

Informatyka temat nr 3	Computer and information sciences topic No. 3
Odporność modeli uczenia maszynowego na ataki adwersarialne	Robustness of machine learning models to adversarial attacks
<p>PhD supervisor: dr hab. Agnieszka Nowak-Brzezińska, prof. UŚ https://orcid.org/0000-0001-7238-1170</p>	
<p>Krótką charakterystyką założeń i celów badawczych</p> <p>Głównym celem projektu jest opracowanie nowatorskich algorytmów uczenia maszynowego nienadzorowanego (klasteryzacji oraz wykrywania anomalii), które posiadają matematycznie dowiedzioną odporność na celowe manipulacje rozkładem danych (adversarial distribution shifts). Badania koncentrują się na stworzeniu modeli zdolnych do stabilnego działania w warunkach aktywnego przeciwdziałania ze strony inteligentnych adwersarzy, uniemożliwiając im skuteczne przeprowadzenie ataków typu poisoning (zatrucie danych) oraz evasion (omijanie detekcji).</p> <p>W trakcie badań zrealizowane będą następujące założenia projektowe:</p> <ul style="list-style-type: none"> - Paradygmat Secure-by-Design: Systemy detekcji nie mogą być jedynie reaktywne; muszą być projektowane z założeniem istnienia inteligentnego, adaptacyjnego przeciwnika. - Stabilność matematyczna: Wynik klasteryzacji lub detekcji musi pozostać stabilny nawet w obliczu wprowadzenia przez napastnika celowych, subtelnych zmian w danych wejściowych. - Rozróżnialność dryftu: Algorytmy muszą precyzyjnie odróżniać naturalny dryft danych (wynikający np. ze zmian rynkowych czy technologicznych) od zmanipulowanego przesunięcia rozkładu generowanego przez autonomiczne agenty. 	<p>Brief description of research assumptions and goals</p> <p>The primary objective of this project is to develop innovative unsupervised machine learning algorithms (for clustering and anomaly detection) that possess mathematically proven robustness against intentional data distribution manipulations (adversarial distribution shifts). The research focuses on creating models capable of stable operation in environments characterized by active interference from intelligent adversaries, thereby preventing the effective execution of poisoning and evasion attacks.</p> <p>To achieve this, the following design assumptions will be implemented throughout the research:</p> <ul style="list-style-type: none"> - Secure-by-Design Paradigm: Detection systems must not be merely reactive; they must be designed with the fundamental assumption of an intelligent, adaptive opponent. - Mathematical Stability: The output of clustering or detection must remain stable even when faced with intentional, subtle modifications to the input data introduced by an attacker. - Drift Discernibility: Algorithms must precisely distinguish between natural data drift (resulting from market or technological shifts) and adversarial distribution shifts generated by autonomous AI agents.



Planowany wkład w rozwój dyscypliny

Proponowany projekt badawczy wnosi istotny wkład w rozwój dziedziny uczenia maszynowego oraz cyberbezpieczeństwa systemów autonomicznych w następujących obszarach:

- Rozszerzenie teorii odpornego uczenia nienadzorowanego (Robust Unsupervised Learning): Opracowanie nowych modeli klasteryzacji i detekcji anomalii, które posiadają wbudowane mechanizmy odporności na ataki adversarialne. Wkładem jest przejście od modeli statycznych do modeli uwzględniających dynamikę inteligentnego przeciwnika (adversarial-aware models).

- Matematyczna formalizacja "Adversarial Distribution Shifts": Zdefiniowanie i sformalizowanie różnic między naturalnym dryftem danych a celową manipulacją rozkładem dokonywaną przez agenty AI. Pozwoli to na stworzenie nowych metryk oceny stabilności modeli nienadzorowanych, które dotychczas były traktowane marginalnie w porównaniu do modeli klasyfikacyjnych (LLM).

- Opracowanie metod detekcji "Shadow Agents": Wprowadzenie nowej metodologii wykrywania aktywności autonomicznych agentów i rojów botów (AI Predator Swarms), które operują poniżej progu detekcji tradycyjnych systemów Outlier Detection.

- Wzmocnienie paradygmatu Secure-by-Design w systemach Agentic SOC: Dostarczenie fundamentów teoretycznych i algorytmicznych dla nowej generacji centrów operacji bezpieczeństwa (SOC), w których agenci AI muszą ufać integralności danych wejściowych, aby autonomicznie neutralizować zagrożenia.

- Integracja metodyk MITRE ATLAS i OWASP z algorytmami ML: Przełożenie wysokopoziomowych ram bezpieczeństwa na konkretne parametry matematyczne modeli, co wypełnia lukę między teoretycznym opisem zagrożeń (np. Rogue AI) a techniczną implementacją mechanizmów obronnych.

Standardowe podejście do ML (zakładające, że dane pochodzą z rozkładu stacjonarnego) staje się niewystarczające. Wkład realizowanego tematu badań będzie polegał na redefinicji bezpieczeństwa

Planned contribution to the development of the discipline

The proposed research project makes a significant contribution to the fields of Machine Learning and Autonomous Systems Cybersecurity in the following areas:

- Advancing the Theory of Robust Unsupervised Learning: Developing novel clustering and anomaly detection models equipped with built-in mechanisms for resistance against adversarial attacks. This contribution marks a shift from static models to adversarial-aware models that account for the dynamics of an intelligent opponent.

- Mathematical Formalization of "Adversarial Distribution Shifts": Defining and formalizing the distinctions between natural data drift and intentional distribution manipulation orchestrated by AI agents. This will enable the creation of new metrics for assessing the stability of unsupervised models, which have been marginalized to date compared to generative and classification models (LLMs).

- Development of Detection Methods for "Shadow Agents": Introducing a new methodology for identifying the activities of autonomous agents and AI Predator Swarms that operate below the detection thresholds of traditional Outlier Detection systems.

- Strengthening the Secure-by-Design Paradigm in Agentic SOC Systems: Providing the theoretical and algorithmic foundations for a new generation of Security Operations Centers (SOCs), where AI agents must rely on the integrity of input data to autonomously neutralize threats.

- Integration of MITRE ATLAS and OWASP Frameworks with ML Algorithms: Translating high-level security frameworks into specific mathematical model parameters, thereby bridging the gap between theoretical threat descriptions (e.g., Rogue AI) and the technical implementation of defense mechanisms.

The standard approach to ML, which assumes data originates from a stationary distribution, is no longer sufficient. The contribution of this research lies in redefining model security as an integral part of its architecture rather than merely an external layer of protection.





<p>modelu jako integralnej części jego architektury, a nie tylko zewnętrznej warstwy ochrony.</p>	
<p>Opis wymagań – wiedza, umiejętności i kompetencje społeczne kandydata</p> <p>Wiedza:</p> <ul style="list-style-type: none">- Zaawansowana statystyka i matematyka: Głęboka wiedza z zakresu teorii prawdopodobieństwa, algebry liniowej oraz metod optymalizacji (w szczególności optymalizacji odpornej).- Teoria uczenia maszynowego: Znajomość paradygmatów uczenia nienadzorowanego (clustering, density estimation) oraz mechanizmów uczenia głębokiego (Autoenkodery, GANs).- Cyberbezpieczeństwo AI: Znajomość taksonomii ataków adversarialnych (według MITRE ATLAS), rozumienie mechanizmów poisoning, evasion oraz model inversion.- Agentic AI & LLMs: Podstawowa wiedza na temat funkcjonowania agentów autonomicznych i ryzyka związanego z Rogue AI (zgodnie z wytycznymi OWASP 2026). <p>Umiejętności / Kompetencje:</p> <ul style="list-style-type: none">- Programowanie i Frameworki: Biegłość w języku Python oraz bibliotekach dedykowanych AI (PyTorch/TensorFlow).- Modelowanie Matematyczne: Zdolność do formalizacji problemów bezpieczeństwa w języku matematycznym i dowodzenia stabilności algorytmów.- Analiza Danych: Doświadczenie w pracy z dużymi zbiorami danych, w tym umiejętność wykrywania anomalii i analizy dryftu danych (data drift analysis).- Język Angielski: Biegłość na poziomie co najmniej B2, umożliwiającą analizę najnowszych publikacji na arXiv oraz prezentowanie wyników na międzynarodowych konferencjach. <p>Kompetencje społeczne:</p> <ul style="list-style-type: none">- Myślenie analityczne i krytyczne.- Etyka zawodowa.- Samodzielność i inicjatywa: Umiejętność samodzielnego planowania eksperymentów badawczych i krytycznej oceny uzyskanych wyników.- Komunikacja naukowa: Zdolność do przekazywania skomplikowanych zagadnień	<p>Description of requirements – knowledge, skills and social competences of the candidate</p> <p>Knowledge:</p> <ul style="list-style-type: none">- Advanced Statistics and Mathematics: In-depth knowledge of probability theory, linear algebra, and optimization methods (specifically robust optimization).- Machine Learning Theory: Proficiency in unsupervised learning paradigms (clustering, density estimation) and deep learning mechanisms (Autoencoders, GANs).- AI Cybersecurity: Familiarity with adversarial attack taxonomies (e.g., MITRE ATLAS) and a thorough understanding of poisoning, evasion, and model inversion mechanisms.- Agentic AI & LLMs: Foundational knowledge of autonomous agent functionality and the risks associated with Rogue AI (in accordance with OWASP 2026 guidelines). <p>Skills and Competencies:</p> <ul style="list-style-type: none">- Programming and Frameworks: Proficiency in Python and dedicated AI libraries (PyTorch/TensorFlow).- Mathematical Modeling: The ability to formalize security problems in mathematical terms and provide proofs for algorithmic stability.- Data Analysis: Experience working with large datasets, including proficiency in anomaly detection and data drift analysis.- English Language Proficiency: Minimum B2 level, enabling the analysis of the latest publications on arXiv and the presentation of research results at international conferences. <p>Social Competencies:</p> <ul style="list-style-type: none">- Analytical and Critical Thinking: A strong capacity for logical reasoning and problem-solving.- Professional Ethics: A commitment to integrity and the development of responsible AI.- Autonomy and Initiative: The ability to independently plan research experiments and critically evaluate the obtained results.- Scientific Communication: The capacity to convey complex technical issues clearly, both in writing (scientific papers) and orally. <p>Additional Assets (optional):</p>





technicznych w sposób zrozumiały, zarówno w formie pisemnej (artykuły naukowe), jak i ustnej.

Dodatkowe atuty (opcjonalnie):

- Udział w projektach skupionych na AI.
- Publikacje naukowe lub czynny udział w kołach naukowych związanych z cyberbezpieczeństwem/konferencjach AI.

- Participation in AI-focused projects.

- Scientific publications or active involvement in student research groups related to cybersecurity or AI conferences.

