

CYBERBEZPIECZEŃSTWO SZTUCZNEJ INTELIGENCJI

Bartosz Bieganowski
Wydział Matematyki, Informatyki i Mechaniki
Uniwersytet Warszawski, ul. Banacha 2, 02-097 Warszawa
Wydział Matematyki i Informatyki,
Uniwersytet Mikołaja Kopernika, ul. Chopina 12/18, 87-100 Toruń
bartoszb@mimuw.edu.pl, <https://www.mimuw.edu.pl/~bartoszb/>

Abstract. *This paper analyzes the dual role of artificial intelligence (AI) in cybersecurity within the educational context, presenting it as both a defensive tool and a significant threat. The article discusses the practical applications of AI in protecting school systems and student privacy, while also highlighting new risks such as AI-driven cheating, deepfakes, and disinformation. The author concludes by stressing the urgent need for a systemic update of the school curriculum and teacher training to equip students with critical digital competence.*

1. Wstęp

Rozwój technologii informacyjno-komunikacyjnych (ICT) w ostatnich dekadach doprowadził do fundamentalnych przemian w sposobie funkcjonowania społeczeństw, gospodarek i państw. Jednym z najważniejszych i najbardziej przełomowych osiągnięć tej rewolucji technologicznej jest sztuczna inteligencja (ang. Artificial Intelligence, AI), która w coraz większym stopniu staje się integralnym komponentem cyfrowych systemów – od aplikacji mobilnych, przez systemy predykcyjne w medycynie, po narzędzia wykorzystywane w cyberbezpieczeństwie. Jednocześnie jednak, rozwój AI rodzi nowe wyzwania, zwłaszcza w kontekście ochrony informacji, prywatności oraz bezpieczeństwa systemów i sieci. Przedmiotem niniejszej pracy jest analiza wzajemnych relacji pomiędzy cyberbezpieczeństwem a sztuczną inteligencją – zarówno z perspektywy jej wykorzystania jako narzędzia obrony, jak i zagrożenia.

Zanim przejdziemy do pogłębionej analizy tych relacji, konieczne jest uporządkowanie kluczowych pojęć. Cyberbezpieczeństwo, choć często używane intuicyjnie, nie posiada jednej powszechnie akceptowanej definicji. W literaturze przedmiotu spotykamy zarówno podejścia wąskie – koncentrujące się na technicznej ochronie systemów przed atakami – jak i szerokie, uwzględniające aspekty prawne, społeczne i polityczne [2]. Zgodnie z definicją przyjętą przez Międzynarodową Unię Telekomunikacyjną (ITU),

cyberbezpieczeństwo to zbiór narzędzi, polityk, koncepcji bezpieczeństwa, wytycznych, praktyk zarządzania ryzykiem, działań szkoleniowych, najlepszych praktyk, zapewnień i technologii, które mogą być stosowane w celu ochrony zasobów informacyjnych w cyberprzestrzeni (ITU-T X.1205, <https://www.itu.int/rec/T-REC-X.1205>). Tym samym, cyberbezpieczeństwo obejmuje zarówno środki techniczne (np. szyfrowanie, zapory sieciowe), jak i organizacyjne (np. polityki bezpieczeństwa, świadomość użytkowników).

Analogiczne trudności napotykamy przy próbie definicji sztucznej inteligencji. Już klasyczne podejście Russella i Norviga [7] wskazuje, że istnieją cztery główne nurty definiowania AI: systemy myślące jak ludzie, działające jak ludzie, myślące racjonalnie oraz działające racjonalnie. W praktyce, współczesna AI to głównie zestaw algorytmów uczenia maszynowego (machine learning, ML), które potrafią rozpoznawać wzorce w danych, uczyć się na ich podstawie i podejmować decyzje lub przewidywania bez jednoznacznego zaprogramowania reguł [4]. Choć pojęcie AI bywa często używane w sposób marketingowy, warto podkreślić, że na gruncie naukowym i technicznym mówimy przede wszystkim o systemach wyspecjalizowanych, a nie o ogólnej inteligencji maszyn (AGI), która pozostaje w sferze spekulacji.

W ostatnich latach coraz więcej uwagi poświęca się zastosowaniu sztucznej inteligencji w obszarze cyberbezpieczeństwa. AI może pełnić tutaj szereg funkcji: od detekcji anomalii w ruchu sieciowym, przez identyfikację podatności w kodzie, po wspomaganie analizy zagrożeń typu zero-day. Przykładem może być wykorzystanie algorytmów uczenia głębokiego w systemach wykrywania włamań (IDS), jak opisano m.in. w pracy Sommera i Paxsona [11], czy nowsze rozwiązania oparte na analizie behawioralnej użytkowników (UEBA). AI wspiera także użytkowników indywidualnych w ochronie prywatności – umożliwia np. półautomatyczne wykrywanie śladów obecności w sieci, monitorowanie publikacji w mediach społecznościowych oraz podejmowanie prób usuwania danych z repozytoriów i wyszukiwarek (np. narzędzia klasy *OSINT AI* czy *AI reputation management*).

Nie można jednak pominąć ciemnej strony tego postępu. Sztuczna inteligencja staje się także narzędziem ofensywnym, wspierającym nowe formy cyberataków. Przykłady obejmują m.in. generowanie złośliwego kodu dostosowanego do konkretnego środowiska, użycie modeli językowych w atakach phishingowych i social engineering, a także automatyzację kampanii dezinformacyjnych z użyciem botów i tzw. deepfake'ów [3]. AI może również pomagać w omijaniu detekcji przez tradycyjne systemy zabezpieczeń poprzez szybkie rekonfigurowanie wektorów ataku w odpowiedzi na działania obronne. W szczególnie niebezpiecznych scenariuszach mówimy o tzw. *adversarial AI* – sytuacjach, w których algorytmy są celowo wprowadzane w błąd przez odpowiednio spreparowane dane wejściowe.

W tym kontekście należy podkreślić uniwersalną prawidłowość: każda nowa technologia, niezależnie od swojego potencjału, może być użyta zarówno w celach obronnych, jak i ofensywnych. Historia cyberbezpieczeństwa dostarcza wielu przykładów: od

kryptografii (służącej zarówno do zabezpieczania komunikacji, jak i do ukrywania działań przestępczych), przez systemy rozproszone (użyteczne w sieciach CDN, ale też w botnetach), po blockchain (używany do zabezpieczania transakcji, ale też do ukrywania środków z cyberataków). Sztuczna inteligencja nie jest wyjątkiem – i nie powinna być postrzegana ani jako panaceum na problemy bezpieczeństwa, ani jako źródło wszelkiego zła. Jak trafnie zauważa [5], "każde potężne narzędzie może zostać wykorzystane do budowania lub niszczenia – wszystko zależy od rąk, które je kontrolują".

W kontekście edukacyjnym zagadnienie cyberbezpieczeństwa sztucznej inteligencji nabiera szczególnego znaczenia. Szkoła jako instytucja kształtująca kompetencje społeczne i cyfrowe uczniów ma obowiązek nie tylko uczyć obsługi narzędzi cyfrowych, ale także kształtować świadome i odpowiedzialne postawy ich użytkowników. Uczniowie – często intuicyjnie korzystający z zaawansowanych systemów AI, np. filtrów obrazu, tłumaczy, asystentów głosowych czy generatorów tekstów – nie zawsze są świadomi ryzyka związanego z prywatnością, przechowywaniem danych czy możliwościami manipulacji informacjami. Równocześnie nauczyciele, którzy pełnią rolę przewodników po cyfrowym świecie, sami potrzebują wsparcia systemowego i szkoleniowego, aby nadążyć za postępem technologicznym i móc weryfikować źródła, narzędzia i intencje działań AI w procesie dydaktycznym. Coraz częściej pojawiają się pytania o dopuszczalność używania sztucznej inteligencji w pracach domowych, sprawdzianach czy komunikacji z uczniami, a także o etyczne i prawne aspekty jej wykorzystywania w edukacji. Edukacja z zakresu cyberbezpieczeństwa i AI – prowadzona w sposób systemowy, interdyscyplinarny i zorientowany na krytyczne myślenie – staje się więc jednym z kluczowych elementów przygotowania młodego pokolenia do funkcjonowania we współczesnym społeczeństwie informacyjnym.

2. Szkoła jako środowisko funkcjonowania AI i cyberbezpieczeństwa

Współczesna szkoła znajduje się w centrum rewolucji cyfrowej, która zmienia nie tylko sposób przekazywania wiedzy, ale i relacje między nauczycielem a uczniem oraz wymagania wobec kompetencji obywatelskich. Komputery, tablice interaktywne, urządzenia mobilne, a coraz częściej także aplikacje z elementami sztucznej inteligencji stają się codziennymi narzędziami dydaktycznymi. Jednocześnie uczniowie, jako przedstawiciele pokolenia cyfrowego, nierzadko przewyższają swoich nauczycieli w intuicyjnym poruszaniu się po świecie technologii, co wprowadza nową dynamikę w proces nauczania.

Cyfryzacja szkoły nie ogranicza się jednak do sprzętu czy oprogramowania. Obejmuje także zmiany programowe i systemowe, w tym nowe podejścia do nauki informatyki i edukacji medialnej. W tym kontekście coraz większe znaczenie zyskuje wiedza o

sztucznej inteligencji i cyberbezpieczeństwie jako element kompetencji kluczowych XXI wieku.

Technologie cyfrowe, w tym systemy AI, stwarzają szerokie możliwości w zakresie personalizacji nauczania, wspomagania uczniów ze specjalnymi potrzebami edukacyjnymi, automatyzacji oceniania czy analizy postępów. Wprowadzenie narzędzi takich jak inteligentne korepetytory, platformy e-learningowe z mechanizmami adaptacyjnymi czy chatboty edukacyjne pozwala na bardziej zindywidualizowane podejście do ucznia.

Z drugiej strony, brak odpowiednich regulacji i przygotowania pedagogicznego może prowadzić do nadużyć, uzależnienia od technologii, spadku zdolności krytycznego myślenia, a także zagrożeń dla prywatności i integralności danych uczniowskich. W szczególności niekontrolowane wykorzystanie AI w edukacji może prowadzić do sytuacji, w której decyzje podejmowane są przez algorytmy o niejawnej strukturze, niezrozumiałe i niekoniecznie etycznie.

W kontekście rosnącej roli technologii w edukacji, nauczyciel nie jest już wyłącznie przekazicielem wiedzy, lecz staje się przewodnikiem po świecie informacji i kompetencji cyfrowych. Szczególną rolę pełni tu nauczyciel informatyki, który powinien nie tylko uczyć podstaw programowania czy obsługi narzędzi, ale także wprowadzać uczniów w zagadnienia etycznego korzystania z AI, ochrony danych osobowych, bezpieczeństwa sieciowego oraz zagrożeń związanych z dezinformacją.

Równie istotna jest rola wychowawcza. Nauczyciele innych przedmiotów, a zwłaszcza wychowawcy klas, muszą mieć świadomość zagrożeń płynących z cyfrowej obecności uczniów: od cyberprzemocy, przez uzależnienia ekranowe, po manipulację algorytmiczną. Edukacja w zakresie bezpiecznego korzystania z AI i technologii cyfrowych powinna być prowadzona interdyscyplinarnie, w oparciu o współpracę nauczycieli, specjalistów IT i psychologów szkolnych.

3. Sztuczna inteligencja w służbie cyberbezpieczeństwa w edukacji

3.1. Narzędzia AI wspomagające bezpieczeństwo informacyjne (techniczne) w szkole

W kontekście szkolnym AI może być wykorzystana jako narzędzie ochrony infrastruktury cyfrowej oraz danych uczniów i nauczycieli przed zagrożeniami online. Przykładowe zastosowania obejmują:

- systemy filtrujące treści (np. Google SafeSearch, Lightspeed Systems), które wykorzystują AI do klasyfikacji i blokowania niebezpiecznych lub niepożądanych stron internetowych,
- inteligentne firewalle i systemy IDS/IPS stosowane w sieciach szkolnych, które z pomocą AI analizują ruch sieciowy i wykrywają podejrzane aktywności,

- systemy wykrywające próby phishingu lub spoofingu adresów e-mail w korespondencji szkolnej,
- automatyczne audyty polityk bezpieczeństwa danych w środowiskach edukacyjnych typu Google Workspace for Education czy Microsoft 365.

Przykładowy prompt dla AI w systemie zarządzania infrastrukturą IT szkoły:

„Przeanalizuj ruch sieciowy z ostatnich 24 godzin pod kątem prób dostępu do niedozwolonych usług zewnętrznych.”

Odpowiedź AI: *„Wykryto 17 prób połączenia z zewnętrznymi serwerami proxy z adresów IP przypisanych do sali komputerowej 21. Możliwe obejście ograniczeń filtrowania.”*

Zgodnie z analizą Cisco Talos (2023), zastosowanie uczenia maszynowego w systemach IDS znacząco skraca czas reakcji na nowe wektory ataku, co ma bezpośrednie przełożenie na ochronę danych uczniów i nauczycieli. Z kolei analiza akademicka [12] wykazała, że ML znacznie zwiększa dokładność IDS w porównaniu z tradycyjnymi metodami

3.2. AI a prywatność: wykrywanie danych osobowych ucznia w sieci

Sztuczna inteligencja może także służyć do ochrony reputacji cyfrowej uczniów i nauczycieli. Narzędzia typu *AI-powered OSINT* (*Open Source Intelligence*) pozwalają na wyszukiwanie danych osobowych, numerów telefonu, zdjęć lub filmów zawierających wizerunek ucznia, umieszczonych w internecie bez zgody. Systemy takie jak *PimEyes* czy *Clearview AI* pozwalają na rozpoznawanie twarzy w sieci, choć ich stosowanie budzi kontrowersje etyczne.

W przypadku wykrycia naruszenia prywatności, AI może wspomagać proces zgłaszania i usuwania danych z serwisów internetowych.

Prompt dla AI: *"Znalazłem swoje zdjęcie zrobione w szkole na nieznanym blogu. Jak je usunąć?"*

Odpowiedź AI: *"Najpierw spróbuj ustalić, kto jest administratorem strony. Mogę ci pomóc przygotować maila z żądaniem usunięcia zdjęcia na podstawie RODO. Chcesz go napisać teraz?"*

Narzędzia OSINT wspierane AI mogą być wykorzystywane nie tylko przez służby, ale także przez szkoły i rodziców do monitorowania obecności cyfrowej dzieci.

3.3. Systemy wspomagania nauczyciela: analiza plagiatów, kontrola źródeł, monitoring aktywności

Wykorzystanie AI w pracy nauczyciela może obejmować przede wszystkim:

- systemy antyplagiatowe (np. Turnitin, Copyleaks AI Detector),
- analizatory źródeł stosowanych przez uczniów w pracach pisemnych,

- detektory treści wygenerowanych przez AI (np. GPTZero, Winston AI),
- systemy LMS z analizą aktywności uczniowskiej, np. Moodle z wtyczkami AI,
- narzędzia do automatycznego generowania zadań i sprawdzianów, umożliwiające tworzenie materiałów dydaktycznych dopasowanych do poziomu i potrzeb uczniów.

Prompt nauczyciela: *"Przygotuj zestaw 10 pytań zamkniętych i otwartych do sprawdzianu z podstaw programowania w Pythonie dla klasy 7."*

Odpowiedź AI: *"Oto przykładowy zestaw: 1. Co oznacza instrukcja if? ... 10. Napisz funkcję, która przyjmuje listę liczb i zwraca sumę parzystych elementów."*

Prompt nauczyciela: *"Sprawdź, czy ten tekst ucznia mógł być wygenerowany przez ChatGPT. Oto fragment..."*

Odpowiedź AI: *"Prawdopodobieństwo generacji przez model GPT-3.5 szacowane na 91%. Zaleca się rozmowę z uczniem."*

Zastosowanie tych narzędzi może nie tylko wspierać proces nauczania, ale też przeciwdziałać oszustwom i rozwijać kompetencje weryfikacji źródeł. Kluczowe jest jednak zachowanie równowagi i poszanowanie autonomii ucznia [10].

4. Zagrożenia związane z wykorzystaniem AI w edukacji

Szybki rozwój narzędzi generatywnej AI, takich jak *ChatGPT*, *Gemini* czy *Claude*, stwarza uczniom możliwość uzyskiwania gotowych odpowiedzi, wypracowań czy kodu w czasie rzeczywistym. Uczniowie wykorzystują te technologie do obchodzenia samodzielnego myślenia i procesu uczenia się. Kwestie etyczne stosowania AI bardzo szeroko są omówione w [1]. Można spodziewać się narastającego problemu polegającego na tym, że uczniowie szkół będą wykorzystywali wyłączenia AI w przygotowywaniu zadań domowych bez konsultacji z nauczycielem.

Przykładowe formy nadużyć:

- generowanie wypracowań i prac zaliczeniowych,
- automatyczne rozwiązywanie zadań testowych (np. przez aplikacje jak Socratic lub ChatGPT z kamerą),
- wykorzystywanie AI do generowania kodu lub rozwiązań matematycznych bez zrozumienia.

Problem pogłębia się w sytuacji, gdy nauczyciele nie mają narzędzi ani kompetencji, aby odróżnić tekst pisany przez człowieka od wygenerowanego automatycznie. Choć istnieją narzędzia detekcji (np. *GPTZero*, *Copyleaks AI Detector*), badania [9, 13] pokazują, że ich skuteczność może być niska przy dobrze sformułowanych promptach przez uczniów.

Technologia generowania obrazów i wideo (np. za pomocą *DeepFaceLab*, *Face-Swap* lub narzędzi *API Stability AI*) umożliwia tworzenie realistycznych, ale fałszywych materiałów z udziałem uczniów lub nauczycieli. Tego typu treści mogą być wykorzystywane do:

- cyberprzemocy (np. wklejenie twarzy ucznia do obrazu kompromitującego),
- tworzenia fałszywych nagrań audio lub wideo nauczyciela (np. symulacja wypowiedzi),
- szantażu lub rozpowszechniania kompromitujących fałszywych informacji.

Jak wskazuje raport ENISA (2023), technologia deepfake staje się coraz bardziej dostępna, a zagrożenie dezinformacją w szkołach będzie w najbliższych latach rosnąć. Wymaga to wyposażenia nauczycieli i uczniów w narzędzia do weryfikacji treści medialnych (*fact-checking*, rozpoznawanie zmanipulowanych obrazów).

Dużym problemem w edukacji staje się tzw. *prompt injection*, czyli formułowanie poleceń do AI w taki sposób, aby ominąć ograniczenia bezpieczeństwa. Uczniowie uczą się tworzyć promptów umożliwiających uzyskanie „zakazanych” informacji. Zjawisko to jest znane i szeroko analizowane w literaturze: np. [6] opisują, jak *prompt injection* może być wykorzystywane do uzyskiwania informacji sprzecznych z politykami bezpieczeństwa, nawet w systemach zaprojektowanych do ochrony przed tym (np. *ChatGPT*, *Bard*).

Wiele narzędzi AI używanych w edukacji (np. platformy LMS z funkcjami predykcyjnymi, generatory testów, aplikacje wspomagające uczenie się) opiera się na zbieraniu i analizie danych uczniowskich. W praktyce oznacza to:

- monitorowanie aktywności online,
- zapisywanie historii odpowiedzi,
- profilowanie postępów edukacyjnych.

Jak zwracają uwagę [8], wiele takich systemów działa w tzw. "czarnej skrzynce" – nie znamy dokładnych algorytmów podejmowania decyzji ani lokalizacji przetwarzania danych. Z punktu widzenia RODO oraz etyki edukacyjnej budzi to poważne wątpliwości dotyczące zgody, przejrzystości i kontroli nad danymi.

Z tego powodu, szkoły powinny posiadać jasno zdefiniowaną politykę korzystania z AI, uwzględniającą minimalizację przetwarzania danych, informowanie uczniów i rodziców oraz audytowalność działań systemów.

5. Podstawa programowa a rzeczywistość szkolna

W obecnie obowiązującej podstawie programowej z informatyki brak jest jednoznacznych odniesień do sztucznej inteligencji oraz zagadnień związanych z cyberbezpieczeństwem w kontekście AI. Choć uczniowie uczą się podstaw programowania, przetwarzania informacji i krytycznej analizy treści, nie są systemowo przygotowywani do rozumienia zagrożeń wynikających z używania systemów AI czy analizy ich

wiarygodności. Wskazuje to na potrzebę aktualizacji podstawy programowej w kierunku włączenia tematów takich jak generatywne AI, dezinformacja, prywatność algorytmiczna i analiza działania modeli.

Niniejsza praca została poprawiona stylistycznie przez generatywną sztuczną inteligencję (*ChatGPT, model 4o*).

Literatura

1. Akgun S., Greenhow C. *Artificial intelligence in education: Addressing ethical challenges in K-12 settings*. AI Ethics 2, 431–440 (2022). <https://doi.org/10.1007/s43681-021-00096-7>
2. Bada M., Sasse A. M., Nurse J. R. C., *Cyber Security Awareness Campaigns: Why do they fail to change behaviour?*, arXiv:1901.02672
3. Brundage M. et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, arXiv:1802.07228v2
4. Goodfellow I., Bengio Y., Courville, A., *Deep Learning*, Wydawnictwo Naukowe PWN (e-book)
5. Harari Y. N., *Nexus. A Brief History of Information Networks from the Stone Age to AI*, Sandycove 2024
6. Liu X., Yu Z., Zhang Y., Zhang N., Xiao C., *Automatic and Universal Prompt Injection Attacks against Large Language Models*, arXiv:2403.04957
7. Russell S. J., Norvig P., *Artificial Intelligence: A Modern Approach (4th edition)*, Pearson 2020
8. Regan, P.M., Jesse, J. *Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking*. Ethics Inf Technol 21, 167–179 (2019). <https://doi.org/10.1007/s10676-018-9492-2>
9. Sadasivan V. S., Kumar A., Balasubramanian S., Wang W., Feizi S., *Can AI-Generated Text be Reliably Detected?*, arXiv:2303.11156
10. Selwyn N., *Should Robots Replace Teachers? AI and the Future of Education*, Polity Press, 2019
11. Sommer R., Paxson V., *Outside the Closed World: On Using Machine Learning for Network Intrusion Detection*, IEEE Symposium on Security and Privacy, Oakland, CA, USA, 2010, pp. 305–316, doi: 10.1109/SP.2010.25
12. Tripathy S. S., Behera B., *Performance evaluation of Machine learning algorithms for Intrusion Detection System*, arXiv:2310.00594
13. Weber-Wulff D., et al., *Testing of Detection Tools for AI-Generated Text*, arXiv:2306.15666v2