

# CyberSecurity Technologies & Governance



**ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ**

ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS

## “How AI can enhance and evade Cybersecurity”

- A brief overview focusing on how AI technologies can restrict and amplify phishing attacks

---

KATERINA BASMPA, F3312408

PANAGIOTIS MELAS, F3312407

NIKOLETTA TERZIDOU, F3312412

# AI and Cybersecurity

**AI introduces new potential threat vectors and new ways to mitigate them.**

---

The technology of AI:

- Enhances defenders' capacity to move faster with greater efficiency and confidence. [1]

- Lowers the bar even further for low-skill threat actors helping them develop more sophisticated exploits.[1]

## **Defenders advantage of AI :**

- Continuous regulatory compliance,
- Case Management,
- Accelerated Thread Hunting,
- Incident Simulation and Pen Testing
- Data Interpretation (Collates telemetry data across sources, and speeds analysts' understanding of security log data)

## **Threat actors benefit from AI :**

- Social Engineering and fraud
- Data theft
- Identify Data Theft and Impersonation
- AI Jailbreaks (removes the guardrails on the gen AI chatbots)
- Password Cracking.

# Focusing on the Social Engineering field

Phishing, the most common one of online threats and cyber attacks.

---

Further below we explore

## **AI Phishing URLs Detection Systems :**

- Different machine learning and deep learning-based approaches have been proposed for designing defensive mechanisms against various phishing attacks.

## **Advanced Phishing Attack Design and Deployment Using generative AI technology**

- Phishing attacks are sophisticated and multifaceted, with phishers taking advantage of many techniques.
- There is an artificial intelligence application which significantly increases the potential risk associated with using it for such illicit activities and expanding the reach and magnitude of phishing attacks.

# a) How AI can reinforce Phishing Detection Systems

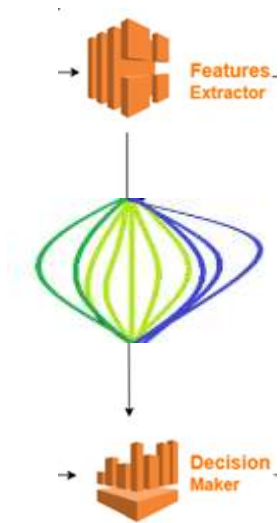
---

## Phishing Detection Systems Before the incorporation of AI technologies

- Most of them rely on signature – based methods. They work by recognising known indicators of phishing URLs. However, they cannot predict and identify attacks that consist of novel features.
- For instance, URL Void detects phishing URLs, by using several engines and blacklists of domains such as Google Safe Browsing, yet it relies heavily on databases of confirmed malicious URLs.
- Comodoro Site Inspector is a tool that allows users to check URLs with a sandbox environment. It creates an environment isolated from the operating system, where the user can check the URL themselves, but this technique relies on the user's suspicion.[9]

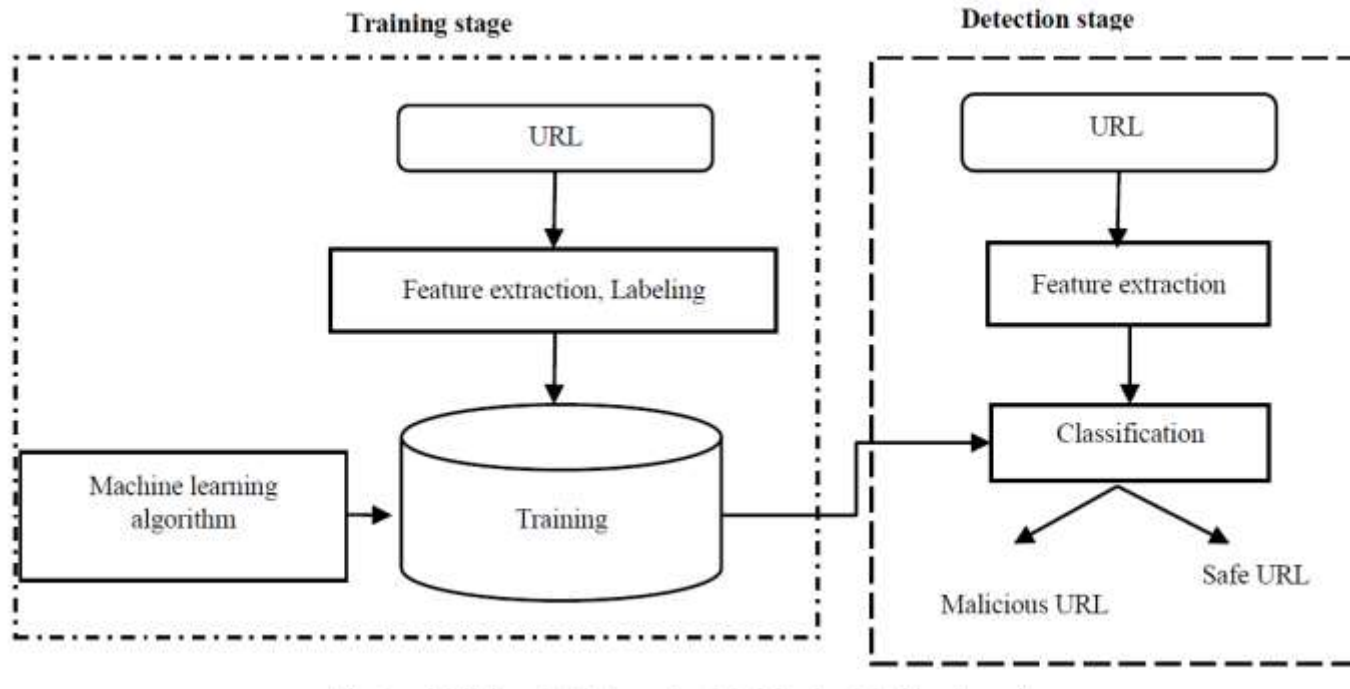
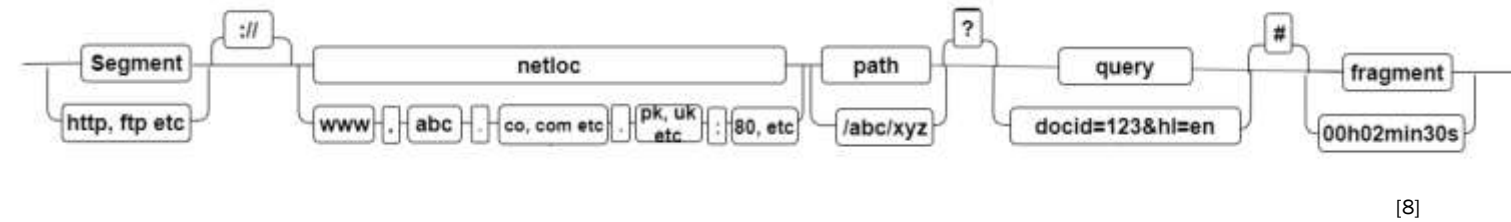
Sameen , Han and Hwang propose [8] PhishHaven – An efficient Real-Time AI Phishing URLs Detection system in response to DeepPhish.

- DeepPhish is an AI-based tool that successfully raises success rates of URL-Phishing attacks created by humans. PhishHaven aims to detect both human and AI-generated phishing URLs[14].
- Using the general model, for the training, it employs 10 different Machine Learning algorithms in a multi-threaded technique. The threads train concurrently and independently from one another.
- Each thread classifies the URL as safe or unsafe and then feeds it to a Decision Maker. The Decision Maker chooses the aggregate result based on number of votes for each side.
- The results showed an average of 98,04% Precision and 98.04% Accuracy in detecting AI-generated and simple Phishing URLs.



Boosting Based	Voting-based decision to make weak learners strong	1. Adaboost 2. Gradient Boosting
Non-Learning Based	Optimization and drawing conclusions	3. Decision Trees 4. Random Forest 5. Extra Tree Classifier 6. Bagging Classifier 7. K-Nearest Neighbour
Learning Based	optimization and drawing- boundaries	8. Logistic Regression 9. Support Vector Machines 10. Neural Networks

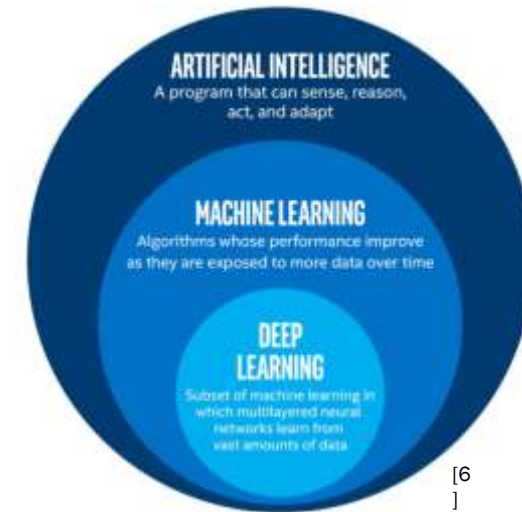
# Contribution of AI Technologies



## A GENERAL MODEL

- Datasets of phishing URLs and safe URLs are collected as input.
- URLs are tokenized and lexical features that indicate markers such as number of subdomain levels or existence of brand name in the domain are collected. Host-based analysis extracts features such as malicious server identity.
- These data are fed into an AI machine where classifiers are trained to decide whether a URL is malicious or safe.
- The machine predicts if a given URL is safe or malicious.[9][8]

- Traditional machine learning models, however effective, require that the features of URLs be extracted manually, a time- consuming process whenever new phishing URLs are introduced. Thus, recent publications focus on Deep Learning Models[4][5][7].
- Alshingiti, Alaqel, Al-Muhtadi, Haq ,Saleem and Faheem test three different Deep Learning- Based algorithms for a proposed Phishing Detection Machine[4]:
- Long short-term Memory (LSTM), Convolutional Neural Network (CNN) and LSTM- CNN. Mimicking the human brain, LSTMs process complete sequence data and eliminate useless information and CNNs assign importance to and differentiate data with weights and biases.
- While all machines are effective, CNNs prove to be the fastest and most effective with accuracy 99.2%. Furthermore, a comparison with other models which use DL Methods like CNN,CAE and DNN, CNN, LSTM, and GRU proves this model as the most successful one.



## Feature of AI in Phishing Detection

- AI- based models are very efficient in detecting phishing attacks and they prevent zero-day attacks. Improving the URL data processing methods and training on larger Databases will assist the efficiency of the machines. Since DL machines require a lot of processing power, advancements in processing will further assist these machines. Finally, the tools need to be re-evaluated as attacks improve.[7][8][5]

## b) THE USE OF CHATGPT FOR PHISING CYBERATTACKS REINFORCEMENT

---

- When thinking of artificial intelligence, ChatGPT often comes to mind. This powerful AI tool, based on machine learning, not only provides knowledge but also generates it. It can filter information, correct input, create code and scripts, and assist with creative ideas.
- Used daily by millions for professional, personal, and operational tasks, ChatGPT is now embedded in our digital lives. However, cybercriminals could exploit its capabilities to generate malware, find vulnerabilities, automate attacks, and conduct scams. This raises concerns about the potential for ChatGPT to enable or create cyber threats.
- As a result, there may be a need for cybersecurity experts to regulate AI and machine learning. The rapid pace of technological evolution demands cybersecurity measures that can proactively prevent threats before they arise.



[10]




- ChatGPT, developed by OpenAI, is a chatbot based on the GPT-3.5 model, trained with Reinforcement Learning from Human Feedback (RLHF), and functions as a virtual assistant offering accurate, high-level information [10][11].
- Its capabilities enable the design of sophisticated phishing attacks, such as mimicking websites, acquiring domains, storing credentials, and distributing fraudulent URLs via spam emails, making phishing efforts more effective [12].
- ChatGPT can generate source code in various languages (HTML, CSS, JS, PHP) to create fully functional phishing websites that mimic trusted brands and use evasive tactics to bypass anti-phishing tools [13].



[15]

## **The Second IEEE Conference on Communications and Network Security 2023 - Cyber Resilience Workshop [12], presents how the attributes described above can be utilized;**

- OpenAI integrates advanced content filtering and safeguards to prevent exploitation, while attackers use jailbreaks, such as the "Do Anything Now" (DAN) prompt, and modular prompt injection to bypass security filters and stealthily assemble phishing kits.
- ChatGPT can generate a Python class using subprocess module to invoke HTTrack, cloning the target site and including an unintended auto-launching web server. It flags potential illegality but produces functional, correctable code.

- 
- Python code uses OpenAI API to optimize site copy, modify login forms to link to an API, and add phishing modals, improving stealth and performance.
  - ChatGPT, unable to directly obfuscate code, was asked about techniques and sample code. Segmenting the source code eliminated model constraints, enabling broader use of obfuscation methods, including character encoding.
  - A Flask API in Python collects victim credentials from a phishing site and sends them directly to Telegram via HTTP requests. It uses BotFather for a token and RawDataBot for a chatID.

- a functional local website was automatically deployed on a cloud instance using Bash scripts and the cloud provider's Python library. ChatGPT created a Bash script for Python 3.11 installation and a Python script using Paramiko for SSH connections and file transfers.
- A domain was associated with a registrar's API and Python library after deploying the phishing kit. ChatGPT generated a registration class and random registrant details, and added a method for automatic name server updates.
- Integrating an online reverse proxy offers benefits like AntiBot services, a valid TLS certificate, and browser security. The provider's Python library and API enable configuration, with ChatGPT generating code for automating domain addition.
- The IEEE conference informs us that ChatGPT can enable automated phishing kits even for those with limited programming skills, but key challenges include token limits and susceptibility to misuse. Future models with larger capacities or improved segmentation could address these limitations.

# In conclusion

---

- ✓ Phishing assaults are becoming more common and complex and artificial intelligence (AI) has become an essential tool for detecting and blocking them.
- ✓ Cybersecurity solutions powered by AI play an important role in enhancing threat detection and prevention.
- ✓ AI in cybersecurity is a huge step forward in the fight against cyber threats that are getting smarter all the time.[3]

## **BUT some aspects of AI could be**

- Not resilient to malicious usage despite extended safeguards and filters.
- Adversaries can leverage them to generate and deploy phishing websites swiftly.
- Significantly increasing the potential risk associated with AI for such illicit activities.
- Expanding the reach and magnitude of phishing attacks.[12]

# General concerns about AI's role in cybersecurity and beyond

## Current trends in cyber attacks[2]

- ❖ The biggest shift observed in 2023 was a pronounced surge in cyberthreats targeting identities.
- ❖ Attackers have a historical inclination to choose the path of least resistance in pursuit of their objectives.
- ❖ In this era, the focus has shifted towards **logging in** rather than **hacking in**.
- ❖ Highlighting the relative ease of acquiring credentials compared to exploiting vulnerabilities or executing phishing campaigns.

## Current and future trends in market[2]

- ❑ While more organizations say they are developing AI models, and AI is being used in different solutions, the AI market is currently in a pre-mass market period.
- ❑ Once AI market dominance is established when a single technology approaches 50% market share or the market consolidates to three or less technologies— researchers assess it will trigger the maturity of AI as an attack surface.
- ❑ The result will be that cybercriminals will then further mobilize and increase their investment in attacking AI.

## Our concern

Even though AI is undoubtably a powerful technology with countless usages, the generative abilities that it provides, raise a warning of whether it is capable of surpassing human innovations in the field of cybersecurity and society in general, shaking humanity's balance.

# REFERENCES

- [1] IBM Institute for Business Value | Research Insights (May 2024), Securing Generative AI - What matters now, Chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://d1.awsstatic.com/executive-insights/en\_US/Securing%20Generative%20AI.pdf
- [2] IBM, X-Force Threat Intelligence Index 2024 (February 2024), chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.ibm.com/downloads/documents/us-en/107a02e952c8fe80
- [3] Nicolas Guzman Camacho (March 2023), The Role of AI in Cybersecurity: Addressing Threats in the Digital Age, ISSN: 3006-4023 (Online), Vol. 3, Issue 1 Journal of Artificial Intelligence General Science (JAIGS), <https://ojs.boulibrary.com/index.php/JAIGS/article/view/75/46>
- [4] Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q.E.U., Saleem, K. and Faheem, M.H. (2023). A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN. *Electronics*, 12(1), p.232. doi:<https://doi.org/10.3390/electronics12010232>.
- [5] Asiri, S., Xiao, Y., Alzahrani, S., Li, S. and Li, T. (2023). A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks. *IEEE Access*, 3237798(3237798), pp.1–1. doi:<https://doi.org/10.1109/access.2023.3237798>.
- [6] Foundations of AI & ML. (2018). *What is DL*. [online] Available at: <https://mylearningsinaiml.wordpress.com/what-is-dl/> [Accessed 8 Dec. 2024].
- [7] Khonji, M., Iraqi, Y. and Jones, A. (2023). Phishing Detection: a Literature Survey. *IEEE Communications Surveys & Tutorials*, [online] 15(4), pp.2091–2121. doi:<https://doi.org/10.1109/surv.2013.032213.00009>.
- [8] Sameen, M., Han, K. and Hwang, S.O. (2020). PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System. *IEEE Access*, 8(2991403), pp.83425–83443. doi:<https://doi.org/10.1109/access.2020.2991403>.
- [9] Xuan, C.D., Dinh, H. and Victor, T. (2020). Malicious URL Detection based on Machine Learning. *International Journal of Advanced Computer Science and Applications*, 11(1). doi:<https://doi.org/10.14569/ijacsa.2020.0110119>.
- [10] OpenAI, “Introducing ChatGPT,” <https://openai.com/blog/chatgpt>.
- [11] OpenAI, “GPT-3.5,” <https://platform.openai.com/docs/models/gpt-3-5>.
- [12] Nils Begou, J’er’emy Vinoy, Andrzej Duda, Maciej Korczyński, “Exploring the Dark Side of AI: Advanced Phishing Attack Design and Deployment Using ChatGPT”, Second IEEE Conference on Communications and Network Security 2023 - Cyber Resilience Workshop.
- [13] Sayak Saha Roy, Krishna Vamsi Naragam, Shirin Nilizadeh, ” Generating Phishing Attacks using ChatGPT”, The University of Texas at Arlington.
- [14] Bahnsen, A.C., Torroledo, I., Camacho, L.D. and Villegas, S. (2018). *DeepPhish : Simulating Malicious AI*. [online] [www.semanticscholar.org](https://www.semanticscholar.org/paper/DeepPhish-%3A-Simulating-Malicious-AI-Bahnsen-Torroledo/ae99765d48ab80fe3e221f2eedec719af80b93f9). Available at: <https://www.semanticscholar.org/paper/DeepPhish-%3A-Simulating-Malicious-AI-Bahnsen-Torroledo/ae99765d48ab80fe3e221f2eedec719af80b93f9>.
- [15] *Telecom Review Africa*, “ChatGPT and the Future of Cybersecurity: An AI Perspective.”, Available at : [www.telecomreviewafrica.com/articles/features/3283-chatgpt-and-the-future-of-cybersecurity-an-ai-perspective/](http://www.telecomreviewafrica.com/articles/features/3283-chatgpt-and-the-future-of-cybersecurity-an-ai-perspective/). Accessed 9 Dec. 2024.