dr inż. Karol Chlasta
Akademia Leona Koźmińskiego
ORCID: 0000-0002-6539-566X

https://doi.org/10.29316/9788368103205_3

SZTUCZNA INTELIGENCJA W CYBERBEZPIECZEŃSTWIE – WYZWANIA I MOŻLIWOŚCI

ARTIFICIAL INTELLIGENCE IN CYBERSECURITY – CHALLENGES AND OPPORTUNITIES

Streszczenie

W dobie rosnącej roli sztucznej inteligencji (AI) w cyberbezpieczeństwie, generatywna sztuczna inteligencja (GenAI) staje się zarówno narzędziem obronnym, jak i potencjalnym zagrożeniem. Niniejsza praca przedstawia przegląd metod sztucznej inteligencji, w tym generatywnej sztucznej inteligencji, w kontekście cyberbezpieczeństwa, uwzględniając zarówno nowe techniki ataków, jak i rozwiązania obronne, wspierane przez automatyczne modelowanie tematów. Rozdział zawiera wyniki przeglądu literatury naukowej dotyczącej cyberbezpieczeństwa i sztucznej inteligencji w latach 2017-2024, obejmującego 2420 publikacji z IEEE Xplore Digital Library i SpringerLink. Publikacje koncentrują się głównie na następujących tematach: (1) phishing (407 publikacji; 16,82%), (2) techniki adversarial

Summary

In an era marked by the increasing influence of artificial intelligence (AI) in cybersecurity, generative AI (GenAI) has emerged as both a formidable defensive tool and a potential threat. This chapter examines new AI methods, with a particular focus on generative AI, within the context of cybersecurity, addressing both emerging attack strategies and defensive solutions, as informed by automated topic modelling. The chapter presents a comprehensive review of scientific literature on cybersecurity and AI, spanning the years 2017 to 2024, drawing from a total of 2,420 publications sourced from the IEEE Xplore Digital Library and SpringerLink. The publications primarily address the following topics: (1) phishing (407 publications; 16.82%), (2) adversarial techniques (347 publications; 14.34%), (3) ChatGPT

(347 publikacji; 14,34%), (3) ChatGPT (352 publikacje; 14,55%), (4) duże modele językowe (LLM) (352 publikacje; 14,55%) oraz podatności (77 publikacji; 3,18%). Zidentyfikowane zagrożenia związane z wykorzystaniem LLM w generowaniu phishingu, automatycznym hakowaniu oraz tworzeniu złośliwego oprogramowania zostały szczegółowo podsumowane. Ponadto, zaobserwowano wzrost znaczenia narzędzi opartych na GenAI w obszarze detekcji zagrożeń, analizy kodu oraz automatyzacji cyberobrony. Zdecydowanie widać, że tematy związane z szeroko rozumianą generatywną sztuczną inteligencją (LLM i ChatGPT) stanowią większy zbiór (704 publikacje) niż tradycyjne metody uczenia maszynowego, w tym zagrożenia związane z manipulowaniem danymi wejściowymi do systemów SI (tematyka adversarial obejmuje tylko 347 publikacji). Rozdział podsumowuje również zidentyfikowane regulacje prawne oraz standardy dotyczące cyberbezpieczeństwa, takie jak NIS2, NIST AI Risk Management Framework oraz standardy ISO/IEC 22989, 23053, 23984 i 42001. Praca ukazuje rosnącą i istotną rolę różnych metod generatywnej sztucznej inteligencji w kształtowaniu przyszłości cyberbezpieczeństwa, podkreślając konieczność opracowania nowych strategii ochrony przed nowymi zagrożeniami w zmieniającym się środowisku organizacyjnym, w którym coraz bardziej polegamy na publicznej infrastrukturze chmurowej.

Słowa kluczowe: Generatywna Sztuczna Inteligencja (AI), Duże Modele Językowe (LLM), Zagrożenia oparte na AI, Cyberbezpieczeństwo, Symulacja ataków, Antagonistyczne Uczenie Maszynowe, zarządzanie ryzykiem cyberbezpieczeństwa (NIS2, NIST AI RMF), Normy ISO/IEC bezpieczeństwa AI

(352 publications; 14.55%), (4) large language models (LLMs) (352 publications; 14.55%), and vulnerabilities (77 publications; 3.18%). The risks associated with the use of LLMs in phishing campaigns, automated hacking, and malware generation are systematically summarised. Furthermore, the growing prominence of GenAI-based tools in threat detection, code analysis, and the automation of cyber defence is highlighted. Notably, the body of work dedicated to GenAI, comprising of LLMs and ChatGPT, outnumbers the traditional machine learning studies, with 704 papers on the former compared to just 347 on adversarial topics. The chapter also explores key cybersecurity regulations and standards identified in the literature, including NIS2, the NIST AI Risk Management Framework, and a range of ISO/IEC standards (22989, 23053, 23984, and 42001). This review underscores the increasing and pivotal role of various generative AI methods in shaping the future landscape of cybersecurity, while emphasising the urgent need for new strategies to defend against evolving threats in an environment increasingly dependent on public cloud infrastructure.

Keywords: Generative AI, Large Language Models (LLMs), AI-Driven Threats, Cybersecurity, Offensive Security, Adversarial Methods, Cybersecurity Risk Management (NIS2, NIST AI RMF), ISO/IEC Standards for AI Security

Wstęp

Bezpieczeństwo w cyberprzestrzeni stanowi jedno z kluczowych wyzwań współczesnego świata, szczególnie w kontekście dynamicznego rozwoju technologii informacyjno-komunikacyjnych oraz narastającej skali i złożoności zagrożeń cyfrowych. W dobie transformacji cyfrowej rośnie znaczenie nowoczesnych i zintegrowanych mechanizmów ochrony danych oraz infrastruktury teleinformatycznej, przy jednoczesnym uwzględnieniu wymogów regulacyjnych i międzynarodowych standardów bezpieczeństwa, takich jak ISO/IEC 27001 czy NIST Cybersecurity Framework[1].

Jednym z istotnych czynników wpływających na kształtowanie współczesnego ekosystemu cyberbezpieczeństwa jest rozwój sztucznej inteligencji (AI), która z jednej strony stanowi skuteczne narzędzie wspierające ochronę systemów informatycznych, a z drugiej – może być wykorzystywana jako wektor ataku przez podmioty złośliwe. Szczególne zainteresowanie nauki i przemysłu wzbudzają obecnie modele generatywne, w tym architektury oparte na Generative Adversarial Networks (GAN)[2] oraz bazujące na architekturze Transformer[3]. Wspomniane modele „przeciwnikowe" typu GAN działają na zasadzie „gry" pomiędzy dwiema sieciami neuronowymi[4]: generatorem oraz dyskryminatorem. Generator tworzy dane imitujące zbiór treningowy, natomiast dyskryminator uczy się rozróżniać dane syntetyczne od rzeczywistych. Z kolei architektura Transformer, oparta na mechanizmie samo-uwagi (self-attention), umożliwia równoległe przetwarzanie sekwencji danych, co czyni ją wyjątkowo efektywną w zadaniach, takich jak analiza języka naturalnego (NLP), detekcja anomalii czy klasyfikacja zagrożeń.

Architektury te umożliwiają nie tylko tworzenie realistycznych danych syntetycznych (np. obrazów, tekstów czy dźwięków), ale także znajdują

---

1 *Framework for improving critical infrastructure cybersecurity*, version 1.1, NIST,2018, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf, [dostęp: 10.05.2025].

2 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. *Advances in neural information processing systems 27,* 2014.

3 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Advances in neural information processing systems 30*, Attention is all you need, 2017.

4 A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1), 2018, 53-65.

zastosowanie w złożonych atakach, takich jak deepfake, spear-phishing z wykorzystaniem metod NLP, czy manipulacją danymi treningowymi w procesie uczenia maszynowego[5].

Postęp w dziedzinie generatywnej sztucznej inteligencji (GenAI) przejawia się w dynamicznej ewolucji architektur: od wprowadzenia Generative Adversarial Networks (GAN) w 2014 r.[6], poprzez pojawienie się architektury Transformer w 2017 r.[7], modelu BERT (2018)[8], GPT-2 (2019)[9], aż po coraz bardziej zaawansowane duże modele językowe, takie jak Claude[10], Gemini[11], GPT-4[12], rozwijane w latach 2022-2024. Można zatem przypuszczać, że GenAI i duże modele językowe (LLM) będą odgrywać coraz bardziej istotną rolę zarówno w obszarze cyberobrony – na przykład poprzez automatyzację detekcji zagrożeń – jak i w działaniach ofensywnych, takich jak generowanie kodu złośliwego oprogramowania, tworzenie treści dezinformacyjnych czy realizacja innych strategii ataków cybernetycznych.

---

5 M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersle, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. C. Allen, J. Steinhardt, C. Flynn, S. Ó hÉigeartaigh, SJ Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, D. Amodei, The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Technical report. University of Oxford, 2018, https://maliciousaireport.com/, [dostęp: 10.05.2025].

6 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Advances in neural information processing systems 27* Generative adversarial nets, 2014.

7 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Advances in neural information processing systems 30*, Attention is all you need, 2017.

8 J. Devlin, Ch. Ming-Wei, L. Kenton Lee, K. Toutanova Bert, Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, 4171-4186.

9 I. Solaiman, M. Brundage, J. Clark, *Release strategies and the social impacts of language models*, arXiv preprint arXiv:1908.09203, 2019.

10 Anthropic's All Models Overview: https://docs.anthropic.com/en/docs/ about-claude/models/all-models, [dostęp: 10.05.2025].

11 Gemini: A family of highly capable multimodal models, Google, Team, 2024, https://arxiv.org/abs/2312.11805, [dostęp: 10.05.2025].

12 OpenAI (2023) Gpt-4 technical report.

Materiały i metody

Niniejszy rozdział stanowi przegląd literatury naukowej dotyczącej zastosowań sztucznej inteligencji w obszarze cyberbezpieczeństwa, ze szczególnym uwzględnieniem technik generatywnych, takich jak duże modele językowe. Analiza obejmuje publikacje z lat 2017-2024, które zostały pozyskane z dwóch uznanych baz danych: *IEEE Xplore Digital Library* oraz *Springer Link*. W celu opracowania tematycznej mapy publikacji wykorzystano narzędzie *litstudy*[13], które umożliwiło automatyczne modelowanie tematów, analizę n-gramów w streszczeniach artykułów i wizualizację wyników.

Celem przeprowadzonej analizy literaturowej było określenie, w jakim zakresie generatywna sztuczna inteligencja, a w szczególności LLM, od czasu pojawienia się architektury Transformer w 2017 r. znajduje zastosowanie w kontekście cyberbezpieczeństwa. W tym celu autor przeprowadził systematyczne wyszukiwanie w bazach *IEEE Xplore* oraz *Springer Link*, stosując słowa kluczowe: „cybersecurity" oraz „AI". Z bazy IEEE Xplore pozyskano łącznie 1422 publikacje, wyeksportowane w formacie CSV (jeden dokument na plik). Z bazy Springer Link pobrano dodatkowo 1000 dokumentów. Po scaleniu danych uzyskano zbiór zawierający łącznie 2420 artykułów. Bazy te wybrano na podstawie wcześniejszych doświadczeń autora, wskazujących na to, że publikacje dostępne w IEEE Xplore cechują się silniejszym naciskiem na aspekty techniczne (np. detekcja zagrożeń, klasyfikacja danych), podczas gdy Springer Link prezentuje częściej ujęcia społeczno-etyczne, w tym kwestie prywatności, zgodności z regulacjami oraz percepcji technologii AI.

W ramach przygotowania danych do analizy przeprowadzono proces oczyszczania korpusu. Usunięto dwie nieistotne publikacje załadowane w formacie RIS, które nie spełniały kryteriów merytorycznych. Ostateczny korpus danych został załadowany i przetworzony w narzędziu *litstudy*, co umożliwiło wygenerowanie statystyk opisowych (m.in. histogramów częstości występowania n-gramów) oraz eksplorację tematyczną dokumentów przedstawionych na rysunkach 1, 2, 3 i 4.

Do budowy korpusu zastosowano funkcję *build_corpus* z parametrem *ngram_threshold=0.8*, co umożliwiło uwzględnienie popularnych fraz jako spójnych jednostek leksykalnych (np. artificial_intelligence, cyber_security).

[13] H. Stijn, A. Sclocco, H. Dreuning, B. Van Werkhoven, P. Hijma, J. Maassen, R. V van Nieuwpoort, *litstudy: A python package for literature reviews*. SoftwareX, 20, 101207, 2022.

Następnie, wykorzystując funkcję *compute_word_distribution*, przeprowadzono analizę częstości występowania n-gramów w korpusie. Pozwoliło to zidentyfikować kluczowe pojęcia obecne w literaturze, takie jak: „machine_learning", „deep_learning", „attack_detection", „data_privacy" czy „threat_intelligence".

Rezultaty automatycznego modelowania tematów przedstawiono w postaci wizualizacji krajobrazu tematycznego, opartego o landscape plot, który wykorzystuje nieliniowe metody redukcji wymiarowości, takie jak tSNE[14] i UMAP[15], do rozmieszczenia dokumentów na płaszczyźnie dwuwymiarowej. Mapy takie umożliwiają intuicyjne rozpoznanie grup tematycznych, w których blisko siebie zlokalizowane dokumenty reprezentują podobną treść semantyczną. Należy przy tym podkreślić, że odległości między punktami na wizualizacji mają charakter ilustracyjny i nie powinny być interpretowane w sensie metrycznym.

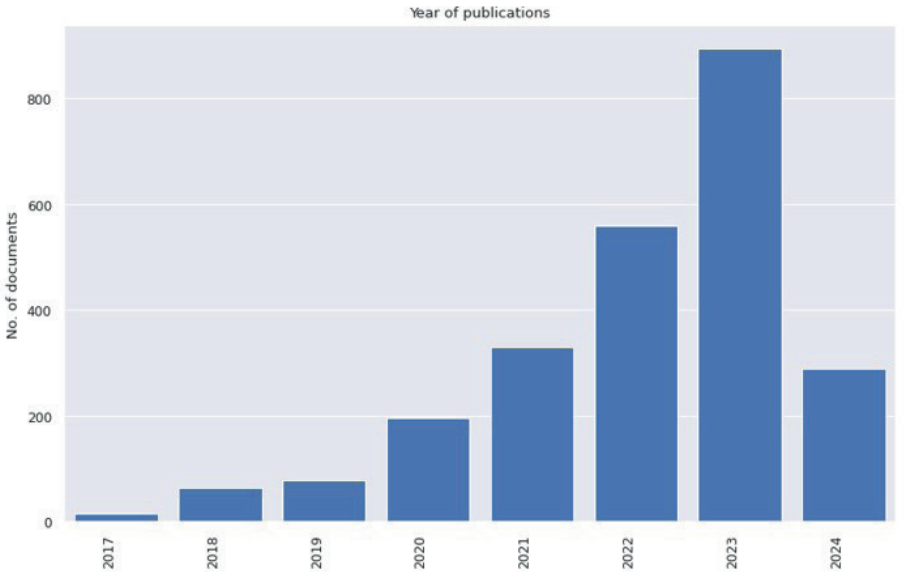
Rezultaty

Zgromadzone wyniki prezentują kompleksowy obraz aktualnego stanu badań na styku sztucznej inteligencji i cyberbezpieczeństwa, podkreślając rosnącą rolę technik generatywnych – zarówno w zakresie detekcji zagrożeń, analiz podatności, symulacji scenariuszy ataków, jak i automatyzacji reakcji na incydenty.

Analiza liczby publikacji w latach 2017-2024 (rysunek 1) ukazuje wyraźny trend wzrostowy. Począwszy od 20 publikacji w 2017 r., ich liczba wzrastała niemal liniowo aż do roku 2020 (200 publikacji), po czym tempo przyrostu uległo znacznemu przyspieszeniu. Szczególnie istotny skok zaobserwowano w latach 2022-2024, gdzie liczba publikacji osiągnęła ponad 400 rocznie. Taki przyrost wskazuje nie tylko na rosnące zainteresowanie tematyką sztucznej inteligencji w kontekście cyberbezpieczeństwa, ale także na zwiększoną świadomość badaczy w zakresie potencjalnych zastosowań obronnych i ofensywnych tych technologii.

[14] L. Van der Maaten, G. Hinton, *Visualizing data using t-sne*. Journal of machine learning research, 9(11), 2008.

[15] L. McInnes, J. Healy, J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction,* arXiv preprint arXiv:1802.03426, 2018.

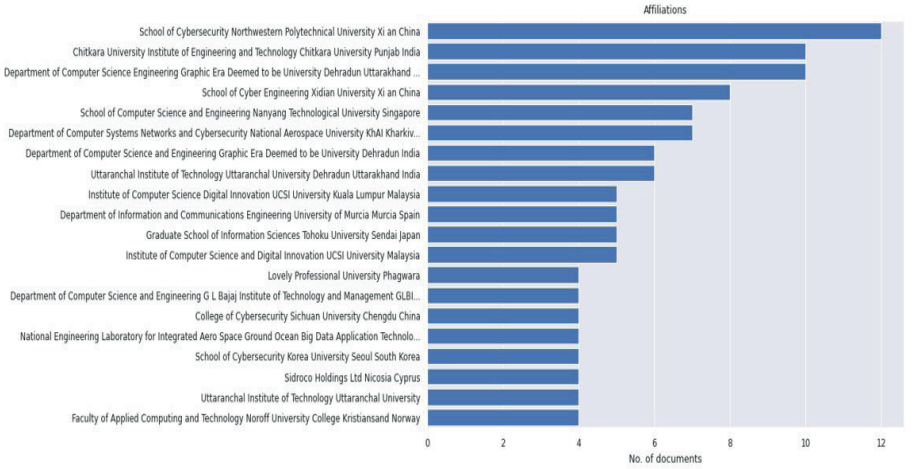


Rysunek 1. Sztuczna inteligencja i cyberbezpieczeństwo, ilość publikacji w latach 2017-2024

Źródło: opracowanie własne.

Analiza danych wskazuje też potencjalnie najaktywniejsze instytucje naukowe w zakresie badań nad zastosowaniami sztucznej inteligencji w cyberbezpieczeństwie. Wśród 20 najczęściej pojawiających się afiliacji dominują ośrodki z Chin i Indii, co odzwierciedla rosnącą rolę Azji w globalnym ekosystemie badawczym. Znaczącą obecność mają również uczelnie i instytuty badawcze z Japonii, Korei Południowej, Singapuru i Malezji. W zestawieniu znalazły się również wybrane europejskie jednostki z Hiszpanii, Norwegii, Cypru oraz Ukrainy (Charków).

Rysunek 2 ukazuje szczegóły 20 najczęściej reprezentowanych afiliacji badawczych w analizowanym korpusie literatury. Widoczna dominacja Azji, zwłaszcza Chin i Indii, może wynikać z dużej liczby publikacji technicznych oraz intensywnego rozwoju krajowych programów AI. Udział Europy, choć mniej liczny, obejmuje ośrodki prowadzące badania nad aspektami etycznymi i regulacyjnymi, co sugeruje zróżnicowanie podejść badawczych zależnie od regionu. Należy zwrócić uwagę na obecność Ukrainy, co może świadczyć o dynamicznym rozwoju kompetencji w obszarze cyberbezpieczeństwa, również wśród naukowców z krajów Europy Wschodniej, gdzie obecnie toczy się wojna.

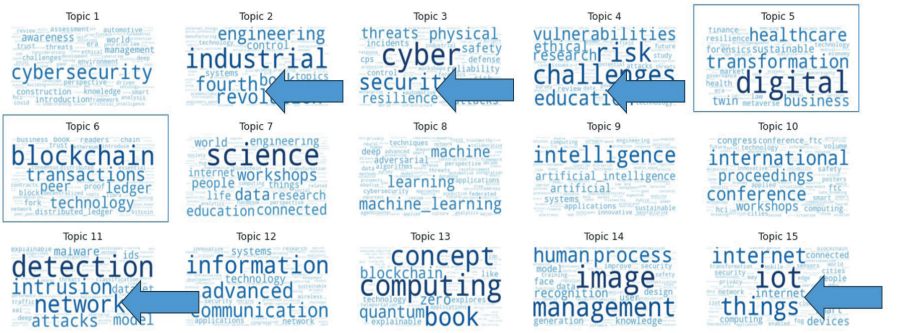


Rysunek 2. Dwadzieścia najczęstszych afiliacji w zebranym zbiorze danych
Źródło: opracowanie własne.

Wyniki automatycznego modelowania tematów w analizowanym korpusie publikacji pozwalają na wyodrębnienie kilku kluczowych obszarów badawczych w dziedzinie cyberbezpieczeństwa. Najsilniej reprezentowane są zagadnienia związane z wykrywaniem intruzów w sieci (intrusion detection) oraz z odpornością systemów informatycznych na różnego rodzaju ataki. W tym kontekście szczególną rolę odgrywają badania poświęcone zarządzaniu ryzykiem, identyfikacji podatności oraz redukcji powierzchni ataku. Istotną część dyskursu zajmują również kwestie bezpieczeństwa infrastruktury przemysłowej oraz zastosowań technologii Internetu Rzeczy (IoT) i przetwarzania brzegowego (edge computing) w kontekście cyberzagrożeń.

Rysunek 3 przedstawia wizualizację tematów uzyskanych za pomocą modelowania nienadzorowanego, gdzie zidentyfikowano kilka wyraźnych klastrów tematycznych. Jeden z najsilniejszych skupia publikacje dotyczące wykrywania intruzów (klaster 11), co potwierdza wagę tego zagadnienia w bieżących badaniach. Widoczne są także klastry związane z bezpieczeństwem fizycznym i infrastrukturą przemysłową (np. klaster 3 – fizyczna ochrona, klaster 15 – IoT), wskazujące na coraz większe zainteresowanie obszarem systemów cyber-fizycznych. Sposób rozmieszczenia dokumentów na rysunku potwierdza również zróżnicowanie podejść – od technicznych po systemowe i organizacyjne.

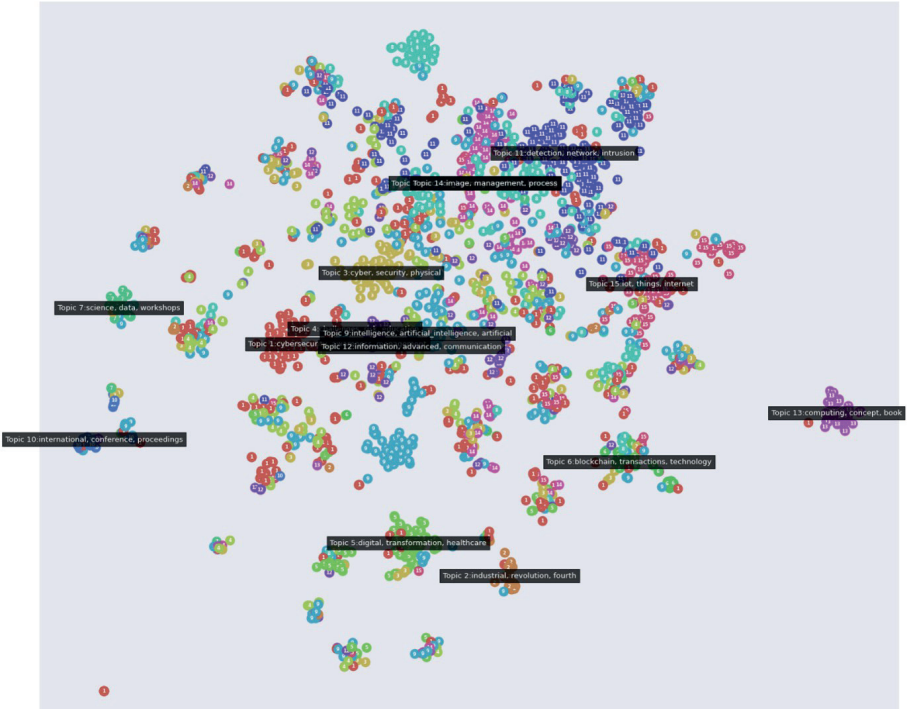
Co interesujące, wśród analizowanych publikacji zauważalna jest obecność tematów związanych z technologią blockchain oraz cyfrową transformacją, szczególnie w sektorze usług zdrowotnych (klaster 5).



Rysunek 3. Automatyczne modelowanie tematów w zbiorze danych
Źródło: opracowanie własne.

Zagadnienia związane z dużymi modelami językowymi, generatywną sztuczną inteligencją oraz dezinformacją nie tworzą jeszcze jednoznacznie wyodrębnionego klastra tematycznego w ramach analizowanego korpusu publikacji. Wydaje się, że publikacje dotyczące tych obszarów są rozproszone w innych obszarach badawczych. Może to świadczyć o ich przekrojowym charakterze, jak również o relatywnie wczesnym etapie integracji metod generatywnej sztucznej inteligencji w badaniach nad cyberbezpieczeństwem.

Natomiast rysunek 4 ilustruje rozkład klastrów tematycznych w przestrzeni dwuwymiarowej zredukowanej przy użyciu technik t-SNE i UMAP. Klaster 1 (kolor czerwony) charakteryzuje się największą dyspersją, co oznacza, że publikacje w jego obrębie są semantycznie heterogeniczne i nie koncentrują się wokół jednego konkretnego zagadnienia. Obecność tematów LLM, GenAI i dezinformacji właśnie w tym klastrze sugeruje ich wieloaspektowość oraz interdyscyplinarny charakter, który przenika zarówno sfery techniczne (np. bezpieczeństwo modeli), jak i społeczne (np. manipulacja informacją). Brak wyraźnego skupienia wskazuje na konieczność dalszych badań w celu lepszego zrozumienia i klasyfikacji tych zjawisk w kontekście cyberbezpieczeństwa.



Rysunek 4. Automatyczne modelowanie tematów w zbiorze danych.
Źródło: opracowanie własne.

Jak wynika z analizy zawartej w tabeli 1, phishing jest najczęściej występującym tematem w badanych publikacjach, odpowiadającym za 16,82% całkowitej liczby prac. Zbliżony udział mają również zagadnienia związane z dużymi modelami językowymi (LLM) oraz narzędziami typu ChatGPT – oba występują z częstotliwością 14,55%. Nieco niższy, choć nadal znaczący odsetek, dotyczy tematyki ataków typu adversarial (14,34%), natomiast kwestie podatności (vulnerabilities) zajmują relatywnie mniejszy udział (3,18%).

Tabela 1. Wyniki analizy tematów (Top 5) w publikacjach dotyczących cyberbezpieczeństwa

Temat	Liczba Publikacji	Procentowy Udział
Phishing	407	16,82%
Adversarial	347	14,34%
ChatGPT	352	14,55%
LLMs	352	14,55%
Vulnerabilities	77	3,18%

Źródło: opracowanie własne.

Tabela 1 przedstawia zestawienie pięciu najczęściej podejmowanych tematów w publikacjach naukowych analizujących zastosowania sztucznej inteligencji w cyberbezpieczeństwie. Dominacja phishingu potwierdza, że zagrożenia związane z inżynierią społeczną nadal są jednym z głównych wyzwań w środowisku cyfrowym. Co interesujące, niemal równy udział tematów związanych z LLM i ChatGPT sugeruje, że badacze rozróżniają specyficzne wyzwania generatywnych chatbotów od ogólnego zagadnienia dużych modeli językowych. Wysoka pozycja tematów związanych z antagonistycznym uczeniem maszynowym (ang. adversarial machine learninig) świadczy o rosnącej świadomości zagrożeń wynikających z manipulowania danymi wejściowymi do systemów SI. Niewielki udział tematów związanych z podatnościami może z kolei świadczyć o niedoreprezentowaniu tej kategorii w analizowanej literaturze lub o trudności klasyfikacyjnej tych prac w ramach bi-gramowego modelowania tematycznego.

Wpływ LLM i GenAI na cyberbezpieczeństwo

W ostatnich latach generatywna sztuczna inteligencja zaczęła odgrywać coraz istotniejszą rolę zarówno w działaniach defensywnych, jak i ofensywnych w cyberprzestrzeni[16]. Modele generatywne, takie jak sieci przeciwnikowe (GAN), architektury typu Transformer oraz duże modele językowe (LLM) – w tym GPT-4 czy Claude – znajdują szerokie zastosowanie w analizie zagrożeń, automatyzacji testów penetracyjnych oraz detekcji podatności w kodzie źródłowy[17]. Jednocześnie, te same mechanizmy są wykorzystywane przez podmioty złośliwe do generowania realistycznych ataków phishingowych, automatycznego przełamywania zabezpieczeń czy tworzenia złożonego złośliwego oprogramowania[18].

Warto podkreślić, że dwuznaczność zastosowań GenAI wymaga nowego podejścia do klasyfikacji zagrożeń – wiele z tych samych narzędzi może być zarówno wsparciem w cyberobronie, jak i komponentem do przeprowadzania

[16] Y. Yigit, W.J. Buchanan, M.G. Tehrani, L. Maglaras, *Review of generative ai methods in cybersecurity*, arXiv preprint arXiv:2403.08701, 2024.

[17] M. Fu, Ch. Tantithamthavorn, Van Nguyen, T. Le, *Chatgpt for vulnerability detection, classification, and repair: How far are we?* In: 2023 30th Asia-Pacific Software Engineering Conference (APSEC) , 632-636. IEEE, 2023.

[18] M. Gupta, A. CharanKumar, A. Kshitiz, E. Parker, L. Praharaj, *From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy,* IEEE Access 11: 80218-80245, 2023.

zautomatyzowanego ataku. Obserwowany rozwój takich technologii zmusza organizacje do ponownego przemyślenia modeli ryzyka i strategii zabezpieczeń, uwzględniających nie tylko skuteczność techniczną, ale również odporność na manipulacje i nadużycia. Pojawia się potrzeba dynamicznych ram regulacyjnych oraz rozwiązań z zakresu cyberodporności, które będą w stanie adaptować organizacje do zmieniającego się charakteru zagrożeń.

Zastosowania sztucznej inteligencji w cyberprzestępczości

Analiza wskazuje, że w ostatnich latach obserwuje się istotną ewolucję złośliwego oprogramowania, katalizowaną przez rozwój dużych modeli językowych. Modele takie jak ChatGPT umożliwiają tworzenie metamorficznego malware[19], dynamicznie modyfikującego swoją strukturę w celu unikania detekcji sygnaturowej i behawioralnej. Przykładowo, malware może zmieniać algorytmy szyfrowania, generować losowo zaciemniony kod przy każdej kompilacji lub rekonfigurować strukturę plików, co znacząco utrudnia analizę i neutralizację. Automatyzacja tych działań przez LLM prowadzi do powstania inteligentnych agentów malware, zdolnych do autonomicznej adaptacji w zależności od środowiska. W połączeniu z powszechnie dostępną infrastrukturą chmurową, złośliwe oprogramowanie może dynamicznie pobierać instrukcje z LLM w czasie rzeczywistym, zyskując wyjątkową elastyczność i trudność wykrycia w systemach produkcyjnych.

Badania[20] dowodzą, że generatywna sztuczna inteligencja może być wykorzystywana do tworzenia zautomatyzowanych narzędzi ransomware, które nie tylko lokalizują i szyfrują dane, lecz także modyfikują swój kod w czasie rzeczywistym, aby unikać detekcji. Szczególnie niepokojące jest osadzanie interpretera Pythona w złośliwym kodzie, umożliwiające dynamiczną komunikację z LLM (np. ChatGPT) w celu generowania nowych fragmentów malware („on-the-fly malware generation"). Powstają również specjalistyczne frameworki, takie jak ThreatGPT, oparte na jailbreakowanych wersjach modeli, które umożliwiają generowanie kodu, scenariuszy ataków, a nawet

[19] P. Madani, *Metamorphic malware evolution: The potential and peril of large language models,* In: 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) , 74-81. IEEE, 2023.

[20] M. Gupta, A. CharanKumar, A. Kshitiz, E. Parker, L. Praharaj, *From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy*, IEEE Access 11: 80218-80245, 2023.

symulację socjotechnicznych interakcji. W połączeniu z lukami sprzętowymi (Meltdown, Spectre), cyberprzestępcy mogą zautomatyzować niemal każdy etap ataku – od wykonania rozpoznania, po jego automatyczne, lub półautomatyczne wykonanie.

Rozwój otwartych modeli językowych (np. Chiński 01.ai[21] oraz publicznych repozytoriów takich jak PentestGPT, ChatGPT_DAN[22] uwidacznia zagrożenia związane z jailbreakami typu „Do Anything Now" (DAN). Umożliwiają one stosunkowo łatwe obchodzenie czasami minimalnych jeszcze zabezpieczeń modeli i wykorzystanie ich jako komponentów złośliwego oprogramowania, np. jak w przypadku wspomnianego ThreatGPT. To w połączeniu z istniejącymi lukami sprzętowymi i wcześniej opisanymi atakami typu ransomware[23] (szyfrującymi dane, np. WannaCry, Ryuk, REvil), czy eraser (NotPetya), tworzą nową klasę zagrożeń.

Podsumowując, LLM-y mogą być więc już obecnie wykorzystywane nie tylko do tworzenia i modyfikacji kodu źródłowego, lecz także do automatyzacji całych łańcuchów ataku, np. poprzez generowanie realistycznych wiadomości phishingowych, czy omijanie systemów ochrony (np. Web Application Firewall). Jako że znacząco ułatwiają one tworzenie zaawansowanych narzędzi ofensywnych i automatyzację działań, zmniejszają one bariery wyjścia i umożliwiają działanie nawet mniej doświadczonym w technologiach informatycznych cyberprzestępcom.

Rozwój mechanizmów obronnych

Wykorzystanie generatywnej sztucznej inteligencji niesie ze sobą również istotne korzyści w dziedzinie cyberbezpieczeństwa. Jednym z kluczowych aspektów jest automatyzacja analizy incydentów bezpieczeństwa. Narzędzia takie jak SecurityLLM i FalconLLM umożliwiają szybkie identyfikowanie wzorców złośliwego kodu w rozległych zbiorach danych[24]. Ponadto, przy pomocy LLMów implementowane są mechanizmy wykrywania deepfake'ów

[21] *Yi Foundation Models*, https://www.01.ai/#yi-foundation-models, [dostęp: 10.05.2025].

[22] *ChatGPT DAN*, https://github.com/0xk1h0/ChatGPT_DAN, [dostęp: 10.05.2025].

[23] *Cloudflare "What are Petya and NotPetya?"*, https://www.cloudflare.com/learning/ security/ ransomware/petya-notpetya-ransomware/, [dostęp: 10.05.2025].

[24] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, T. Lestable, *Revolutionizing cyber threat detection with large language models*, *arXiv preprint arXiv:2306.14263*: 2023, 195-202,

i treści phishingowych, co przyczynia się do ograniczenia skuteczności ataków socjotechnicznych[25].

Sztuczna inteligencja znacząco wspomaga też analizę podatności. Narzędzia pokroju DeepExploit[26] i PentestGPT[27] pozwalają na automatyczne skanowanie i ocenę bezpieczeństwa aplikacji oraz infrastruktury IT. Wraz z postępem technologicznym, coraz większe znaczenie zyskuje automatyczna detekcja kodu malware, realizowana za pomocą systemów AI uczonych na obszernych bazach danych złośliwego oprogramowania[28].

Automatyzacja zadań w cyberbezpieczeństwie obejmuje również analizę logów (LogGPT), automatyczne tworzenie raportów o incydentach oraz wsparcie testów penetracyjnych z wykorzystaniem systemów, takich jak DeepExploit i PentestGPT[29]. Coraz istotniejszą rolę odgrywają także narzędzia wspomagające statyczną i dynamiczną analizę kodu (np. GitHub Copilot, SAST/DAST) oraz systemy do syntetycznego generowania danych na potrzeby uczenia maszynowego (np. PAC-GPT3).

Wzrost liczby defensywnych rozwiązań opartych na GenAI dowodzi, że technologia ta, mimo generowania nowych zagrożeń, stanowi kluczowy element nowoczesnych strategii cyberobrony. Integracja tych systemów z normami zarządzania ryzykiem – takimi jak ISO/IEC 42001 czy dyrektywa NIS2 – jest niezbędna dla zapewnienia zarówno skuteczności operacyjnej, jak

---

[25] S. Enathe VP, Ch. Mouli S, R. Dheepthi, *Llm-enhanced deepfake detection: Dense cnn and multi-modal fusion framework for precise multimedia authentication*. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), IEEE. 2024, 1-6.

[26] A. Sychugov, M. Grekov, *Automated penetration testing based on adversarial inverse reinforcement learning*, In: 2024 International Russian Smart Industry Conference (SmartIndustryCon) , IEEE, 2024, 373-377.

[27] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, S. Rass, {PentestGPT}: Evaluating and harnessing large language models for automated penetration testing. In: 33rd USENIX Security Symposium (USENIX Security 24) , 2024, 847–864.

[28] O. Veprytska, V. Kharchenko, AI *powered attacks against ai powered protection: classification, scenarios and risk analysis*, In: 2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), IEEE, 2022, 1-7.

[29] Q. Jiaxing, S. Huang, Z. Luan, S. Yang, C. Fung, H. Yang, D. Qian, J. Shang, Z. Xiao, Z. Wu, *Loggpt: Exploring chatgpt for log-based anomaly detection*. In: 2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), IEEE, 2023, 273-280.

i zgodności z ramami prawnymi i etycznymi. Nowoczesne podejście do bezpieczeństwa informacji wymaga zatem równoległego rozwoju technologicznego oraz wdrożenia mechanizmów kontroli organizacyjnej i audytowalności działań. Wspomniane zastosowania to:

- automatyzacja analizy incydentów (SecurityLLM/FalconLLM),
- generowanie i wykrywanie złośliwego oprogramowania (MalGPT),
- automatyzacja testów penetracyjnych (DeepExploit/PentestGPT),
- automatyzacja analizy logów (LogGPT),
- generowanie i wykrywanie bezpiecznego kodu (SAST/DAST, GitHub Copilot),
- identyfikacja cyberataków: (SecurityLLM/FalconLLM/SecurityBERT),
- generowanie zbiorów danych: PAC-GPT3 do generowania pakietów dla uczenia maszynowego.

Rozwój Standardów Cyberbezpieczeństwa

Wraz z dynamicznym rozwojem i rosnącym znaczeniem generatywnej sztucznej inteligencji i dużych modeli językowych w cyberbezpieczeństwie, imperatywem staje się opracowanie adekwatnych ram regulacyjnych i standardów bezpieczeństwa, koncentrujących się na zarządzaniu ryzykiem związanym z implementacją systemów AI. Kluczowe jest zarówno dostosowanie istniejących norm prawnych i standardów do specyfiki, jak i wypracowanie nowych regulacji dedykowanych temu obszarowi. Niezbędna jest również intensyfikacja współpracy międzynarodowej oraz wymiana doświadczeń w celu skutecznej prewencji zagrożeń wynikających z ewolucji metod sztucznej inteligencji.

Na poziomie Unii Europejskiej, dyrektywa NIS2 oraz norma ISO/IEC 42001 stanowią istotne ramy zarządzania ryzykiem związanym z wdrażaniem SI w organizacjach. Fundamentalnym wyzwaniem pozostaje zabezpieczenie systemów przed potencjalnymi nadużyciami oraz zapewnienie transparentności procesów decyzyjnych realizowanych przez modele generatywne.

W odpowiedzi na zmieniające i zwiększające się ryzyka, obserwuje się implementację nowych regulacji prawnych i standardów bezpieczeństwa. Dyrektywa NIS2 redefiniuje wymogi dotyczące ochrony infrastruktury krytycznej i cyfrowej, natomiast normy ISO/IEC 22989, 23053, 23984 oraz 42001 precyzują zasady zarządzania ryzykiem w kontekście wdrażania AI.

Niniejsza sekcja analizuje kluczowe zmiany w standardach mających implikacje dla cyberbezpieczeństwa. Normy i standardy odgrywają fundamentalną rolę we wspieraniu organizacji w projektowaniu i implementacji bezpiecznych rozwiązań opartych na sztucznej inteligencji. Każda z wymienionych norm ISO odnosi się, w różnym zakresie, do problematyki ryzyka związanego z AI. W kontekście projektowania i wdrażania systemów opartych na AI, zwłaszcza w obszarze cyberbezpieczeństwa, rekomenduje się szczegółowe zapoznanie z wyżej wymienionymi standardami. Zawarte w nich wytyczne i zalecenia mogą zostać przełożone na konkretne mechanizmy kontrolne, stanowiące integralną część dokumentacji polityki bezpieczeństwa. Niezbędna jest systematyczna aktualizacja polityk bezpieczeństwa i ich transpozycja na szczegółowe procedury zarządzania ryzykiem informatycznym, uwzględniająca specyfikę i wykorzystanie nowych metod generatywnych i LLM – w tym zarządzanie kontekstem organizacyjnym oraz wdrożenie własnych, zabezpieczonych i odseparowanych od sieci Internet dużych modeli językowych.

Istotnym wkładem w systematyzację zarządzania ryzykiem związanym z AI jest opracowany przez NIST Artificial Intelligence Risk Management Framework, stanowiący kluczowy dokument dla organizacji zaangażowanych w tworzenie i wdrażanie „bezpiecznej i godnej zaufania" sztucznej inteligencji.

W kontekście regulacji prawnych, dyrektywa NIS2, obowiązująca od 2024 r., wprowadza znaczące zmiany mające na celu podniesienie poziomu bezpieczeństwa sieci i systemów informacyjnych w Unii Europejskiej. Implementuje nowe wymogi dotyczące zarządzania ryzykiem oraz raportowania incydentów, wzmacniając odporność na cyberataki. Sankcje za nieprzestrzeganie dyrektywy mogą sięgać maksymalnie 10 000 000 EUR lub do 2% całkowitego rocznego światowego obrotu danego przedsiębiorstwa, w zależności od tego, która kwota jest wyższa. Dyrektywa obejmuje podmioty istotne (np. dostawców usług cyfrowych, sektory produkcji, przetwórstwa i dystrybucji żywności) oraz niezbędne (w tym administrację publiczną).

Podsumowanie kluczowych, według autora nowych regulacji prawnych i standardów dotyczących sztucznej inteligencji i cyberbezpieczeństwa:

- NIS2: Redefinicja standardów bezpieczeństwa sieci i systemów informacyjnych.
- ISO/IEC 22989:2022: Technologia informacyjna – Sztuczna inteligencja – Pojęcia i terminologia.

- ISO/IEC 23053:2022: Ramy dla systemów SI wykorzystujących uczenie maszynowe.
- ISO/IEC 23984:2023: Technologia informacyjna – Sztuczna inteligencja – Wytyczne dotyczące zarządzania ryzykiem.
- ISO/IEC 42001:2023: Technologia informacyjna – Sztuczna inteligencja – System zarządzania.
- NIST Artificial Intelligence Risk Management Framework.

Należy zauważyć, że NIST Artificial Intelligence Risk Management Framework jest w początkowej fazie rozwoju, a wiele organizacji publikuje własne wytyczne i zasady dotyczące AI, co wskazuje na potrzebę dalszej systematyzacji i standaryzacji regulacji w tym obszarze.

Podsumowanie

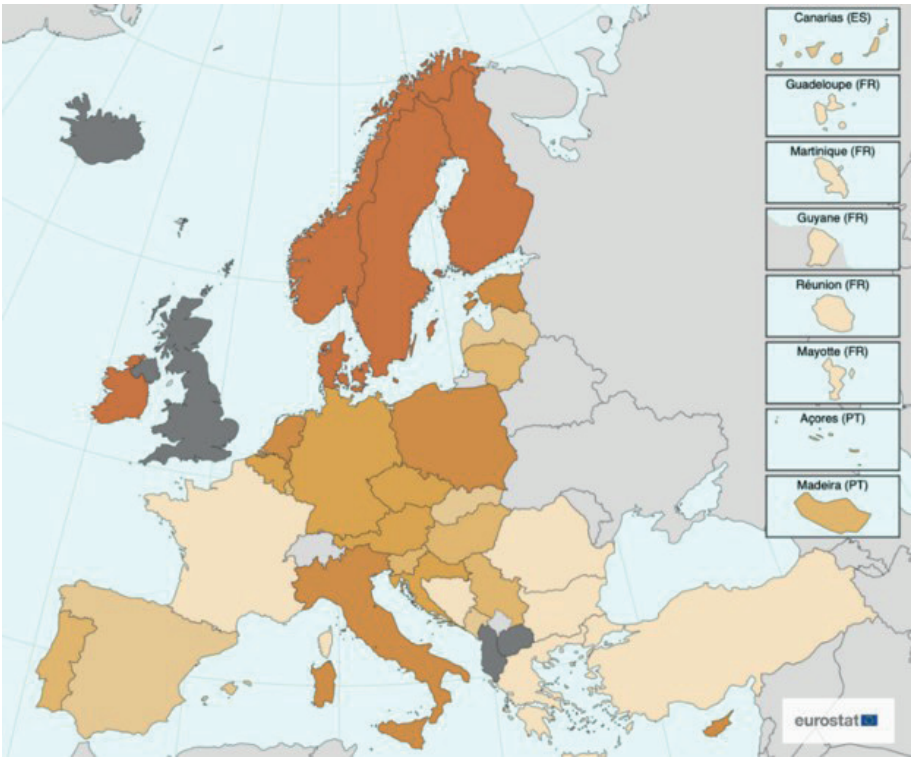
W niniejszym rozdziale przedstawiono wybrane aspekty rozwoju sztucznej inteligencji i jej wpływu na bezpieczeństwo systemów informatycznych, zwracając szczególną uwagę na jej nowe generatywne odmiany. Zaprezentowano przegląd literatury z lat 2017-2024, obejmujący automatyczne modelowanie tematów oraz analizę trendów w publikacjach naukowych dotyczących sztucznej inteligencji i cyberbezpieczeństwa. Dokonano przeglądu ponad 2420 publikacji naukowych z zasobów IEEE Xplore Digital Library i SpringerLink. Wykonana analiza ilustruje wykładniczy wzrost zainteresowania badawczego tematyką sztucznej inteligencji i cyberbezpieczeństwa.

Zdaniem autora, w obliczu rosnącej liczby ataków wykorzystujących sztuczną inteligencję, organizacje powinny inwestować w technologie zabezpieczające systemy oparte na AI oraz rozwijać kompetencje pracowników w zakresie analizy zagrożeń i reagowania na incydenty przy użyciu tych samych metod.

Współpraca publiczno-prywatna oraz wymiana informacji o zagrożeniach będą kluczowe dla skutecznej ochrony. Istotne będzie też dostosowanie istniejących norm zarządzania ryzykiem do specyfiki generatywnej sztucznej inteligencji i LLM oraz opracowanie nowych standardów zarządzania ryzykiem związanych z jej szerokim wdrażaniem. W wyniku analizy ustalono, że współczesne zagrożenia cybernetyczne coraz częściej wykorzystują zaawansowaną sztuczną inteligencję i inżynierię społeczną. Phishing i zautomatyzowane techniki hakowania, np. oferowane przez PentestGPT, stają się bardziej

wyrafinowane, umożliwiając tworzenie bardziej przekonujących wiadomości i skuteczniejszych wektorów ataku. Generowanie złośliwych ładunków, inspirowanych podejściem WannaCry, NotPetya, Ryuk czy Revili ransomware staje się łatwiejsze, co ilustruje potencjał AI w działaniach ofensywnych cyberbezpieczeństwa. Szczególnie groźne stają się techniki polimorficznego malware, skutecznie unikające detekcji, tworzonego przy użyciu (jailbreakowanych) modeli LLM (np. metoda DAN). Ponadto, sztuczna inteligencja znajduje zastosowanie w dekonstrukcji algorytmów kryptograficznych, w tym AES i CHAM, podważając bezpieczeństwo fundamentalnych mechanizmów ochrony danych. Te zjawiska wskazują na zacieranie się granic między defensywą, a ofensywą w cyberprzestrzeni oraz rosnącą rolę nowych metod w tym ekosystemie.

W odpowiedzi na rosnące zagrożenia cybernetyczne, coraz większe znaczenie zyskują zautomatyzowane metody detekcji i analizy anomalii w systemach IT. Przykładem jest wykorzystanie skryptów PowerShell do monitorowania infrastruktury, np. identyfikowania zapytań do baz danych Microsoft SQL obciążających procesor, mogących wskazywać na nieoptymalne działanie lub atak. Rozwijają się również rozwiązania do automatycznej analizy danych o zagrożeniach z wielu źródeł, wspierające szybkie reagowanie i podejmowanie decyzji w zakresie cyberbezpieczeństwa. Integralną częścią proaktywnej ochrony aplikacji są techniki SAST i DAST, pozwalające wykrywać podatności na etapie kodowania i działania aplikacji. Wspólne zastosowanie tych narzędzi zwiększa odporność systemów i ogranicza powierzchnię ataku.



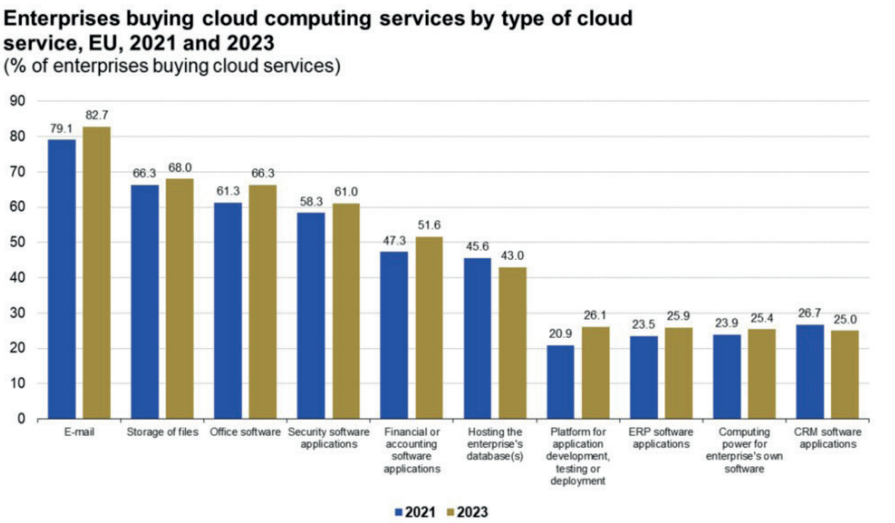
Rysunek 5. Adopcja chmury obliczeniowej w Unii Europejskiej w latach 2021-2023
Źródło: Raport Eurostat (2023) "EU Survey on ICT and e-commerce usage in enterprises".

W tym miejscu warto wspomnieć, że w Polsce w latach 2021-2023 odnotowano zauważalny wzrost wykorzystania publicznej chmury obliczeniowej. W tym czasie na polskim rynku IT nastąpiły znaczące zmiany, w tym powstanie regionu Google Cloud (2021 r.), strefy lokalnej Amazon Web Services (2022 r.) oraz regionu Microsoft Azure (2023 r.) w Warszawie, co według autora mogło przyczynić się do wzrostu zainteresowania chmurą obliczeniową w naszym kraju.

Rysunek 5 przedstawia wyniki badania usług IT i handlu elektronicznego Eurostat (2023 r.) w Unii Europejskiej przeprowadzonego przez Eurostat. Statystyki uzyskano z badań przeprowadzonych przez krajowe urzędy statystyczne w pierwszych miesiącach 2023 r. na próbie około 161 000 przedsiębiorstw zatrudniających 10 lub więcej pracowników lub osób samozatrudnionych (spośród 1,5 miliona takich organizacji w całej UE). Wśród nich 83% stanowiły małe przedsiębiorstwa (10-49 pracowników), 14% średnie (50-249)

i 3% duże (250 lub więcej). Badanie to ukazuje, że w 2023 r. 55,7% polskich przedsiębiorstw korzystało z usług chmury obliczeniowej, co stanowi wzrost o 10,5 punktu procentowego w porównaniu do 2021 r. Polska zajmuje obecnie 10. miejsce w UE pod względem wykorzystania chmury obliczeniowej. Najwyższy odsetek przedsiębiorstw korzystających z usług cloud computing odnotowano w Finlandii (78,3%), Szwecji (71,6%), Danii (69,5%) i na Malcie (66,7%), a najniższy w Grecji (23,6%), Rumunii (18,4%) i Bułgarii (17,5%).

Z badania wiemy też, że zaawansowane usługi przetwarzania w chmurze zakupiło 83,6% dużych przedsiębiorstw, w porównaniu do 77,4% średnich i 74,2% małych przedsiębiorstw korzystających z tych usług. Najczęściej kupowanymi usługami w chmurze obliczeniowej w 2023 r. były proste usługi poczty e-mail (82,7%), przechowywania plików (68%) i oprogramowanie biurowe (66,3%), a także aplikacje zabezpieczające (61,0%), finansowe lub księgowe (51,6%) oraz hosting baz danych (43,0%). Około jedna czwarta firm zakupiła platformy obliczeniowe do tworzenia, testowania lub wdrażania aplikacji (26,1%), oprogramowanie ERP (25,9%), moc obliczeniową do uruchamiania własnego oprogramowania (25,4%) lub oprogramowanie CRM (25,0%), co widać na przytaczanym rysunku 6.



Rysunek 6. Konsumpcja usług chmury obliczeniowej w Unii Europejskiej w latach 2021 i 2023

Źródło: Raport Eurostat (2023) "EU Survey on ICT and e-commerce usage in enterprises".

Autor uważa, że redefiniuje to typowe „granice organizacyjne", które wymagają ochrony. Można więc założyć, że w tak zmieniającym się środowisku, w związku z nowymi metodami i narzędziami SI (GenAI/LLM) większość danych organizacji w UE będzie wkrótce generowanych i przetwarzanych poza tradycyjnymi centrami danych, koncentrując potencjalnie ataki na API systemów i urządzeniach tzw. „przetwarzania brzegowego" tych organizacji. Przetwarzanie brzegowe to model rozproszonych obliczeń, przesuwający obliczenia i przechowywanie danych bliżej ich źródeł, w bardziej rozproszony sposób, co oczywiście dla zespołów cyberbezpieczeństwa generuje nowe wyzwania.

Wracając do wyników, dokonano przeglądu ponad 2420 publikacji naukowych z zasobów IEEE Xplore Digital Library i SpringerLink, koncentrują się one głównie na tematyce (1) phishingu (407 publikacji, 16,82%), (2) adversarial (347 publikacji, 14,34%), (3) ChatGPT (352 publikacji, 14,55%), (4) LLMs (352 publikacji, 14,55%) i podatności (77 publikacji, 3,18%).

Widać wyraźnie, że obecność tematów związanych z szeroko rozumianą generatywną sztuczną inteligencją (LLM i ChatGPT) jest sumarycznie większa (704 dokumenty) niż nacisk kładziony na bardziej tradycyjne metody uczenia maszynowego, w tym zagrożenia związane z manipulowaniem danymi wejściowymi do systemów AI. Tematyka Adversarial to tylko 347 dokumentów.

Sztuczna inteligencja wykazuje rosnący potencjał w cyberbezpieczeństwie, zarówno w obronie, jak i ataku. Wraz z rozwojem tych technologii, organizacje muszą dostosować strategie ochrony przed nowymi zagrożeniami. Kluczowe będzie także opracowanie i wdrożenie nowych regulacji prawnych oraz standardów bezpieczeństwa uwzględniających AI. Czego więc możemy się spodziewać w przyszłości?

- Duże modele językowe mogą zostać zinfiltrowane ze względu na potencjalne niedostateczne zabezpieczenie baz danych.
- Narzędzia oparte na generatywnej sztucznej inteligencji mają potencjał znaczącej zmiany sposobu prowadzenia operacji dezinformacyjnych poprzez masowe tworzenie fałszywych treści.
- Ułatwiona dostępność nowych narzędzi wspiera rozproszone ataki, takie jak DoS, wycieki danych, podmiany stron internetowych i próby zakłócenia infrastruktury krytycznej.
- Zmieniające się środowisko cyberprzestrzeni w Polsce – wzrost znaczenia usług chmury publicznej i przetwarzania brzegowego.

W związku z tym konieczne staje się wdrażanie nowych standardów bezpieczeństwa, uwzględniających nowe metody AI, oraz zmieniające się środowisko działania organizacji (wspomniana adopcja chmury publicznej i przetwarzania brzegowego). Czeka nas również wdrożenie dyrektywy NIS2 i nowych standardów takich jak ISO/IEC 22989:2022, ISO/IEC 23053:2022, ISO/IEC 23984:2023, ISO/IEC 42001:2023, czy NIST Artificial Intelligence Risk Management Framework.

Autor serdecznie dziękuje Wiktorowi Szymanikowi za cenne uwagi, które przyczyniły się do rozwoju niniejszego rozdziału.

Literatura

1. Creswell, Antonia; Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta & Anil A Bharath, *Generative adversarial networks: An overview,* IEEE signal processing magazine 35(1), 2018.
2. Deng, Gelei; Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger & Stefan Rass, *{PentestGPT}: Evaluating and harnessing large language models for automated penetration testing*. In: 33rd USENIX Security Symposium (USENIX Security 24), 2024.
3. Devlin, Jacob; Ming-Wei Chang, Kenton Lee & Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019.
4. Eurostat, *Cloud computing – statistics on the use by enterprises from the 2023 EU survey on ICT usage and e-commerce in enterprises*. Technical report. Eurostat 2023, https://ec.europa.eu/eurostat/statisticsexplained/index.php
5. Ferrag, Mohamed Amine; Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C Cordeiro, Merouane Debbah & Thierry Lestable, Revolutionizing cyber threat detection with large language models. *arXiv preprint arXiv:2306.14263*, 2023.
6. Fu, Michael; Chakkrit Kla Tantithamthavorn, Van Nguyen & Trung Le, *Chatgpt for vulnerability detection, classification, and repair: How far are we?,* In: 2023 30th Asia-Pacific Software Engineering Conference (APSEC), IEEE, 2023.
7. Goodfellow, Ian J; Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville & Yoshua Bengio, *Generative adversarial nets,* Advances in neural information processing systems 27, 2014.
8. Google, Team, *Gemini: A family of highly capable multimodal models 2024*, https://arxiv.org/abs/2312.11805
9. Gupta, Maanak; CharanKumar Akiri, Kshitiz Aryal, Eli Parker & Lopamudra Praharaj, *From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy*. IEEE Access 11, 2023.

10. Heldens, Stijn; Alessio Sclocco, Henk Dreuning, Ben Van Werkhoven, Pieter Hijma, Jason Maassen & Rob V van Nieuwpoort, *litstudy: A python package for literature reviews*. SoftwareX 20, 2022.
11. Madani, Pooria, *Metamorphic malware evolution: The potential and peril of large language models*. In: 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), IEEE, 2023.
12. McInnes, Leland; John Healy & James Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv preprint arXiv:1802.03426, 2018.
13. Miles Brundage, Shahar Avin, Jack Clark Helen Toner Peter Eckersley Ben Garfinkel Allan Dafoe Paul Scharre Thomas Zeitzoff Bobby Filar Hyrum Anderson Heather Roff Gregory C. Allen Jacob Steinhardt Carrick Flynn Seán Ó SJ Beard Haydn Belfield Sebastian Farquhar Clare Lyle Rebecca Crootof Owain Evans Michael Page Joanna Bryson Roman Yampolskiy Dario Amodei, *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Technical report. University of Oxford 2018, https://maliciousaireport.com/.
14. *Framework for improving critical infrastructure cybersecurity*, version 1.1, NIST 2018, https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP. 04162018.pdf.
15. Qi, Jiaxing; Shaohan Huang, Zhongzhi Luan, Shu Yang, Carol Fung, Hailong Yang, Depei Qian, Jing Shang, Zhiwen Xiao & Zhihui Wu, *Loggpt: Exploring chatgpt for log-based anomaly detection*. In: 2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), IEEE, 2023.
16. Solaiman, Irene; Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, *Release strategies and the social impacts of language models*. arXiv preprint arXiv:1908.09203, 2019.
17. Sychugov, Alexey & Mikhail Grekov, *Automated penetration testing based on adversarial inverse reinforcement learning*. In: 2024 International Russian Smart Industry Conference (SmartIndustryCon), IEEE, 2024.
18. Van der Maaten, Laurens & Geoffrey Hinton, *Visualizing data using t-sn*e. Journal of machine learning research 9(11), 2008.
19. Vaswani, Ashish; Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin, *Attention is all you need*. Advances in neural information processing systems 30. 2017.
20. Veprytska, Olena & Vyacheslav Kharchenko, *Ai powered attacks against ai powered protection: classification, scenarios and risk analysis*. In: 2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), IEEE, 2022.
21. VP, Samesh Enathe; R Dheepthi*, Llm-enhanced deepfake detection: Dense cnn and multi-modal fusion framework for precise multimedia authentication*. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), IEEE, 2024.