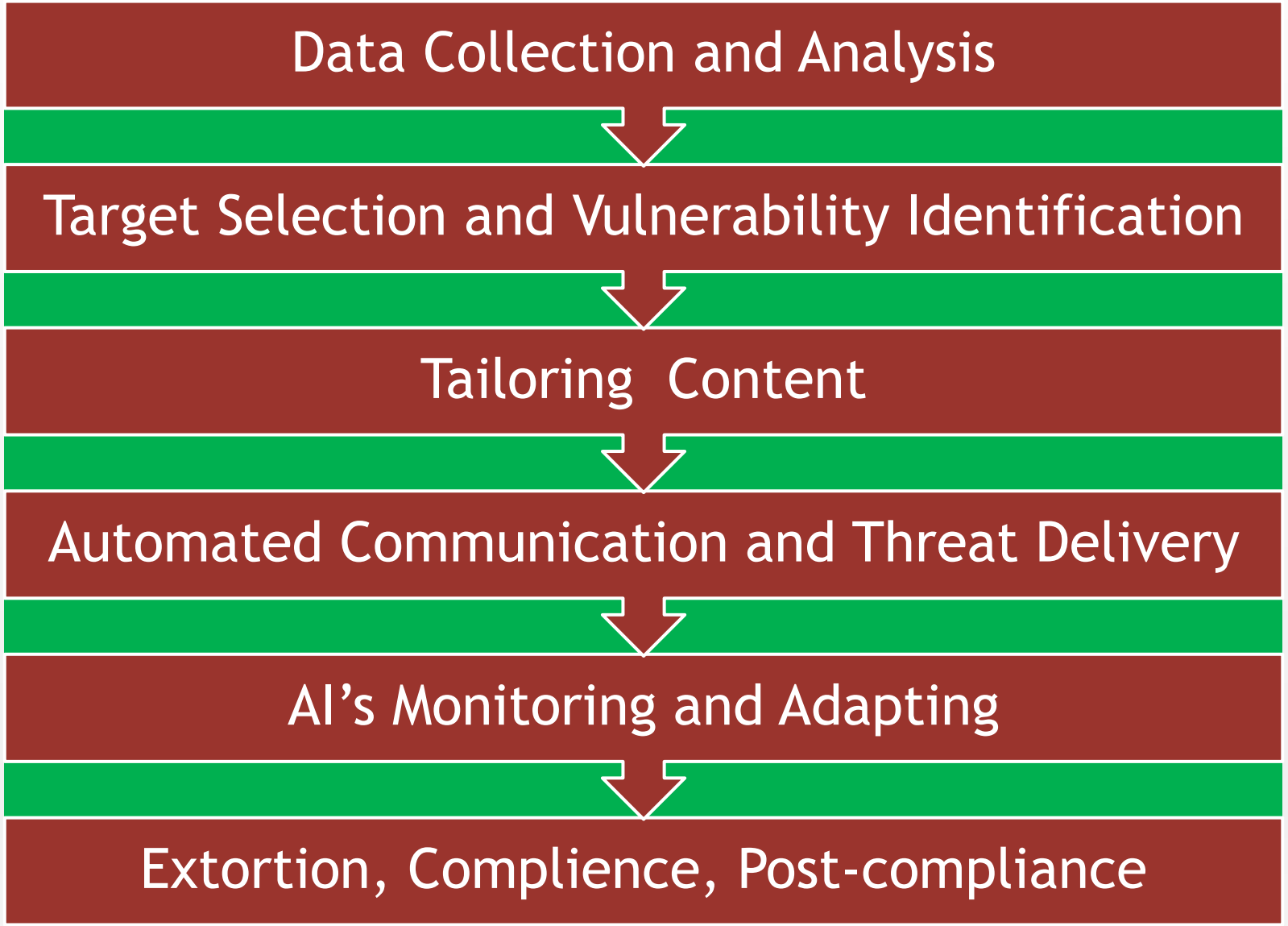
LARGE-SCALE BLACKMAIL VIA AI – KEY ASPECTS

[5]

- AI's Role in Enhancing Blackmail Efficiency and Automation.
- Data Breaches: The Fuel for AI-Enabled Blackmail Schemes
- Scaling Targets: High-Profile Targets or Mass Targeting.
- Personalized Blackmail at Scale: Tailoring Channel of communication, Content of Threat, and Targets.
- Is Automated

Wroclaw University of Science and Technology

LARGE-SCALE BLACKMAIL – TYPICAL FLOW[5]



CASE STUDIES OF AI-ASSISTED CYBER ATTACKS

[12] [13][14]

[Img8]

Over 20 000 WordPress sites infected with a botnet-style cyber attack

[Img9]

IG (August 2019) and (November 2019) cyber attack and data breach

[Img10]

TaskRabbit - Attack (2018), botnets carry ddos attack leading to the temporary suspension

[Img11]

Hong Kong Bank Heist (2020) - deepfake AI was used to clone the voice of a company director

CYBERSECURITY MEASURES AGAINST AI SYSTEM DISRUPTION – DEFENSE STRATEGIES[17]

Robust Encryption
- Protecting data integrity

Adversarial Training
- Preparing AI for attack scenarios.

Anomaly Detection Systems
- Identifying unusual activities

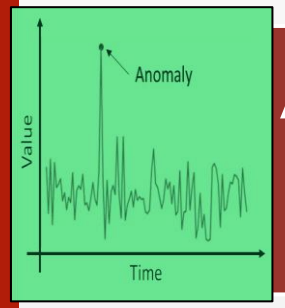
Regular Audits and Updates
- Ensuring AI behaves as expected

User Education and Awareness
- Training individuals on security best practices.

Access Control and Monitoring
- Implementing strict access policies and surveillance.

Legal and Ethical Considerations
- Guiding responsible AI development

CYBERSECURITY MEASURES AGAINST AI SYSTEM DISRUPTION– AI IN DEFENSE[17]



AI-Powered Anomaly Detection

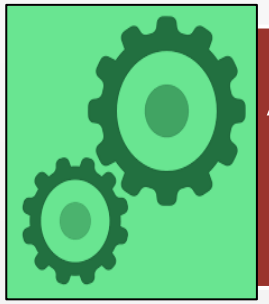
[Img12]

Utilizing AI to monitor for unusual behavior.

Predictive Threat Intelligence

[Img12]

AI for forecasting potential security incidents.



Automated Incident Response

[Img12]

Employing AI to respond to threats swiftly.

ML for Real-Time Threat Detection

[Img12]

ML algorithms for immediate threat identification.

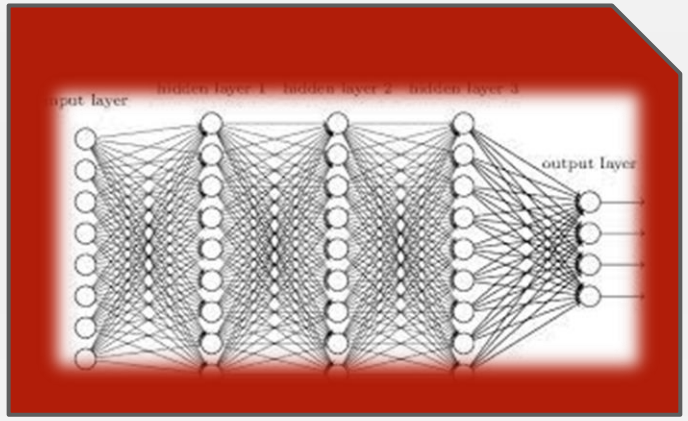
Wroclaw University of Science and Technology

CYBERSECURITY MEASURES AGAINST AI SYSTEM DISRUPTION– COLLABORATIVE EFFORTS [15] [18]

Public-Private Partnerships	• Uniting government and industry efforts in cybersecurity.
International Cooperation	• The role of global alliances in standardizing AI security measures.
Information Sharing Frameworks	• Establishing protocols for sharing threat intelligence.
Cross-Sector Collaboration	• Leveraging expertise across various industries for a unified defense.
Initiatives	• Cybersecurity Exercise/Simulation training

FUTURE CHALLENGES I [5][9]

- Increasing Complexity of AI Systems and Outpacing Cybersecurity Measures
- Quantum Computing impact
- Anticipatory Security Measures
- Cyber-Physical Attacks



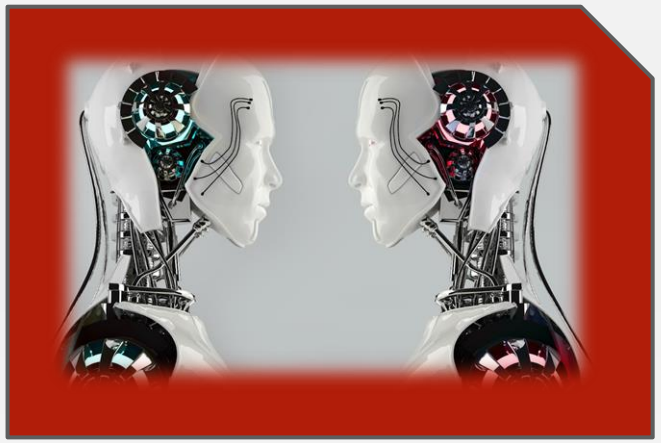
[Img13]

[Img14]

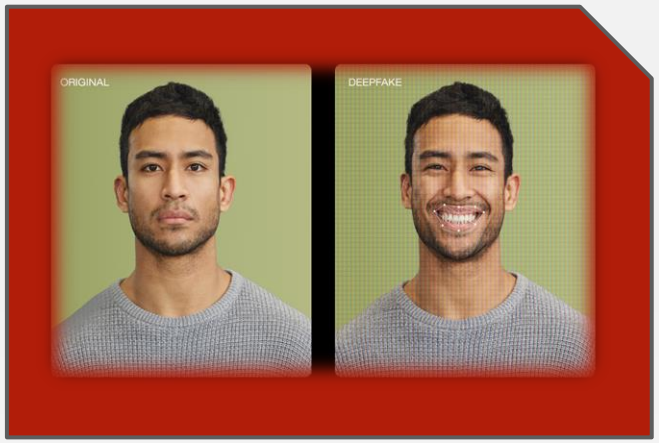
[Img15]

FUTURE CHALLENGES II[5][9]

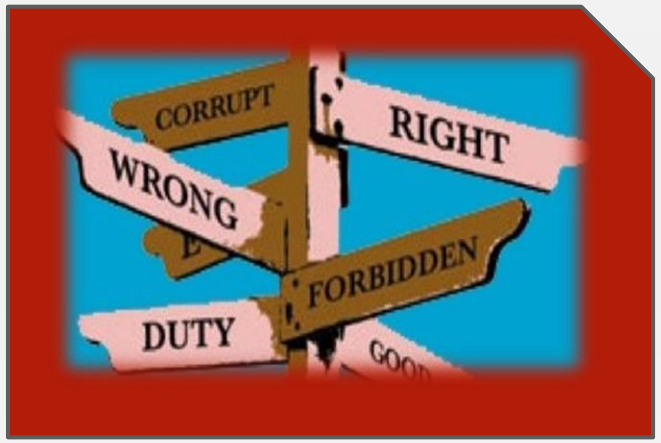
- AI vs. AI Scenarios
- Synthetic Media as a Vector for Disinformation (example: Deepfakes…)
- Ethical/Moral Considerations
- Legal/Regulatory/Privacy Consequences



[Img16]



[Img17]



[Img16]

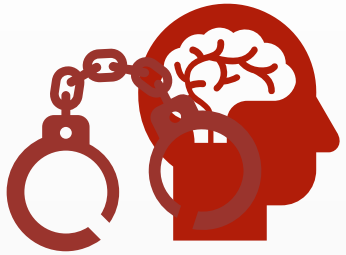
Wroclaw University of Science and Technology

CONCLUSIONS

- Discussed – AI systems as a potential target of malicious attacks and potential threat, source of attack. Intelligent Crimes and common culprits.
- Talked – Early Exploits, More sophisticated attacks and Recent Trends.
- Presented – Classification of methods and vulnerabilites
- Go – into details of some attacks.
- Explore – Real life Cases.
- Analyze – Large-Scale Blackmail
- Get to know – defensive measures.
- Highlighted – challanges ahead.

Wrocław University of Science and Technology

REFERENCES



LITERATURE/ARTICLES:

- [1] - Russell, Stuart J., and Peter Norvig. Artificial intelligence a modern approach. London, 2010.
- [3] - Wang, P. (2019). On defining artificial intelligence. Journal of Artificial General Intelligence, 10,
- [5] - Caldwell, M., Andrews, J.T.A., Tanay, T. et al. AI-enabled future crime. Crime Sci, 2020.
- [7] – Eggers, Shannon Leigh, 2020 . Vulnerabilities in Artificial Intelligence and Machine Learning Applications and Data. United States.
- [8] - NIST (2023). AI 100-2 E2023, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. National Institute of Standards and Technology.
- [9] - Pupillo, L., Fantin, S., Ferreira, A., & Polito, C. (2021). Artificial Intelligence and Cybersecurity: Technology, Governance and Policy Challenges. Centre for European Policy Studies (CEPS).
  https://eda.europa.eu/docs/default-source/documents/ceps-tfr-artificial-intelligence-and-cybersecurity.pdf (date: 14.11.2023)
- [11] - Blessing Guembe, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz & Vera Pospelova (2022) The Emerging Threat of Ai-driven Cyber Attacks: A Review, Applied Artificial Intelligence
- [16] - European Economic and Social Committee, Ensuring awareness and resilience of the private sector across Europe in face of mounting cyber risks, 2018

BLOG/WEBSITE SOURCES:

- [2] - https://www.smartinsights.com/managing-digital-marketing/marketing-innovation/artificial-intelligence-adoption-different-sectors/ (date: 15.11.2023)
- [4] - https://surfshark.com/research/chart/statistics-of-ai-incidents (date: 14.11.2023)
- [6] - https://atlas.mitre.org/studies/ (date: 11.11.2023)
- [10] - https://www.belfercenter.org/publication/AttackingAI (date: 14.11.2023)
- [12] - https://www.infoq.com/articles/ai-cyber-attacks/ (date: 15.11.2023)
- [13]- https://proprivacy.com/privacy-news/deepfake-technology-used-in-hong-kong-bank-heist (date: 15.11.2023)
- [14] - https://www.linkedin.com/pulse/new-vulnerability-popular-wordpress-plugin-exposes-over-2-million (date: 15.11.2023)
- [15] - https://www.thalesgroup.com/en/worldwide/security/press_release/french-mod-challenge-thales-performs-successful-sovereign-a[illegible]hack([illegible].11.202[illegible]
- [17] - https://www2.deloitte.com/us/en/insights/focus/tech-trends/2022/future-of-cybersecurity-and-ai.html (date: 14.11.2023)

OTHER SOURCES

IMAGE SOURCES:

- [Img1] – Image genere via DALL-E https://chat.openai.com/ (date: 15.11.2023)
- [Img2] – https://www.smartinsights.com/managing-digital-marketing/marketing-innovation/artificial-intelligence-adoption-different-sectors/ (date: 15.11.2023)
- [Img3] - Data collected on June 16th, 2023 Data source: Aincidentdatabase.ai https://surfshark.com/research/chart/statistics-of-ai-incidents (date: 14.11.2023)
- [Img4] - https://www.money.pl/ (date: 16.11.2023)
- [Img5] - https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (date: 16.11.2023)
- [Img6] - https://translate.google.pl/?hl=pl (date: 16.11.2023)
- [Img7] - https://www.id.me/ (date: 16.11.2023)
- [Img8] – https://idhosting.pl/ (date: 15.11.2023)
- [Img9] – https://www.cpomagazine.com/cyber-security/instagram-bug-allows-account-takeover-attacks-turns-mobile-devices-into-spying-tools/ (date: 15.11.2023)
- [Img10] – https://www.taskrabbit.com/ (date: 15.11.2023)
- [Img11] - https://www.tripwire.com/state-of-security/deepfake-voice-technology-phishing-strategies (date: 15.11.2023)
- [Img12] - https://hashstudioz.com/blog/benefits-of-using-artificial-intelligence-in-cyber-security/ (date: 15.11.2023)
- [Img13] - https://bmc.com/blogs/neural-network-introduction/ (date: 14.11.2023)
- [Img14] - https://www.pap.pl/aktualnosci/news%2C1459236%2Cwywiad-wojskowy-ukrainy-rosyjskie-wojska-zaminowaly-kachowska-elektrownie (date: 14.11.2023)
- [Img15] - https://www.science.org/content/article/quantum-computers-take-key-step-toward-curbing-errors (date: 14.11.2023)
- [Img16] - https://www.freepik.com/ (date: 14.11.2023)
- [Img17] - https://vimeo.com/blog/post/video-deepfakes/ (date: 14.11.2023)

TEMPLATE:

- https://pwr.edu.pl/ (date: 08.11.2023)

Wroclaw University of Science and Technology

Q&A