CyberPOLICY
NASK

Cyberbezpieczeństwo AI

AI w cyberbezpieczeństwie

CyberPOLICY
NASK

Cyberbezpieczeństwo AI.
AI w cyberbezpieczeństwie.

Szanowni Państwo,

Z ogromną radością przedstawiam Państwu publikację *Cyberbezpieczeństwo AI. AI w cyberbezpieczeństwie*, która została przygotowana w Dziale Strategii i Rozwoju Bezpieczeństwa Cyberprzestrzeni NASK-PIB przy udziale ekspertów zewnętrznych. Jestem głęboko przekonany, że niniejsze opracowanie dostarczy Państwu wszechstronnej wiedzy na temat szeroko rozumianego bezpieczeństwa sztucznej inteligencji.

Sztuczna inteligencja w ostatnich latach stała się fundamentem rewolucji technologicznej. Znajduje zastosowanie zarówno w dziedzinach takich jak zdrowie czy transport, ale również w wielu innych obszarach naszego codziennego życia. Jej dynamiczny rozwój przynosi nie tylko coraz więcej innowacyjnych rozwiązań i usprawnień, ale stwarza zupełnie nowe wyzwania i zagrożenia, także w dziedzinie cyberbezpieczeństwa.

Wraz z upowszechnianiem stosowania sztucznej inteligencji rośnie ryzyko nadużywania tej technologii przez cyberprzestępców – na przykład do tworzenia fałszywych informacji, wykradania danych czy infiltrowania systemów zabezpieczeń. Wobec tego istotne staje się zagadnienie bezpieczeństwa i solidności AI, przy równoczesnym poszanowaniu praw podstawowych i zapewnieniu takich warunków jak wyjaśnialność i rozliczalność tych systemów w całym cyklu ich życia.

Wierzę, że poprzez zwiększanie świadomości społecznej i rozwijanie interdyscyplinarnej współpracy możemy skutecznie przeciwdziałać zagrożeniom i korzystać z zalet, jakie niesie ze sobą sztuczna inteligencja.

Życzę Państwu miłej lektury.

Krzysztof Silicki

Dyrektor ds. Strategicznego Rozwoju
Cyberbezpieczeństwa NASK-PIB

Wstęp

Niezwykle szybko postępujący w ostatnich latach rozwój nowych technologii, w tym sztucznej inteligencji, przyniósł rewolucję w różnych dziedzinach, od medycyny przez bankowość po kwestie związane z obronnością, porządkiem publicznym czy egzekwowaniem prawa. Jednak wraz z rozwojem sztucznej inteligencji pojawiają się także nowe wyzwania, związane m.in. z jej cyberbezpieczeństwem.

Celem niniejszej pracy zbiorowej było przedstawienie wyników analizy zagadnień dotyczących cyberbezpieczeństwa sztucznej inteligencji oraz możliwych zastosowań tej technologii w cyberbezpieczeństwie. Przedstawione zostały różne aspekty i wyzwania, z jakimi mierzą się organizacje, naukowcy, czy specjaliści ds. bezpieczeństwa. Do zaprezentowania swoich punktów widzenia na tak określoną problematykę zaproszeni zostali przedstawiciele różnych instytucji i ośrodków naukowych. Ta różnorodność autorów poszczególnych artykułów pozwala na wieloaspektowe porównanie często odmiennych od siebie podejść i holistyczne spojrzenie na prezentowany problem.

Na niniejszą pracę zbiorową składają się zarówno teksty dotyczące technicznych aspektów cyberbezpieczeństwa sztucznej inteligencji, jak i artykuły zawierające analizę z obszaru *cyberpolicy*.

Publikację otwiera opracowanie pt. *Cyberbezpieczeństwo systemów wykorzystujących sztuczną inteligencję w świetle raportów ENISA*, autorstwa Krzysztofa Silickiego, Dyrektora ds. Strategicznego Rozwoju Cyberbezpieczeństwa w NASK. Formułuje on wnioski dotyczące systemów ICT, które wykorzystują technologie sztucznej inteligencji, w kontekście zabezpieczeń wynikających ze specyficznych zagrożeń dla samouczących się algorytmów.

Następnie, przedstawiciele Śląskiego Centrum Inżynierii Prawa, Technologii i Kompetencji Cyfrowych CYBER SCIENCE analizują wyzwania i zagrożenia z zakresu cyberbezpieczeństwa podczas projektowania lub wykorzystywania AI. Skupiają się oni na relacjach między wyjaśnialnym uczeniem maszynowym a cyberbezpieczeństwem, wskazując także na cyberzagrożenia rozwiązań wykorzystujących uczenie maszynowe. Przedstawiają także etyczne wyzwania pozostające w związku ze sposobami wykorzystania powodującymi zagrożenie.

Trzeci artykuł, autorstwa dr. hab. Jerzego Surmy, profesora Szkoły Głównej Handlowej w Warszawie, przybliża ataki na systemy uczenia maszynowego. Omawia on wektory ataków, przedstawia ich taksonomię oraz szczegółowo analizuje atak na integralność. Dla klarowności tekstu te zagadnienia zostały przedstawione na przykładzie systemów uczących się pod nadzorem realizujących zadanie klasyfikacji.

Kolejny tekst, zatytułowany *Weryfikacja wiarygodności systemów w erze uczenia maszynowego*, przygotowany przez Mateusza Krzysztonia z Centrum Badań i Rozwoju NASK-PIB, porusza kwestie związane z techniczną weryfikacją wiarygodności systemów. Autor skupia się na potencjalnych źródłach braku technicznej wiarygodności, problemie sprawiedliwości, interpretowalności i wytłumaczalności, a także bezpieczeństwie tego rodzaju systemów.

Następne opracowanie dotyczy zagadnienia antagonistycznego uczenia maszynowego i zostało przygotowane przez Mateusza Bursiaka, również reprezentującego NASK-PIB. Oprócz przeglądu klasyfikacji uczenia maszynowego i kwestii związanych z uczeniem nadzorowanym, w tekście znalazł się również przegląd wybranych metod ataków na algorytmy uczenia maszynowego. Niezwykle interesujący jest również zaprezentowany przykład ataku na algorytmy uczenia maszynowego nadzorowanego.

W dalszej kolejności, dr Jarosław Greser z Politechniki Warszawskiej dokonał analizy cyberbezpieczeństwa medycznej AI. Sklasyfikował i opisał cyberzagrożenia dla tego rodzaju systemów. W swoim artykule skupia się on również na kwestiach regulacji medycznej sztucznej inteligencji.

Tę część publikacji zamykają Krzysztof Zieliński, dyrektor Departamentu Cyberbezpieczeństwa Urzędu Komisji Nadzoru Finansowego oraz Agata Ślusarek z CSIRT KNF, którzy poruszają kwestie wykorzystania sztucznej inteligencji przez cyberprzestępców w atakach na klientów rynku finansowego, ze szczególnym uwzględnieniem technologii *deepfake*. W bardzo ciekawy sposób zaprezentowali oni narzędzia, jakimi posługują się sprawcy w celu manipulacji ofiarą oraz przykłady wykorzystania tego typu aplikacji w konkretnych atakach.

Wraz z artykułem *Strategia Sztucznej Inteligencji dla NATO*, autorstwa Piotra Słowińskiego z Działu Strategii i Rozwoju Bezpieczeństwa Cyberprzestrzeni NASK-PIB następuje płynne przejście do części poświęconej

cyberpolicy. Autor prezentuje politykę, jaką NATO przyjęło wobec sztucznej inteligencji, skupiając się przede wszystkim na zasadach odpowiedzialnego wykorzystania AI oraz zadaniach i celach Rady Eksperckiej NATO ds. Danych i Sztucznej Inteligencji.

Następnie Robert Kroplewski w swoim tekście *Odporność AI dla odpornej wspólnoty* przybliża kwestie konwergencji cyberbezpieczeństwa i odporności w obszarze sztucznej inteligencji. Analizując dokumenty strategiczne i akty prawne organizacji międzynarodowych, przedstawia problem sprawowania pieczy nad godną zaufania SI oraz systemowej odporności w tym zakresie. Skupia się również na kwestiach zarządzania w kontekście odporności i zasobów systemów sztucznej inteligencji.

Dziesiąty artykuł, autorstwa Moniki Stachoń z Działu Strategii i Rozwoju Bezpieczeństwa Cyberprzestrzeni NASK-PIB, prezentuje unijne podejście do sztucznej inteligencji. Na podstawie trzech dokumentów strategicznych, przyjętych przez Komisję Europejską, przedstawia wnioski dotyczące zasad i głównych wytycznych, jakimi kieruje się UE w obszarze sztucznej inteligencji. Uzupełnieniem jest analiza wytycznych HLEG w sprawie etyki oraz polityki i inwestycji w SI.

Pozostając na gruncie prawa europejskiego, w dalszej części Aleksandra Szczęsna z NASK-PIB przybliża główne założenia projektu AI Act, czyli rozporządzenia ustanawiającego zharmonizowane przepisy dotyczące sztucznej inteligencji. Skupia się na wymogach dla systemów AI, zakazanych praktykach, obowiązkach w zakresie przejrzystości czy systemach stwarzających ryzyko. Punktem wyjścia dla Autorki tekstu pozostaje podejście ogólne Rady dotyczące wniosku w sprawie tego aktu, zatwierdzone 6 grudnia 2022 roku.

Publikację zamyka wywiad przeprowadzony przez Emilię Zalewską-Czajczyńską z NASK-PIB z dr Gabrielą Bar, partnerką zarządzającą Szostek_Bar i Partnerzy Kancelarii Prawnej oraz członkinią organizacji *Women in AI*. Tematem wywiadu są wyzwania dla cyberbezpieczeństwa sztucznej inteligencji w kontekście AI Act, w tym trudności regulatorów w nadążaniu za technologią czy próby stworzenia jednolitej definicji sztucznej inteligencji na gruncie unijnym.

Publikacja, którą przekazujemy na Państwa ręce, ma na celu przedstawienie różnorodnych aspektów i wyzwań związanych z cyberbezpieczeństwem

sztucznej inteligencji. Wierzymy, że stanowi cenne źródło wiedzy i prezentuje różne perspektywy, w tym ekspertów w dziedzinie cyberbezpieczeństwa, analityków i prawników. Mamy nadzieję, że dostarczy ona odpowiednich informacji dla czytelników zainteresowanych cyberbezpieczeństwem w erze AI. Zachęcamy do zapoznania się i życzymy owocnej lektury.

Dział Strategii i Rozwoju Bezpieczeństwa
Cyberprzestrzeni NASK-PIB

Cyberbezpieczeństwo systemów wykorzystujących sztuczną inteligencję w świetle raportów ENISA

Krzysztof Silicki

NASK-PIB | Dyrektor ds. Strategicznego Rozwoju Cyberbezpieczeństwa

Wprowadzenie

W styczniu 2017 roku ponad stu prominentnych naukowców i praktyków zajmujących się problematyką sztucznej inteligencji spotkało się na mającej doniosłe znaczenie konferencji Asilomar Conference on Beneficial AI, zorganizowanej przez Future of Life Institute. Jej efektem było sformułowanie dwudziestu trzech tzw. zasad z Asilomar (Asilomar AI Principles) [1]. Zgodnie z nimi, rozwój sztucznej inteligencji powinien opierać się na określonych pryncypiach zapewniających, że powstające rozwiązania w tej dziedzinie będą dobroczynne dla ludzkości. Według zasady szóstej, systemy AI powinny być bezpieczne w całym cyklu ich życia, a tam, gdzie ma to zastosowanie i jest wykonalne, powinno być to możliwe do zweryfikowania. Do roku 2023 pod deklaracją z Asilomar podpisało się ponad 5000 osób z całego świata, w tym badaczy sztucznej inteligencji i robotyki, prominentnych przedstawicieli różnych dziedzin nauki, szefów dużych i małych firm technologicznych, członków ciał standaryzacyjnych, polityków i reprezentantów innych zawodów.

W marcu 2023 roku ten sam Future of Life Institute wystosował list otwarty [2] do wszystkich laboratoriów pracujących nad sztuczną inteligencją, aby wstrzymać na sześć miesięcy prace nad wszelkimi systemami, które mogą być mocniejsze niż algorytm GPT-4. Model ten z jednej strony stał się synonimem potężnych możliwości sztucznej inteligencji już teraz, a nie w przyszłości, a z drugiej – przykładem konkretnego zagrożenia dla spełniania zasad z Asilomar. List ten podpisało do czerwca 2023 roku ponad 30 000 osób, w tym m.in. Elon Musk, który w 2015 roku zainwestował środki w powstanie laboratorium OpenAI – twórcę czatu opartego o model GPT-4.

Zachodzi pytanie, czy list ten należy traktować jako przejaw zbiorowej histerii. W mojej ocenie – niekoniecznie. Na przykład w bazie incydentów publikowanej przez AIID (AI Incident Database) [3] w czerwcu 2023 roku można zapoznać się z ponad 2700 rzeczywistymi zgłoszonymi incydentami różnego typu, związanymi z użyciem technologii AI, które dotyczą ponad 1100 podmiotów lub konkretnych osób[1].

Relacje pomiędzy cyberbezpieczeństwem a sztuczną inteligencją

Europejska Agencja ds. Cyberbezpieczeństwa (ENISA) od 2005 roku publikuje raporty na temat bezpieczeństwa sieci teleinformatycznych i informacji w różnych dziedzinach w kontekście bieżących i przyszłych zagrożeń i ich wpływu na funkcjonowanie gospodarki. Od kilku lat pojawiają się publikacje związane z cyberbezpieczeństwem sztucznej inteligencji. Są to między innymi: „Looking into the crystal ball: A report on emerging technologies and security challenges" (styczeń 2018 r.) [4], „Artificial Intelligence Cybersecurity Challenges. Threat Landscape for Artificial Intelligence" (grudzień 2020 r.) [5], „Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving" (luty 2021) [6], „Securing Machine Learning Algorithms" (grudzień 2021 r.) [7], „Cybersecurity of AI and Standardisation, ENISA Foresight Cybersecurity Threats for 2030" (marzec 2023 r.) [8].

1 Stan na 5 czerwca 2023 roku.

W pierwszym z nich [4], sztuczna inteligencja i uczenie maszynowe (AI/ML) zostały przedstawione jako przyszłościowe narzędzia w dziedzinie zapewnienia wysokiego poziomu cyberbezpieczeństwa, w szczególności obszaru *threat intelligence* oraz detekcji ataków czy zarządzania cyberbezpieczeństwem. Jednak z drugiej strony podkreśla się, że systemy wykorzystujące algorytmy AI/ML same staną się celem ataków, a to z kolei otwiera zgoła nowe przestrzenie do manipulacji i tworzenia metod cyberataków.

W raporcie drugim [5], zdefiniowane zostały trzy wymiary relacji pomiędzy sztuczną inteligencją a cyberbezpieczeństwem:

1. Cyberbezpieczeństwo sztucznej inteligencji (ang. *cybersecurity for AI*).

2. Sztuczna inteligencja w służbie cyberbezpieczeństwa (ang. *AI to support cybersecurity*).

3. Szkodliwe wykorzystanie sztucznej inteligencji (ang. *malicious use of AI*).

Z kolei inny raport [7] koncentruje się na bezpieczeństwie samych algorytmów wykorzystywanych w systemach sztucznej inteligencji. W ramach tego zagadnienia, określono w nim taksonomię tych systemów i zaproponowano obszerny katalog środków bezpieczeństwa, które można – a nawet powinno się – zastosować. Te środki czy metody cyberbezpieczeństwa dzielą się w istocie na dwie główne kategorie: specyficzne dla algorytmów czy modeli AI/ML oraz standardowe środki techniczne i organizacyjne stosowane generalnie w obszarze bezpieczeństwa teleinformatycznego.

Widzimy więc, że obszar związków pomiędzy technologiami czy systemami wykorzystującymi sztuczną inteligencję a zagadnieniem cyberbezpieczeństwa jest bardzo szeroki i wielowymiarowy. Wymusza to, z jednej strony, całościowe patrzenie na zagadnienia technologiczne – ze świadomością, że w kontekście sztucznej inteligencji mamy także do czynienia z zagadnieniami etycznymi, prawnymi czy regulacyjnymi. Z drugiej jednak skłania to do prowadzenia wyspecjalizowanych analiz, badań i rozwoju w każdej z trzech dziedzin, które wymienia wspomniany wyżej raport ENISA.

Zagrożenia cyberbezpieczeństwa technologii stosujących sztuczną inteligencję

Opisując wyzwania, jakie stoją przed nami w kwestii zapewniania odpowiedniego poziomu bezpieczeństwa systemom wykorzystującym algorytmy sztucznej inteligencji, należy zastanowić się nad tym, co tak naprawdę powinniśmy chronić i przed jakimi zagrożeniami. Jeśli zgodzimy się, że warto zastosować całościowe podejście technologiczne do bezpieczeństwa, możemy mówić o kilku obszarach czy kategoriach zagrożeń:

- skierowanych na same algorytmy czy modele sztucznej inteligencji,
- w procesie zarządzania i przetwarzania danych, którymi posługują się algorytmy,
- dla procesu trenowania algorytmów zbiorami danymi,
- dla implementacji programowej modeli oraz systemu sztucznej inteligencji,
- wynikających z istniejących podatności infrastruktury teleinformatycznej (zwirtualizowanej, chmurowej lub fizycznej), na której działa system AI.

W taksonomii przedstawionej w raporcie ENISA [5] opisano kilkadziesiąt typów możliwych ataków w podziale na kategorie, takie jak:

- przeprowadzanie ataków wykorzystujących niedojrzałość i podatności technologiczne;
- nieintencjonalne powodowanie szkód czy błędów;
- naruszenia prawa, umów, regulaminów;
- błędy lub awarie systemów AI;
- przechwytywanie danych, niedozwolone ujawnianie danych lub modeli;
- ataki fizyczne;
- utrata łączności;
- katastrofy lub zjawiska środowiskowe.

Rozpatrywanie tematyki cyberbezpieczeństwa we wszystkich powyższych kategoriach jest wyrazem całościowego podejścia do obszaru zagrożeń wszystkich technologii teleinformatycznych i systemy AI nie są tu wyjątkiem pod względem reguł bezpieczeństwa. Warto też przypomnieć, że zgodnie z wyrażonym w dyrektywie NIS [9] i kolejnych aktach prawnych przyjmowanych w UE od roku 2016 podejściem, cyberincydenty to nie

tylko klasyczne ataki na infrastrukturę czy usługi cyfrowe, takie jak DDoS, phishing, ransomware, łamanie haseł czy wycieki danych osobowych. Są to wszelkie zagrożenia, które powodują (lub mogą spowodować) przerwę lub nieprawidłowe działanie usług cyfrowych. Dlatego awarie systemów, błędy w konfiguracji, utrata łączności z aplikacją chmurową czy ataki fizyczne znajdują się na palecie zagrożeń, również w kontekście systemów sztucznej inteligencji.

W tym kontekście warto zauważyć, że mówiąc o zagrożeniach czy atakach na systemy sztucznej inteligencji będziemy mieli do czynienia z ogromną grupą istniejących albo przewidywanych zagrożeń, mniej lub bardziej znanych z dotychczasowej praktyki osób zajmujących się bezpieczeństwem teleinformatycznym oraz nowy obszar zagrożeń specyficznych dla algorytmów, modeli i danych wykorzystywanych przez AI.

Trudno jest wymienić wszystkie możliwe typy zagrożeń czy ataków, tak więc posłużmy się pojedynczymi przykładami w każdej z powyższych kategorii:

- ataki wykorzystujące niedojrzałość i podatności technologiczne. Przykładem tego rodzaju ataków są szeroko opisywane w literaturze ataki antagonistyczne (*adversarial examples*). Takie ataki często przedstawia się jako wprowadzanie niewielkich zaburzeń do danych (np. obrazów), które – niezauważalne dla ludzkiego oka – powodują nieprawidłowe (nieskuteczne) działanie modeli opartych na sztucznej inteligencji [10].

  Innym przykładem w tej kategorii mogą być ataki lub słabości procesu etykietowania danych w systemach uczenia nadzorowanego. Manipulowanie etykietami (modyfikowanie etykiet w procesie uczenia, losowe wprowadzanie perturbacji), szczególnie przy częściowej lub pełnej wiedzy na temat docelowego modelu przez atakującego, spowoduje nieprawidłowe wyniki działania algorytmu;

- nieintencjonalne powodowanie błędnego działania modeli. Przykładem takiego działania może być tzw. bias (czyli pewien brak neutralności danych) na jakich dany model był trenowany. Często przywoływany przypadek dla zilustrowania tego typu zagrożenia polega np. na wykorzystywaniu do uczenia rozpoznawania obrazów twarzy człowieka w większości zdjęć osób jednej płci lub jednego

koloru skóry lub tylko w określonym przedziale wieku, co może powodować niewłaściwe rozpoznawanie twarzy w przekroju całej populacji ludzkiej;

- naruszenia prawa, umów, regulaminów. W tej kategorii często przywoływanym zagrożeniem jest naruszenie prywatności danych w czasie przechowywania lub przetwarzania.

  Jako inny przykład w tej kategorii można przywołać możliwość ujawnienia danych osobowych bezpośrednio albo poprzez korelację różnych danych ze zbiorów używanych przez algorytmy. Zagrożeniom tego typu „sprzyjają" braki w procedurach weryfikacji źródeł danych czy niedoskonałe mechanizmy pseudonimizacji danych. Ilustracją zagrożeń tej kategorii mogą być także naruszenia SLA (*service level agreement*) ze strony kontrahentów, którzy oferują konkretne usługi (dostarczanie danych, przestrzeni wirtualnej do obliczeń, modeli itp.), aby dany system AI działał prawidłowo;

- błędy lub awarie systemów AI. Tego typu zdarzenia mają i będą miały miejsce. W przypadku systemów AI nieprawidłowe działanie lub przerwy w działaniu np. systemów chmurowych, systemów dostawców danych, operatorów *AI as a service*, sieci telekomunikacyjnych czy własnej infrastruktury oraz braki w dokumentowaniu parametrów modelu czy w ogólności dokumentacji systemu AI mogą spowodować nie tylko przerwy w działaniu systemów (szczególnie groźne w przypadku systemów monitorujących krytyczne procesy), ale także utratę zaufania do wyników danego systemu AI;

- przechwytywanie danych, niedozwolone ujawnianie danych lub modeli. Oprócz klasycznych zagrożeń nieuprawnionego ujawniania danych, w tej kategorii jest także mowa o celowym ujawnianiu lub atakowaniu wewnętrznych parametrów modelu przez nieuprawnione osoby [11].

Trzy ostatnie kategorie, znane z praktyki zarządzania czy operowania systemami teleinformatycznymi nie wymagają raczej szczególnego rozwijania i nie są jakimiś specyficznymi zagrożeniami dla sztucznej inteligencji, ale dla całościowego obrazu zagrożeń, wymagają odnotowania:

- ataki fizyczne;

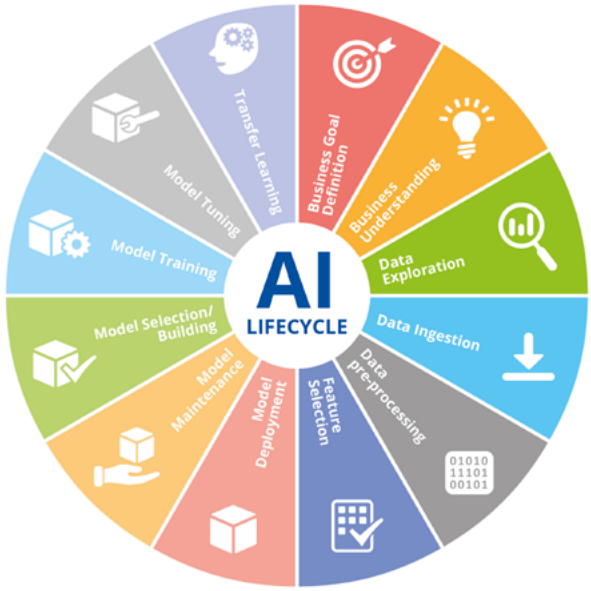
- utrata łączności;
- katastrofy lub zjawiska środowiskowe.

W praktyce, istotnym działaniem w celu przeprowadzenia analizy zagrożeń i potem zarządzania ryzykiem jest umiejętność modelowania tychże zagrożeń dla danego przypadku użycia systemu sztucznej inteligencji. Oznacza to proces prawidłowej identyfikacji zagrożeń czy ryzyk, nadanie im właściwej wagi czy priorytetu, a w efekcie – podjęcie właściwej decyzji co do wdrożenia odpowiednich technicznych, organizacyjnych czy umownych środków bezpieczeństwa. W tym celu niezbędne jest posłużenie się odpowiednimi metodykami modelowania zagrożeń. Takie metodyki, w odniesieniu do AI, powinny wziąć pod lupę zarówno tradycyjne cechy bezpieczeństwa teleinformatycznego, takie jak: poufność, integralność, dostępność czy, dodatkowo, autentyczność lub niezaprzeczalność, ale także cechy, które będą specyficzne dla systemów sztucznej inteligencji. Wśród nich należy wyróżnić odporność modeli na ataki czy niepożądane zmiany, wyjaśnialność sposobu, w jaki wynik działania algorytmu został uzyskany, skuteczność działania w odniesieniu do spodziewanych efektów, audytowalność (możliwość badania cyberbezpieczeństwa systemu opartego na AI w każdej fazie jego cyklu życia) czy w ogóle cechy odnoszące się do zaufania, jakie możemy mieć do danego eksploatowanego systemu.

Fazy cyklu życia sztucznej inteligencji oraz zasoby podlegające ochronie

Innym elementem całościowego podejścia do cyberbezpieczeństwa sztucznej inteligencji jest, godne poparcia, dążenie do rozpatrywania tego zagadnienia w kontekście całego cyklu życia modeli AI – w odróżnieniu od koncentrowania się jedynie na jednej z faz, np. pozyskiwania danych, trenowania czy wdrożenia modelu sztucznej inteligencji.

Fazy cyklu ilustruje poniższy rysunek.



RYS. 1. Fazy cyklu życia sztucznej inteligencji [5]

Można zatem, w ślad za powyższą propozycją ENISA, wyodrębnić szereg faz związanych z powstawaniem i funkcjonowaniem modeli sztucznej inteligencji, takich jak:

- definiowanie celu biznesowego, jaki chcemy osiągnąć przy wykorzystaniu AI;
- pozyskiwanie danych z różnych źródeł;
- weryfikowanie i walidacja danych;
- wstępne przetwarzanie danych, integracja danych pochodzących z różnych źródeł, interpolacja, pseudonimizacja i inne;
- wybór wymiarów danych najbardziej znaczących w kontekście danego modelu;
- wybór i budowanie najodpowiedniejszego typu modelu sztucznej inteligencji dla danego zastosowania biznesowego;
- trenowanie modelu;
- strojenie modelu;
- implementacja danego modelu do postaci konkretnego oprogramowania zainstalowanego na konkretnej infrastrukturze sprzętowej i połączonego z danymi produkcyjnymi;
- stałe monitorowanie i konserwacja modelu w obliczu zmian koniecznych do wprowadzenia w czasie pracy systemu sztucznej inteligencji;

- analiza skuteczności zastosowanego modelu i stopnia realizacji celów biznesowych;
- wycofanie systemu z eksploatacji.

Ważne jest także, aby uświadomić sobie, jakie konkretne zasoby czy elementy systemu sztucznej inteligencji powinny podlegać ochronie. Tym razem, ponownie, wyzwaniem jest wyobrażenie sobie (zmapowanie) wszystkich koniecznych kategorii zasobów, które powinny podlegać ochronie – tych klasycznych, znanych z dotychczasowej praktyki ochrony systemów teleinformatycznych oraz tych specyficznych dla technologii wykorzystujących AI. Próbę takiego podejścia znajdujemy również we wspomnianym raporcie unijnej Agencji ds. Cyberbezpieczeństwa, co zostało zobrazowane na poniższym rysunku.

RYS. 2. Taksonomia aktywów wykorzystywanych przez AI [5]

Zakończenie

Wnioski płynące z poszczególnych raportów ENISA wyraźnie wskazują, iż konwencjonalne zabezpieczenia systemów ICT, które wykorzystują technologie sztucznej inteligencji, muszą zostać wzbogacone o cały katalog zabezpieczeń wynikających ze specyficznych zagrożeń dla samouczących się algorytmów. Im bardziej całościowo podejdziemy do tego zagadnienia, tym lepiej dla ostatecznego efektu poziomu cyberbezpieczeństwa systemów AI. Warto też dodać, że na świecie powstają normy [12] definiujące standardowe działania, jakimi powinniśmy się kierować przy określaniu ryzyka systemowego, badania bezpieczeństwa i stosowania środków zapobiegawczych. Normy te również generalnie potwierdzają, że wszystkie dotychczas przyjęte dokumenty (np. seria norm ISO 27000 [13]) pozostają w dalszym ciągu bazą dla określania systemów teleinformatycznych wykorzystujących sztuczną inteligencję, ale definiują także część nową, specyficzną dla tejże technologii. Można przewidywać, iż w przyszłości podejście oparte o normy przyniesie dodatkowy efekt w postaci możliwości certyfikowania produktów, usług czy procesów pod kątem cyberbezpieczeństwa. Ważny krok w kierunku zdefiniowania ram certyfikacji cyberbezpieczeństwa przyniósł w Unii Europejskiej Akt o cyberbezpieczeństwie [14] z 2019 roku.

Niemniej jednak, ważne dla uzyskania pozytywnego efektu są inicjatywy legislacyjne związane z AI – w szczególności z zapewnieniem warunków do bezpieczeństwa i zaufania do rozwijanych technologii sztucznej inteligencji. Europa jest, podobnie jak w przypadku zagadnienia ochrony danych osobowych, niewątpliwym liderem w tej dziedzinie.

Dokumentem, który podejmuje tę tematykę jest proponowany Akt w sprawie sztucznej inteligencji [15]. Jednym z głównych celów przyjęcia tego rozporządzenia jest właśnie zapewnienie bezpieczeństwa sztucznej inteligencji. Znajdziemy w nim szereg motywów oraz postanowień, które wspierają tezy przedstawione wyżej, np. to, iż cyberataki na systemy sztucznej inteligencji mogą polegać na wrogim wykorzystaniu określonych zasobów, takich jak zbiory danych treningowych lub trenowane modele (np. wprowadzanie do modelu szkodliwych danych) lub wykorzystaniu konkretnych podatności systemu sztucznej inteligencji bądź w infrastrukturze ICT, na której opiera się dany system. Żeby zatem zapewnić poziom cyberbezpieczeństwa adekwatny do przeanalizowanego ryzyka, dostawcy systemów sztucznej inteligencji powinni wdrożyć

proporcjonalne środki bezpieczeństwa, uwzględniając również to, na jakiej infrastrukturze ICT funkcjonuje dany system.

Bibliografia

[1] Future of Life Institute (2017). *AI Principles* [Online]. Dostęp: https://futureoflife.org/open-letter/ai-principles/

[2] Future of Life Institute (22 marca 2023). *Pause Giant AI Experiments: An Open Letter* [Online]. Dostęp: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[3] *AI Incident Database* [Online]. Dostęp: https://incidentdatabase.ai/

[4] ENISA (31 stycznia 2018). „Looking into the crystal ball: A report on emerging technologies and security challenges" [Online]. Dostęp: https://www.enisa.europa.eu/publications/looking-into-the-crystal-ball

[5] ENISA (15 grudnia 2020). „Artificial Intelligence Cybersecurity Challenges. Threat landscape for Artificial Intelligence" [Online]. Dostęp: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[6] ENISA (11 lutego 2021). „Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving" [Online]. Dostęp: https://www.enisa.europa.eu/publications/enisa-jrc-cybersecurity-challenges-in-the-uptake-of-artificial-intelligence-in-autonomous-driving/

[7] ENISA (14 grudnia 20210. „Securing Machine Learning Algorithms" [Online]. Dostęp: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.

[8] ENISA (29 marca 2023). „ENISA Foresight Cybersecurity Threats for 2030" [Online]. Dostęp: https://www.enisa.europa.eu/publications/enisa-foresight-cybersecurity-threats-for-2030.

[9] Dyrektywa Parlamentu Europejskiego i Rady (UE) 2016/1148 z dnia 6 lipca 2016 r. w sprawie środków na rzecz wysokiego wspólnego poziomu bezpieczeństwa sieci i systemów informatycznych na terytorium Unii.

[10] I. Goodfellow, J. Shlens, C. Szegedy (20 grudnia 2014). „Explaining and Harnessing Adversalial Examples" [Online]. Dostęp: https://arxiv.org/abs/1412.6572.

[11] E. Spafford (29 listopada 1988. „The Internet Worm Program: An Analysis", Purdue Technical Report CSD-TR-823, IN 47907-2004 [Online]. Dostęp: https://spaf.cerias.purdue.edu/tech-reps/823.pdf

[12] ENISA (14 marca 2023). „Cybersecurity of AI and Standardisation" [Online]. Dostęp: https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation.

[13] Information security, cybersecurity and privacy protection, ISO/IEC 27000:2018.

[14] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2019/881 z dnia 17 kwietnia 2019 r. w sprawie ENISA (Agencji Unii Europejskiej ds. Cyberbezpieczeństwa) oraz certyfikacji cyberbezpieczeństwa w zakresie technologii informacyjno-komunikacyjnych oraz uchylenia rozporządzenia (UE) nr 526/2013 (akt o cyberbezpieczeństwie).

[15] Wniosek – Rozporządzenie Parlamentu Europejskiego i Rady z 21 kwietnia 2021 ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji (Akt w sprawie sztucznej inteligencji) i zmieniające niektóre akty ustawodawcze Unii.

Wyzwania i zagrożenia z zakresu cyberbezpieczeństwa podczas projektowania lub wykorzystywania AI

dr hab. Dariusz Szostek, prof. ucz.
dr hab. inż. Paweł Kasprowski, prof. ucz.
dr hab. Jan Kozak, prof. ucz.
dr inż. Adrian Kapczyński
dr n. pr. inż. Rafał Prabucki

Śląskie Centrum Inżynierii Prawa,
Technologii i Kompetencji Cyfrowych CYBER SCIENCE

Wprowadzenie

W ostatnim czasie przeżywamy niezwykle intensywny rozwój systemów informatycznych oraz algorytmów wspierających funkcjonowanie człowieka w podejmowaniu decyzji bądź podejmujących je za niego – sztucznej inteligencji [1]. Równocześnie mamy do czynienia z niespotykaną dotychczas ilością incydentów, w tym mających znamiona krytycznych w zakresie cyberbezpieczeństwa. Co 11 sekund przeprowadzany jest atak z użyciem oprogramowania typu ransomware, koszty ataków cybernetycznych w 2021 roku szacowane są na 5,5 biliona EUR i raczej będą rosły [2]. Zarówno do ataków, jak i do obrony przed nimi coraz częściej wykorzystywane są narzędzia oparte o AI. Także nowoczesne, tworzone w chwili obecnej w Unii Europejskiej prawo obejmuje zarówno

możliwość wykorzystania AI w zakresie cyberbezpieczeństwa, ale także, co stanowi istotne *novum*, konieczność zapewnienia rozliczalności oraz *compliance* w odniesieniu do AI tzw. wysokiego ryzyka. Warto jednakże zwrócić uwagę, iż ograniczenia wynikające z wniosku w sprawie rozporządzenia AI (ang. *AI Act*) w założeniu projektodawcy nie będą miały zastosowania do systemów AI opracowywanych lub wykorzystywanych wyłącznie do celów wojskowych oraz bezpieczeństwa narodowego, jak również do działalności badawczo – rozwojowej, jeżeli ta działalność nie prowadzi do wprowadzenia systemu na rynek lub do użytku. Inaczej rzecz ujmując, UE wyraźnie dopuszcza systemy AI, mało tego – poza kontrolą i wymogami przewidywanymi we wniosku wspomnianych przepisów. Równocześnie zakazuje się wprowadzania do obrotu systemów AI, które mogłyby w sposób istotny i bezpośredni zagrozić infrastrukturze krytycznej, czy być wykorzystane w celu ataku terrorystycznego. Unia Europejska w zakresie wykorzystania AI na gruncie cyberbezpieczeństwa przyjmuje zarówno akty prawne, jak i strategie oraz dokumenty dotyczące standaryzacji.

Niniejszy rozdział został przygotowany przez interdyscyplinarny zespół naukowców funkcjonujących w ramach Śląskiego Centrum Inżynierii Prawa, Technologii i Kompetencji Cyfrowych CYBER SCIENCE i jest próbą transdycyplinarnego naszkicowania wyzwań i zagrożeń z zakresu cyberbezpieczeństwa podczas projektowania lub wykorzystywania AI.

Podstawowe zagadnienia w zakresie projektowania i wykorzystania AI

Zacznijmy od tego czym właściwie jest AI, czyli sztuczna inteligencja. Termin ten stał się w ostatnich czasach swego rodzaju *buzzword* i używany jest często do określania urządzeń czy programów, w których nie ma mowy o inteligencji (np. „inteligentne żelazka"). Zaznaczyć też przy tym należy, że wciąż trwa dyskusja na temat tego, czy w ogóle możemy używać pojęcia „inteligencja" w stosunku do maszyn i systemów komputerowych. Pomijając jednak te wątpliwości, uznać można, że system mający cechy inteligencji charakteryzuje się możliwością uczenia – nabywania wiedzy w trakcie działania. Tak więc typowe użycie systemu AI rozpoczyna się od jego uczenia, czyli tak zwanego trenowania. Dopiero wytrenowany system jest w stanie rozwiązywać prawdziwe problemy.

Istnieje kilka rodzajów systemów AI. Najpopularniejsze z nich to systemy nadzorowane, które uczą się na przykładach – tak zwanych zbiorach treningowych. Taki system można na przykład nauczyć rozpoznawania sygnatur znanych wirusów komputerowych. Tego typu systemy mogą też rozpoznawać znanych przestępców na filmach z monitoringu.

Inny rodzaj systemów AI to systemy nienadzorowane. Tego typu systemy nie są instruowane co do tego, jak należy zaklasyfikować dane wejściowe, one same uczą się rozróżniać dane. Przykładowe zastosowania to wykrywanie anomalii w ruchu sieciowym lub wykrywanie intruzów czy nietypowych sytuacji w danych z monitoringu.

Kolejnym bardzo szybko się rozwijającym się działem AI jest *reinforcement learning* (RL), co na język polski można przetłumaczyć jako „uczenie ze wzmacnianiem". W tego typu zastosowaniach AI nie otrzymuje zbioru danych treningowych, ale dostęp do środowiska, w którym uczy się podejmować kolejne decyzje tak, żeby zmaksymalizować końcowy wynik. Typowe zastosowania systemów RL to aplikacje grające w różnego rodzaju gry (np. szachy czy grę Go). W przypadku cyberbezpieczeństwa można przykładowo wyobrazić sobie agenta, który na różne sposoby próbuje skompromitować zabezpieczenia danego mu jako środowisko systemu komputerowego.

Podstawową zaletą systemów sztucznej inteligencji jest fakt, że jeśli dostarczymy im odpowiednich danych treningowych lub środowiska, w którym mogą trenować, są one w stanie nauczyć się realizacji bardzo skomplikowanych celów – często takich, które przekraczają możliwości pojedynczych ludzi. Nie trzeba też dodawać, że systemy komputerowe są w stanie podejmować decyzje znacznie szybciej niż ludzie.

Wyjaśnialne uczenie maszynowe a cyberbezpieczeństwo

Obecnie, w celu zwiększenia swojej wydajności i efektywności, wiele firm i instytucji wykorzystuje systemy sztucznej inteligencji. Jednakże ze wzrostem zastosowań technologii AI pojawiają się nowe zagrożenia, które wymagają ochrony sieci i systemów informatycznych przed atakami. Dlatego też coraz więcej organizacji zaczyna interesować się wyjaśnialnym uczeniem maszynowym, jako sposobem na lepsze zabezpieczenie swoich danych i systemów przed nieautoryzowanym dostępem i atakami.

Eksperci ds. bezpieczeństwa muszą stawić czoła jednemu z głównych wyzwań, jakim jest zrozumienie złożonych algorytmów uczenia maszynowego, które często działają w sposób niewyjaśniony. Podejście związane z wyjaśnialną sztuczną inteligencją (ang. *explainable artificial intelligence*, XAI), a właściwie wyjaśnialnym uczeniem maszynowym odnosi się do procesu tworzenia algorytmów uczenia maszynowego, które umożliwiają zrozumienie oraz interpretację decyzji podejmowanych przez systemy sztucznej inteligencji. Aby skutecznie wykorzystać podejście XAI w dziedzinie cyberbezpieczeństwa, eksperci muszą poświęcić czas na dokładne poznanie działania systemów AI, a następnie opracować strategię wykorzystania XAI. Dzięki temu eksperci ds. bezpieczeństwa sieci mogą szczegółowo monitorować swoje systemy oraz podjąć odpowiednie kroki w przypadku wykrycia nieprawidłowości, co przyczynia się do zwiększenia poziomu bezpieczeństwa sieci oraz systemów informatycznych. Będą oni w stanie zrozumieć, jakie czynniki wpłynęły na decyzję systemu AI i w jaki sposób doszło do podjęcia danej decyzji, co pozwoli na bardziej skuteczną reakcję na potencjalne zagrożenia [3].

Tym samym wyjaśnialne uczenie maszynowe może pomóc w wykrywaniu i identyfikowaniu zagrożeń w sieciach informatycznych. Dzięki temu organizacje będą w stanie zwiększyć efektywność swoich systemów obrony przed cyberatakami, co przyczyni się do minimalizacji ryzyka naruszenia bezpieczeństwa danych oraz innych poważnych zagrożeń dla ich działalności.

Trzeba zwrócić uwagę, że wykorzystanie wyjaśnialnego uczenia maszynowego w dziedzinie cyberbezpieczeństwa niesie ze sobą pewne zagrożenia. Istnieje ryzyko ataku na system z wykorzystaniem wiedzy o działaniu wytrenowanego algorytmu przez osoby nieuprawnione. Ten problem został dokładniej poruszony w pracy [4].

Wyjaśnialne uczenie maszynowe jest ważną dziedziną, która szybko się rozwija i może odegrać kluczową rolę w zapewnieniu bezpieczeństwa sieci i systemów informatycznych. W związku z tym, coraz więcej organizacji zwraca uwagę na XAI jako narzędzie pozwalające zwiększyć bezpieczeństwo swoich danych i zasobów sieciowych. Niemniej jednak wciąż potrzeba dalszych badań i eksperymentów, aby dokładniej zrozumieć, jak skutecznie wykorzystać wyjaśnialne uczenie maszynowe w dziedzinie cyberbezpieczeństwa.

Cyberzagrożenia rozwiązań wykorzystujących uczenie maszynowe

Uczenie maszynowe to gałąź sztucznej inteligencji, która zajmuje się zdolnością do nauki na podstawie danych, co pozwala na dokonywanie predykcji i podejmowanie decyzji na podstawie wzorców, które są wykrywane w danych. Uczenie maszynowe może być stosowane w wielu dziedzinach, takich jak rozpoznawanie obrazów, analiza języka naturalnego, analiza danych medycznych i wiele innych [5]. Wachlarz potencjalnych rozwiązań wykorzystujących uczenie maszynowe jest bardzo duży – można wyjść od szerokiego spektrum zagadnień, np. „rozpoznawanie obrazów", a zakończyć na szczegółowym zagadnieniu dotyczącym „identyfikacji biometrycznej na podstawie obrazu naczyń krwionośnych palca".

Warto podać kilka przykładów rozwiązań, w których kluczową rolę pełni uczenie maszynowe. Pierwszym z nich mogą być systemy rekomendacyjne, w których algorytmy uczenia maszynowego analizują preferencje użytkowników i na ich podstawie proponują kolejne filmy lub produkty, które mogą ich zainteresować. Dalej, potencjał uczenia maszynowego jest widoczny w realizacji systemów rozpoznawania mowy i przetwarzania języka naturalnego stanowiących element interfejsu głosowego pozwalającego na zrozumienie tego, o czym mówi użytkownik. Wreszcie, systemy kontroli fizycznej mogą wykorzystywać cechy anatomii twarzy człowieka starającego się uzyskać dostęp do budynku, strefy czy pomieszczenia. Literatura przedmiotu [6] wskazuje na to, jak szerokie zastosowanie ma uczenie maszynowe i jak wiele dziedzin może z niego czerpać w celu osiągnięcia lepszych wyników lub usprawnienia procesów.

Warto zwrócić uwagę nie tylko na potencjał uczenia maszynowego, w tym na korzyści, jakie oferuje, ale także na inne aspekty, w tym koszty czy zagrożenia. Odnośnie do kosztów, jest to temat na odrębne opracowanie, ale warto jest wymienić choćby te podstawowe, do których należą: koszt infrastruktury teleinformatycznej, szkolenia modelu uczenia maszynowego czy też utrzymania rozwiązania wykorzystującego uczenie maszynowe.

Przejdźmy teraz do najważniejszych z punktu widzenia niniejszego podrozdziału zagadnień związanych z uczeniem maszynowym, a mianowicie – do cyberzagrożeń. Warto podkreślić, iż skupimy uwagę wyłącznie na zagrożeniach w sferze cyberbezpieczeństwa, mając świadomość panoramicznego charakteru kategorii nadrzędnej, która traktuje o zagrożeniach

rozwiązań wykorzystujących uczenie maszynowe. Za ilustrację zagrożeń należących do kategorii nadrzędnej niech posłuży przykład wspomnianych wcześniej systemów rekomendacyjnych, które oparte są na danych pochodzących od użytkowników, z czym wiąże się zagrożenie naruszenia prywatności.

Sięgając do Słownika Języka Polskiego [7] można dowiedzieć się, że cyberzagrożenie ma związek ze środkami komunikacji elektronicznej. Dalsza kwerenda bibliograficzna [8] pozwala na zbudowanie bardziej szczegółowego obrazu tego zjawiska, jako wewnętrznego lub zewnętrznego źródła negatywnego wpływu na pracę systemu, w tym na przetwarzanie, przesyłanie, przechowywanie lub prezentowanie informacji. Przykłady zagrożeń w cyberprzestrzeni obejmują wirusy komputerowe, programy szpiegujące, ataki wymuszające okup, ataki DDoS (ang. *Distributed Denial of Service*), czy ataki na systemy krytycznej infrastruktury. Cyberzagrożenia są wszechobecne i złożone, co wymaga stosowania skutecznych narzędzi ochrony i działań prewencyjnych. Upowszechniana jest konieczność stosowania zasad cyberhigieny, wśród których wymienia się regularną aktualizację oprogramowania, stosowanie silnych haseł czy korzystanie z oprogramowania antywirusowego i zapór sieciowych.

W dalszych rozważaniach skoncentrujemy się na zagadnieniach dotyczących cyberzagrożeń nie o charakterze ogólnym, lecz szczegółowym, tj. dotyczących rozwiązań wykorzystujących uczenie maszynowe.

Oto lista dziesięciu wybranych zagrożeń dla rozwiązań wykorzystujących uczenie maszynowe [9]:

1. ATAK NA MODEL UCZENIA MASZYNOWEGO. Atakujący mogą próbować wprowadzić fałszywe dane lub zmienić już istniejące, aby zmanipulować wyniki uczenia maszynowego.

2. ATAK ADWERSARYJNY. Atakujący mogą próbować wprowadzić niewielkie zmiany w danych treningowych, aby wpłynąć na wyniki generowane przez model uczenia maszynowego.

3. ATAK ZATRUWAJĄCY. Atakujący mogą próbować wprowadzić fałszywe dane do systemu uczenia maszynowego, aby zmanipulować wyniki modelu.

4. ATAK POPRZEZ TYLNIE WEJŚCIE. Atakujący mogą próbować wykorzystać ukryte funkcjonalności lub punkty wejścia, aby uzyskać nieautoryzowany dostęp do systemu uczenia maszynowego.

5. ATAK NA MECHANIZM OCHRONY DANYCH. Atakujący mogą próbować uzyskać nieautoryzowany dostęp do danych wykorzystywanych przez modele uczenia maszynowego.

6. ATAK TYPU POWTÓRZENIOWEGO. Atakujący mogą próbować przechwycić dane przesyłane między systemem uczenia maszynowego a innymi systemami, a następnie użyć ich do celów niepożądanych.

7. ATAK TYPU ODMOWA ŚWIADCZENIA USŁUGI. Atakujący mogą próbować zablokować lub przeciążyć systemy uczenia maszynowego, aby uniemożliwić ich działanie.

8. KRADZIEŻ MODELU UCZENIA. Atakujący mogą próbować skopiować modele uczenia maszynowego, aby wykorzystać je w niepożądany sposób.

9. ZŁOŚLIWE OPROGRAMOWANIE. Atakujący mogą próbować zainfekować systemy uczenia maszynowego złośliwym oprogramowaniem.

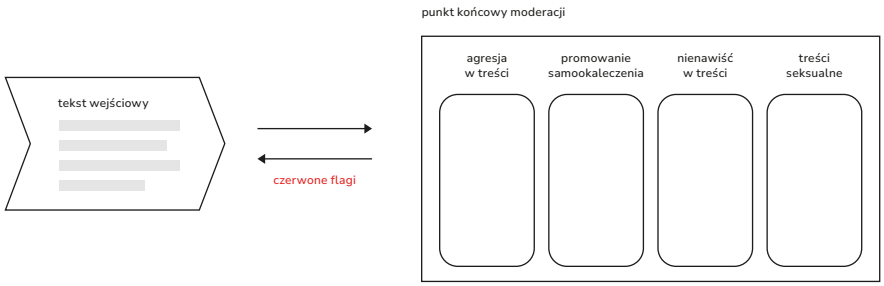
10. NARUSZENIE PRYWATNOŚCI DANYCH. Modele uczenia maszynowego mogą zawierać dane osobowe lub poufne informacje, a atakujący mogą próbować uzyskać do nich dostęp w celach niepożądanych.

Przedstawiona lista zagrożeń stanowi zestaw zagadnień, które należy traktować jako punkty startowe dla dalszych, pogłębionych rozważań, rzecz jasna ukierunkowanych nie tylko na ich uszczegółowienie, ale także na eksplorację komplementarnego obszaru – cyberzabezpieczeń. Bezpieczeństwo rozwiązań wykorzystujących uczenie maszynowe stanowi istotne wyzwanie ze względu na wiele potencjalnych zagrożeń, co wymaga starannego projektowania rozwiązań, ich monitorowania, zabezpieczania, a także stałego aktualizowania wiedzy dotyczącej cyberzagrożeń.

Etyczne wyzwania pozostające w związku ze sposobami wykorzystania powodującymi zagrożenie

Wiele emocji w zakresie systemów AI budzi kwestia ich wykorzystania w taki sposób, że zagrożone stają się wartości, które w codziennym życiu są w pewien sposób chronione, np. prawem. Największe wzburzenie wywołuje jedna z nich – życie ludzkie. Szczególnie modele generatywne sztucznej inteligencji zwiększają pod tym kątem zainteresowanie badaczy, np. w kwestii treści, których następstwem byłaby samobójcza śmierć użytkownika po zapoznaniu się z wygenerowaną treścią [10]. W raporcie Europolu [12] zwrócono uwagę, że istnieją ogólne modele moderacji, które powinny nie dopuszczać pewnych treści (rys. 1). Jak jednak zauważają autorzy, powszechne są już procesy obejścia ograniczeń (ang. *jailbreak*), z których najbardziej popularny był DAN (*Do Anything Now*), który jest podpowiedzią zaprojektowaną specjalnie w celu obejścia zabezpieczeń OpenAI [12].

Procesy niewłaściwego projektowania i wykorzystania AI mogą też zagrozić kwestii prywatności. Raport Instytutu Allana Turinga [11] wskazuje, że rozwój systemów AI będzie często wiązał się też z wykorzystaniem danych osobowych. Konsekwencje takiego działania mogą być daleko idące. Nawet jeżeli prywatność sama w sobie nie jest wartością szczególnie chronioną, dane o pewnej osobie przetworzone przez system AI mogą zagrozić jej prawu do realizowania swoich celów i planów życiowych w sposób wolny od niechcianych wpływów. Przykładem kontrowersyjnego użycia systemu AI w kwestii prywatności był system odpowiadający za rozpoznawanie twarzy w Radio City Music Hall na Manhattanie. Kazus dotyczył prawniczki Kelly Conlon, która nie została wpuszczona na koncert świąteczny w grudniu 2022 roku. System AI rozpoznał ją jako pracownika kancelarii prawnej, która reprezentowała stronę pozywającą operatora sali koncertowej. Adwokatka musiała czekać na swoją córkę przed salą koncertową, bo uznano ją za zagrożenie [13].



RYS. 1. Ogólny schemat moderacji w modelu generatywnej sztucznej inteligencji [12]

Ochrona poszczególnych wartości, czy to od strony prawnej, czy też etycznej, jest uzależniona od kręgu kulturowego i tego, co dla pewnych grup społecznych jest ważne, a co postrzega się jako zagrożenie. Niemniej mimo różnic pomiędzy kulturą zachodnią a Chinami, badacze z Stanford University w raporcie z 2023 roku [14] nawet tam zidentyfikowali zainteresowanie głównie badaniami z zakresu etyki, koncentrującymi się na problemie prywatności oraz równości.

Olbrzymie znaczenie w kreowaniu właściwego projektowania i wykorzystania systemów AI będą odgrywały regulacje prawne. W UE przykładem takiego działania jest projekt unijnego Aktu w sprawie sztucznej inteligencji. Regulacja ta nakierowana jest na skonstruowanie ram prawnych dla osób projektujących i wykorzystujących AI w taki sposób, aby ich działania przyniosły korzyści dla społeczeństwa i przemysłu krajów tworzących UE [15].

Podsumowanie

Konkludując należy zauważyć, że istnieje kilka kwestii, które autorzy uważają za istotne:

- Wzrost cyberzagrożeń powoduje, że odpowiednie projektowanie i wykorzystanie systemów AI jest istotne, co nie znaczy, że same systemy AI nie mogą być wykorzystane do budowy struktur cyberbezpieczeństwa.

- Pojęcie AI może oznaczać różne rozwiązania. Przystępując do bezpiecznego projektowania i wykorzystania AI musimy pamiętać, aby sprawdzić z czym właściwie mamy do czynienia.

- W zakresie projektowania i wykorzystania systemów AI w cyberbezpieczeństwie warto zwrócić uwagę na wyjaśnialne uczenie maszynowe, które zapewnia wysoką przejrzystość w procesie podejmowania decyzji przez AI.

- Mimo szerokiego zakresu zastosowania uczenia maszynowego, może ono być też źródłem cyberzagrożeń.

- Uczenie maszynowe, zarówno w zakresie wyjaśnialności, jak i cyberzagrożeń, wymaga dalszych badań.

- Stosowane są różne zabezpieczenia techniczne w procesie projektowania, które mają zablokować wykorzystanie systemów AI do celów, które mogą nie być etycznymi, jakkolwiek istnieją osoby próbujące te zabezpieczenia przełamywać.

- Etyka nie wyznacza konkretnych ram projektowania i wykorzystania AI, przez co pojawiają się sytuacje, które wzbudzają wątpliwości w kwestii użycia systemów AI do pewnych celów, które pozornie w teorii wydają się słuszne.

- Stosowanie ram etycznych jest też zależne od kręgów kulturowych, w jakich system AI ma być zastosowany, niemniej coraz więcej badaczy na świecie zauważa problem związany z ochroną prywatności, niezależnie od systemu politycznego, czy hierarchii wartości w pewnej kulturze.

- Prawodawcy na całym świecie przedstawiają pierwsze regulacje, które miałyby wskazywać projektującym i wykorzystującym systemy AI jak działać, aby rozwój AI przynosił pozytywne skutki dla pewnego społeczeństwa, gdzie dane prawo ma obowiązywać.

Bibliografia

[1] „Artificial Intelligence", *Encyklopedia Britannica*, dostęp: https://www.britannica.com/ technology/artificial-intelligence.

[2] Komisja Europejska (15 września 2022). *Cyber Resilience Act – Factsheet* [Online]. Dostęp: https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act-factsheet

[3] S. Hariharan, A. Velicheti, A. S. Anagha, C. Thomas, N. Balakrishnan, „Explainable Artificial Intelligence in Cybersecurity: A Brief Review", in *4th International Conference on Security and Privacy (ISEA-ISAP)*, Dhanbad, India, 2021, pp. 1-12.

[4] N. Capuano, G. Fenza, V. Loia and C. Stanzione, „Explainable Artificial Intelligence in CyberSecurity: A Survey", *IEEE Access*, vol. 10, pp. 93575-93600, 2022.

[5] C. Janiesch, P. Zschech, & K. Heinrich, „Machine learning and deep learning," *Electron Markets* 31, pp. 685–695, 2021.

[6] I. H. Sarker, „Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Computer Science* vol. 2, pp.160, 2021.

[7] „Cyberzagrożenie", *Słownik Języka Polskiego*, dostęp: https://sjp.pl/cyberzagro%C5%BCenie.

[8] M. Marczyk, „Cyberprzestrzeń jako nowy wymiar aktywności człowieka – analiza pojęciowa," *Przegląd Teleinformatyczny* vol. 6(24), pp. 59–72, 2018.

[9] J. Surma (red.), *Hakowanie sztucznej inteligencji*. Warszawa: Wydawnictwo Naukowe PWN, 2020.

[10] M. Coeckelbergh (29 marca 2023). *Chatbots can kill* [Online]. Dostęp: https://coeckelbergh.medium.com/chatbots-can-kill-d82fde5cf6ca.

[11] D. Leslie (11 czerwca 2019). „Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector", The Alan Turing Institute [Online]. Dostęp: https://doi.org/10.5281/zenodo.3240529.

[12] Europol (17 kwietnia 2023). „ChatGPT. The impact of Large Models on Law Enforcement", Europol's Innovation Lab, Tech Watch Flash reports [Online]. Dostęp: https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement#downloads

[13] The Guardian (21 grudnia 2022). *Facial recognition bars lawyer from Girl Scout trip to Rockettes Christmas show* [Online]. Dostęp: https://www.theguardian.com/us-news/2022/dec/21/facial-recognition-bars-lawyer-rockettes-show.

[14] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. Carlos Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark and R. Perrault, „The AI Index 2023 Annual Report," Institute for Human-Centered AI, Stanford University., Stanford, CA, 2023.

[15] Wniosek – Rozporządzenie Parlamentu Europejskiego i Rady z 21 kwietnia 2021 ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji (akt w sprawie sztucznej inteligencji) i zmieniające niektóre akty ustawodawcze Unii.

Wprowadzenie do ataków na systemy uczenia maszynowego

dr hab. Jerzy Surma, prof. ucz.

Szkoła Główna Handlowa w Warszawie | Instytut Informatyki i Gospodarki Cyfrowej

Wprowadzenie

Systemy sztucznej inteligencji i w szczególności systemy uczenia maszynowego (ang. *machine learning systems*) są obecnie powszechnie stosowane w praktyce [1]. W tym kontekście zagrożenia związane z celowym atakowaniem takich systemów na każdym etapie cyklu życia stają się coraz większym wyzwaniem dla ich bezpiecznego użytkowania. Jest to szczególnie istotne zagadnienie w sytuacji znikomej obecnie świadomości tych realnych zagrożeń. Potwierdza to badanie przeprowadzone przez Shankar z zespołem [2] w firmach, które w zaawansowanym zakresie wykorzystują i rozwijają samodzielnie systemy uczące się. Badani pracownicy (szefowie zespołów programistycznych i menedżerowie odpowiedzialni za bezpieczeństwo IT) wyrażali w zdecydowanej większości opinie o futurystycznym charakterze ataków na systemy maszynowego uczenia i deklarowali brak zasobów do analizy tego typu zagrożeń. Znamienna jest opinia jednego z badanych, który stwierdził, że ataki na modele systemów uczących się to sytuacja, w której: *nie wiemy, że nie wiemy*. W tym kontekście ten artykuł ma szczególne znaczenie nie tylko dla środowiska naukowego, lecz także dla menedżerów zajmujących się rozwojem takich systemów oraz odpowiedzialnych za ryzyko operacyjne i ciągłość działania.

W pierwszej części dokonano przeglądu badań naukowych. Następnie omówiono wektory ataków na systemy uczenia maszynowego. Przedstawiono taksonomię ataków oraz szczegółowo przeanalizowano atak na integralność. Dla klarowności wywodu te zagadnienia przedstawiono dla systemów uczących się pod nadzorem realizujących zadanie klasyfikacji. W ostatniej części przedstawiono reprezentatywne prace badawcze związane z budowaniem tzw. odpornych systemów uczenia maszynowego. Niniejszy artykuł został opracowany jako skrócona i zaktualizowana wersja rozdziału „Wstęp do hakowania systemów uczących się" książki pod redakcją J. Surmy [3].

Przegląd badań naukowych

Problem intencjonalnych ataków na systemy uczenia maszynowego po raz pierwszy w sposób kompleksowy został przedstawiony w artykule Dalviego [4], w którym opisano metodę manipulowania zbiorem uczącym, aby zwiększyć błąd klasyfikacji. Dla ilustracji tego zjawiska wykorzystano „zmanipulowany" filtr antyspamowy, który klasyfikował pocztę elektroniczną ze spamem jako pocztę adekwatną do czytania. Ta tematyka była kontynuowana w pracy Barreno [5] oraz w pracy doktorskiej Nelsona [6]. W artykule [7] po raz pierwszy zaprezentowano systematyczne podejście do klasyfikacji potencjalnych ataków na systemy uczące się oraz zaproponowano teoretyczny model interakcji pomiędzy atakującym i broniącym z wykorzystaniem funkcji kosztu. Od roku 2015, w kontekście spektakularnego sukcesu zastosowań konwolucyjnych sieci neuronowych, tematyka ta staje się jednym z kluczowych obszarów badawczych. Obecnie ten obszar badań jest najczęściej określany jako antagonistyczne maszynowe uczenie się (ang. *adversarial machine learning*)[1]. Do najbardziej innowacyjnych prac badawczych w ostatnim okresie można zaliczyć badania Goodfellowa [8] w zakresie generatywnych sieci GAN (ang. *generative adversarial network*) oraz prace w zakresie tzw. ataków na czarną skrzynkę (ang. *black-box adversary attack*) z wykorzystaniem modeli głębokiego uczenia [9].

1 To pojęcie pojawia się po raz pierwszy w roku 2007 w roboczym opracowaniu pt. *Foundations of Adversarial Machine Learning* dostępnym na stronie: https://www.researchgate.net/publication/228623424_Foundations_of_Adversarial_Machine_Learning

Ten obszar badawczy rozwija się obecnie niezwykle gwałtownie i są już dostępne setki recenzowanych artykułów naukowych. Reprezentatywny przegląd najlepszych badań naukowych zawierają artykuły [10] [11] oraz książka [12]. Niezwykle istotna jest standaryzacja ataków na systemy sztucznej inteligencji podjęta przez NIST (Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations) [13] oraz konsorcjum Mitre (Adversarial Threat Landscape for Artificial-Intelligence Systems) [14].

Atakowanie systemów maszynowego uczenia

DEFINICJA I CELE ATAKUJĄCEGO

Zgodnie z klasyczną definicją Mitchella [15] system jest się w stanie uczyć z doświadczenia (ang. *experience*) E w kontekście realizacji zadań (ang. *tasks*) T i miary jakości działania (ang. *performance measure*) P, jeśli jego działanie w realizacji zadań T, mierzone z wykorzystaniem P, polepsza się wraz ze wzrostem doświadczenia E. W efekcie procesu uczenia się system jest w stanie generować poprawną odpowiedź (wyjście) dla danego obiektu na wejściu. Jeżeli w trakcie uczenia się system otrzymuje informację trenującą (etykiety w zbiorze uczącym), to takie podejście nazywane jest uczeniem nadzorowanym (ang. *supervised learning*). Przykładem tego podejścia są systemy uczenia maszynowego realizujące zadanie regresji oraz zadanie klasyfikacji.

Rozważmy system uczenia maszynowego pod nadzorem, dla którego T jest zadaniem klasyfikacji, miarą jakości działania P jest dokładność klasyfikacji (ang. *accuracy*), a doświadczenie E jest reprezentowane poprzez zbiór uczący. Intencjonalny atak na taki system uczenia maszynowego może mieć następujące cele [3]:

1. OBNIŻENIE JAKOŚCI KLASYFIKATORA (ang. *misclassification*) przez generowanie błędów fałszywie pozytywnych lub fałszywie negatywnych. Konsekwencją tego typu ataku jest spadek dokładności klasyfikacji, co implikuje obniżenie wiarygodności systemu (ang. *cofidence reduction*), a nawet, w sytuacji skrajnej, rezygnację z jego użytkowania. Wynika to m.in. z faktu, że błędne klasyfikacje

generują realne i potencjalne (np. związane z utratą reputacji) koszty będące konsekwencją błędnych decyzji albo ich braku.

2. CELOWY BŁĄD KLASYFIKACJI (ang. *targeted misclassification*) przez uzyskanie błędnej klasyfikacji dla określonych obiektów. W takiej sytuacji klasyfikator niepoprawnie klasyfikuje konkretny obiekt lub zbiór obiektów zgodnie z intencją atakującego. W takim podejściu atakujący jest zainteresowany, aby jakość klasyfikatora była na odpowiednim wysokim poziomie i tym samym, żeby wzbudzał on zaufanie użytkowników. Ten atak jest najczęściej realizowany przez tzw. tylne drzwi w klasyfikatorze (ang. *targeted backdoor attack*).

3. OGRANICZENIE DOSTĘPNOŚCI (ang. *access restriction*), czyli uzyskanie nieakceptowalnie długiego czasu reakcji systemu na dane wejściowe, a w sytuacji skrajnej, zatrzymanie działania systemu. Celem atakującego może też być ograniczenie dostępności w trakcie budowania modelu, tj. w trakcie jego uczenia, aktualizacji i testowania.

TAKSONOMIA ATAKÓW

System maszynowego uczenia jest systemem informatycznym, który podlega takim samym kryteriom oceny cyberbezpieczeństwa jak każdy system informatyczny. W tej perspektywie można wyróżnić standardowo trzy podstawowe kryteria jakości ochrony informacji [16]:

1. POUFNOŚĆ (tajność) (ang. *confidentiality*) – ochrona informacji przed nieuprawnionym dostępem.

2. INTEGRALNOŚĆ (ang. *integrity*) – zapewnienie, że składowane i przetwarzane dane są niezmienione i nie zostały wykonane na nich niedozwolone działania.

3. DOSTĘPNOŚĆ (ang. *availability*) – zapewnienie adekwatnego stopnia dostępności do danych, procesów i aplikacji dla autoryzowanych użytkowników.

Taka specyfikacja kryteriów umożliwia poprawne zarządzanie ryzykiem operacyjnym i ustalenie odpowiednich polityk bezpieczeństwa. To podejście jest zgodne ze specyfikacją amerykańskiego instytutu standardów

technicznych NIST [17], który nawiązuje do standardu oceny ryzyka bezpieczeństwa informacji [18]. W specyfikacji NIST atak jest wykonywany na konkretny cel (ang. *target*) i jego konsekwencje (ang. *consequences*) zależą od przyjętych procedur obrony (ang. *defenses*). Potencjalne konsekwencje są zgodne z wymienionymi wcześniej trzema kryteriami jakości ochrony informacji[2]. W odwołaniu do triady Poufność-Integralność-Dostępność możliwe jest zatem przedstawienie następującej taksonomii ataków na systemy uczące się [19]:

1. ATAK NA POUFNOŚĆ (ang. *confidentiality violation*) – polega na zdobyciu informacji, które dotyczą procesu uczenia, aktualizacji, testowania i użytkowania systemu. Oznacza to zdobycie informacji obejmującej: obiekt wraz z jego specyfikacją cech, zbiór etykiet klas, model klasyfikatora, algorytm wraz z jego parametrami, zbiór uczący, zbiór testujący, użyte biblioteki i środowiska programistyczne, kontekst użycia systemu: intencje, cel użycia systemu, organizacja pracy, zaangażowani pracownicy, itp. Biorąc pod uwagę ten zakres informacji, wiedzę atakującego można podzielić na trzy grupy: pełna wiedza (ang. *white box attack*), wiedza częściowa (ang. *grey box*) i brak wiedzy (ang. *black box*). W każdym wymienionym przypadku, nawet przy całkowitym braku wiedzy, możliwe są skuteczne działania atakującego. W przypadku zdobycia pełnej wiedzy mówimy o pełnej ekstrakcji modelu (ang. *extraction attack*). Natomiast przy wiedzy niepełnej lub jej braku, atakujący próbuje odtworzyć model i buduje jego substytut, bazując na założeniach, domysłach i testach. Atak na poufność, jako rodzaj rozpoznania (rekonesans) i zdobycia wiedzy, zwykle poprzedza atak na integralność lub na dostępność.

2. ATAK NA INTEGRALNOŚĆ (ang. *integrity violation*) – polega na zakłóceniu procesu uczenia, aktualizacji lub testowania zgodnie z celami atakującego omówionymi w poprzednim podrozdziale. Ze względu na ważność i złożoność tego zagadnienia, temu rodzajowi ataków poświęcono następny podrozdział.

3. ATAK NA DOSTĘPNOŚĆ (ang. *availability violation*) – polega na spowolnieniu albo zatrzymaniu pracy systemu, co utrudnia jego praktyczne

2 W specyfikacji NIST kryterium poufności obejmuje również zagadnienie prywatności (ang. *privacy*).

wykorzystanie. Ten rodzaj ataku może być wykonany w trakcie uczenia, testowania czy też przez zakłócenia procesu aktualizacji systemu. Niemniej najczęściej realizowany w fazie użytkowania (funkcjonowania) systemu. W tym kontekście możliwe jest:

A. zakłócenie lub zablokowanie procesu zbierania i przekazywania obiektów do klasyfikacji, np. przez wygenerowanie „sztucznego tłoku" obiektów na wejściu klasyfikatora (ang. *denial of service*);

B. generowanie dużej liczby błędów fałszywie pozytywnych, powodujących zaangażowanie zasobów na obsługę tych błędów (ang. *false positive overload*). To podejście wymaga ingerencji w proces uczenia lub testowania.

ATAK NA INTEGRALNOŚĆ W TRAKCIE BUDOWANIA MODELU

W ramach ataku na proces budowania systemu (uczenie, aktualizacja, testowanie) możliwe jest naruszenie integralności zarówno zbioru uczącego, testującego, jak i procesu uczenia/aktualizacji oraz testowania. Zgodnie z propozycją Chakraborty'ego [11] uzasadnione jest rozróżnienie na:

1. ATAK INFEKCYJNY (ang. *poisoning attack*) – atak na proces uczenia lub aktualizacji, który polega na ingerencji w zbiór uczący. W pracy [20] ten rodzaj ataku jest nazywany atakiem przyczynowym (ang. causative *attack*). Procedura atakującego naruszająca integralność zbioru uczącego może być realizowana przez:

   A. infekowanie danych (ang. *data injection*) – realizowane przez dodawanie do zbioru uczącego fałszywych przykładów, a także modyfikowanie lub usuwanie istniejących elementów zbioru uczącego (ang. *data modification*). Tego typu działania mogą na przykład wpływać na taki rozkład klas w zbiorze uczącym, aby spowodować „stronniczość" (ang. *bias*) systemu;

   B. manipulowanie danymi (ang. *data manipulation*) – realizowane przez wpływanie na strukturę zbioru uczącego zarówno poprzez dodanie, modyfikację lub usunięcie cechy wektora obiektu, jak i etykiety klasy (ang. *label modification*). Należy

wspomnieć, że dane uczące niejednokrotnie reprezentują rzeczywiste obiekty. Możliwa jest zatem sytuacja, w której manipulacja następuje na rzeczywistym obiekcie, np. charakteryzacja osoby, której zdjęcie trafia do zbioru uczącego i w konsekwencji reprezentacja tej osoby jest zniekształcona.

2. ATAK INWAZYJNY (ang. *envasion attack*) – atak na proces testowania, który polega na ingerencji w zbiór testujący. W pracy [20] ten rodzaj ataku jest nazywamy atakiem eksploracyjnym (ang. *exploratory attack*). Procedura atakującego naruszająca integralność zbioru testującego może być realizowana przez infekowanie danych, co jest uzyskiwane przez dodawanie do zbioru testującego fałszywych przykładów, a także modyfikowanie lub usuwanie istniejących elementów zbioru testującego. Aby nie zostać wykrytym, atakujący stara się tak modyfikować zbiór testujący, aby odzwierciedlał charakterystykę (rozkład statystyczny) rzeczywistego zbioru testującego.

3. ATAK NA MODEL (ang. *model logic corruption*) – jest to atak przeprowadzony bezpośrednio na model, tak aby uzyskać jego wersję „zniekształconą". Ten rodzaj ataku może wystąpić w sytuacji, w której użytkownik nieświadomie korzysta z algorytmu maszynowego uczenia się, który został pobrany z publicznie dostępnych zainfekowanych[3] środowisk programistycznych [11]. Innym wektorem ataku może być przejęcie wcześniej zainfekowanego modelu poprzez tzw. uczenie poprzez transfer modelu (ang. *transfer learning*)[4].

Możliwości ataków na proces uczenia/aktualizacji i testowania nie ograniczają się tylko do wymienionych rodzajów. Możliwa jest również manipulacja procesem uczenia i testowania, tak aby pogorszyć jakość działania klasyfikatora przez wykorzystanie klasycznych problemów systemów uczących się, takich jak na przykład uczenie na nieaktualnych zbiorach

3 Zawierających podatności (ang. *vulnerability*) w kodzie programów, które są znane atakującemu. Możliwy jest też scenariusz, polegający na tym, że atakujący podmieni używane przez obrońcę oprogramowanie, na oprogramowanie zawierające złośliwy kod.

4 W tym podejściu do rozwiązania nowego problemu wykorzystuje się wcześniej opracowany model dla innego zagadnienia. Jest to relatywnie często wykorzystywana metoda, w warunkach dostępności relatywnie małych zbiorów uczących, do rozpoczęcia uczenia (ang. *starting point*) konwolucyjnych sieci neuronowych.

danych. Atakujący może też dezinformować lub – przez działania socjotechniczne – doprowadzić obrońcę do prowadzenia procesu uczenia w sposób nierzetelny, co może skutkować na przykład przeuczeniem klasyfikatora.

ATAK NA INTEGRALNOŚĆ W TRAKCIE FUNKCJONOWANIA MODELU

System po zbudowaniu również może podlegać atakom na integralność w pośredni sposób. W przypadku udanego ataku na poufność i zdobycia wiedzy całkowitej lub częściowej, atakujący ma potencjalną możliwość odtworzenia modelu klasyfikatora. Na przykład zdobycie wiedzy o zbiorze uczącym oraz o używanym środowisku programistycznym może umożliwić samodzielne zbudowanie wiarygodnego modelu i analizę jego podatności. Taka próba odtworzenia i zbudowania substytutu rzeczywistego modelu jest szczególnie istotnym wektorem ataku. Wynika to z tego, że uzyskanie wpływu na proces uczenia i testowania klasyfikatora jest zazwyczaj niezwykle trudne do osiągnięcia i będzie potencjalnie generować niewspółmiernie wysokie koszty potencjalnego ataku [21][5]. W sytuacji braku jakiejkolwiek wiedzy, system jest traktowany jako czarna skrzynka, która może podlegać eksperymentom mającym na celu zbadanie, jaka będzie jego reakcja na określone dane wejściowe[6]. Jest to klasyczne zadanie identyfikacji, które mają na celu zbudować model systemu na podstawie badań eksperymentalnych, tj. danych pomiarowych zebranych z wejścia i wyjścia identyfikowanego systemu [22]. Zebranie odpowiednio dużej liczby par wejście–wyjście umożliwia zbudowanie zbioru uczącego i wytworzenie przez atakującego substytutu rzeczywistego klasyfikatora. Mając taki model, atakujący może opracować zainfekowane przykłady (ang. *adversarial examples*), które wykorzysta w ataku na rzeczywiście pracujący system. To podejście wymaga od atakującego dostępu do atakowanego systemu.

5 Analiza rentowności ataku, czyli zestawienie kosztów przeprowadzenia ataku z korzyściami z niego wynikającymi jest ważnym zagadnieniem badawczym, umożliwiającym analizę ryzyka i określenie najbardziej prawdopodobnych wektorów ataków. Oczywiście to podejście jest adekwatne dla grup przestępczych kierujących się zyskiem, co nie jest na przykład kryterium dla aktorów państwowych (ang. *state actor*).

6 Czasami ten rodzaj ataku na czarną skrzynkę jest nazywany nieformalnie atakiem z odwołaniem się do wyroczni (ang. *oracle attack*), w którym dla zadanych pytań (wejście) wyrocznia odpowiada (wyjście).

Uwagi końcowe

Intencjonalne ataki na systemy maszynowego uczenia to realne zagrożenie dla współczesnego świata „zanurzonego" w technologiach cyfrowych. Jesteśmy obecnie w punkcie zwrotnym, w którym obnażone zostały słabości tych systemów i znane są relatywnie proste scenariusze skutecznych ataków. W tym kontekście kluczowe staje się zagadnienie budowania systemów uczących się, które będą odporne na próby ataków (ang. *robust machine learning*). Jest to obecnie jeden z najważniejszych obszarów badań naukowych w zakresie metod sztucznej inteligencji, który docelowo miałby wypracować teorię bezpieczeństwa informacji (ang. *theory of information security*). Obecnie najpoważniej to zagadnienie jest analizowane w kontekście spektakularnych sukcesów systemów opartych na głębokim uczeniu się. W tym zakresie należy wspomnieć prace Mądrego [23] oraz badania Goodfellowa [24]. Próbą całościowego spojrzenia oraz przedstawienia aktualnego stanu prac badawczych jest publikacja Zhanga [25] oraz w szerszym kontekście sztucznej inteligencji, z uwzględnieniem np. dużych modeli językowych (ang. *large language models*), praca Marcusa [26].

Bibliografia

[1] J. Surma. Cyfryzacja życia w erze Big Data: człowiek, biznes, państwo. Warszawa: Wydawnictwo Naukowe PWN, 2017.

[2] R. Shankar, S. Kumar, M. Nyström, J. Lambert, A. Marshall i in. (19 marca 2021). „Adversarial Machine Learning – Industry Perspectives" [Online]. Dostęp: https://arxiv.org/abs/2002.05646.

[3] J. Surma. „Wstęp do hakowania systemów uczących się", w: *Hakowanie Sztucznej Inteligencji*, Warszawa: Wydawnictwo Naukowe PWN, 2020.

[4] N. Dalvi, P. Domingos, M. Sumit, D. Verma. „Adversarial classification", w: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'04)*, ACM Press, 2004.

[5] M. Barreno, B. Nelson, R. Sears, A. Joseph, J. Tygar. „Can machine learning be secure?" w: *ASIACCS'06*, 2006.

[6] B. Nelson, B. *Behavior of Machine Learning Algorithms in Adversarial Environments. Technical Report No. UCB/EECS-2010-140*. Electrical Engineering and Computer Sciences, University of California at Berkeley, 2010.

[7] M. Barreno, B. Nelson, A. Joseph, J. Tygar, „The security of machine learning", *Machine Learning*, vol. 81 no. 2, ss. 121–148, 2010.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, i in. (10 czerwca 2014). „Generative Adversarial Networks", [Online]. Dostęp: https://arxiv.org/abs/1406.2661

[9] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Celik, A. Swami, „Practical Black-Box Attacks against Machine Learning", w: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (ASIA CCS '17). New York: ACM, 2017.

[10] P. McDaniel, N. Papernot, Z. Celik, „Machine Learning in Adversarial Settings", *IEEE Security & Privacy*, vol. 14, np. 3, pp.68-72, maj-czerwiec 2016.

[11] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay (28 września 2018). „Adversarial Attacks and Defences: A Survey" [Online]. Dostęp: https://arxiv.org/abs/1810.00069.

[12] L. Muñoz-González, E. C. Lupu, „The Security of Machine Learning Systems", w: L. Sikos, (ed.) *AI in Cybersecurity*, Intelligent Systems Reference Library, vol 151. Springer: Cham, 2009, ss. 47-79.

[13] A. Oprea, A. Vassilev (8 marca 2023). „Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations" [Online]. Dostęp: https://csrc.nist.gov/publications/detail/white-paper/2023/03/08adversarial-machine-learning-taxonomy-and-terminologydraft

[14] MITRE ATLAS [Online]. Dostęp: https://atlas.mitre.org/

[15] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[16] K. Liderman, *Bezpieczeństwo informacyjne*. Warszawa: Wydawnictwo Naukowe PWN, 2017.

[17] A. Oprea, A. Vassilev (8 marca 2023). „Adversarial Machine Learning: A Taxonomy and Terminology of Attacks andMitigations" [Online]. Dostęp: https://csrc.nist.gov/

publications/detail/white-paper/2023/03/08adversarial-machine-learning-taxonomy--and-terminology/draft

[18] NIST, „Guide for Conducting Risk Assessments", NIST 800-30 [Online], wrzesień 2012. Dostęp: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication-800-30r1.pdf

[19] J. Surma, „*Hacking Machine Learning: Towards The Comprehensive Taxonomy of Attacks Against Machine Learning Systems*", w: *ICIAI 2020: Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence*, May 2020.

[20] L. Huang, A. Joseph, B. Nelson, B. Rubinstein, J. Tygar, „Adversarial machine learning", w: *Proceedings of the 4th ACM workshop on Security and artificial intelligence* (AISec '11). ACM Press, 2011.

[21] J. Surma, „Attack Vectors on Supervised Machine Learning Systems in Business Applications", *Informatyka Ekonomiczna* vol. 3 no. 57, 2020.

[22] Z. Bubnicki, *Identyfikacja obiektów sterowania*. Warszawa: PWN, 1974.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu (19 czerwca 2017), „Towards Deep Learning Model*s Resistant to Adversarial Attacks*" [Online]. Dostęp: https://arxiv.org/abs/1706.06083.

[24] I. Goodfellow, P. McDaniel, N. Papernot, „Making machine learning robust against adversarial inputs", *Communication of the ACM*, vol. 61 no. 7, ss. 56–66, 2018.

[25] J. Zhang, K. Liu, F. Khalid, M. Hanif, S. Rehman, i in., „Building Robust Machine Learning Systems: Current Progress, Research Challenges, and Opportunities", w: *Proceedings of the 56th Annual Design Automation Conference*, 2019.

[26] G. Marcus (14 lutego 2020). „The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence" [Online]. Dostęp: https://arxiv.org/abs/2002.06177.

Weryfikacja wiarygodności systemów w erze uczenia maszynowego

Mateusz Krzysztoń

NASK-PIB | Centrum Badań i Rozwoju | Zakład Systemów Rozproszonych

Wstęp

Obecnie obserwowany rozwój technologii w obszarze sztucznej inteligencji przekłada się na coraz szybciej postępującą cyfryzację gospodarki i społeczeństwa. Zatem, w obliczu wzrostu wpływu systemów informatycznych na kolejne obszary życia, konieczne jest aby tworzone systemy były godne zaufania. Wiarygodne systemy muszą spełniać potrzeby interesariuszy systemu w zakresie bezpieczeństwa, prywatności, niezawodności i integralności biznesowej [1]. W przypadku systemów tradycyjnych znanych jest wiele sprawdzonych metodologii i dobrych praktyk wytwarzania, jak również kryteriów weryfikacji poszczególnych aspektów systemu składających się na ostateczną ocenę wiarygodności [2].

Niestety, wraz z pojawieniem się systemów wykorzystujących techniki uczenia maszynowego (ang. *machine learning*, ML), dotychczasowa wiedza o tworzeniu wiarygodnych systemów przestała być wystarczająca. Wynika to przede wszystkim z faktu, że o ile tradycyjne systemy opierają się na wiedzy eksperckiej, o tyle w systemach uczących się wiedza pochodzi

z danych[1], które charakteryzują się różnym stopniem jakości i, co często pomijane, wrażliwości. Konieczne stało się zatem ponowne przemyślenie takich aspektów wpływających na wiarygodność systemu jak bezpieczeństwo, przewidywalność działania czy ochrona danych wrażliwych. Ponadto, jak pokazały badania [4], systemy informatyczne ML mogą być uprzedzone do wybranych grup społecznych czy etnicznych. Obniżone zaufanie do podstaw działania systemu spowodowało zwiększone zapotrzebowanie na poznanie nie tylko samej decyzji systemu, ale również jej uzasadnienia, co w przypadku większości typów algorytmów uczenia maszynowego jest problematyczne [5] [6]. Biorąc pod uwagę powyższe aspekty wyniki tegorocznego ogólnoświatowego badania, wskazujące na niskie zaufanie społeczne do wykorzystania sztucznej inteligencji, nie mogą dziwić [7].

Lepsze zrozumienie jak działa sztuczna inteligencja, co stoi u podstaw problemów z jej wiarygodnością, a także jak weryfikować czy system ML jest godny zaufania, leży w interesie całego społeczeństwa. Dotyczy to zarówno dostawców systemów, którzy w ten sposób zwiększą zaufanie do swoich produktów, jak i odbiorców, którzy chętniej, świadomiej i bezpieczniej będą wykorzystywać nowe możliwości ery sztucznej inteligencji. Rozwój zagadnień związanych z wiarygodnością systemów dotyczy również podmiotów publicznych odpowiedzialnych za to, aby rozwój sztucznej inteligencji był możliwy i poprawiał jakość życia, a jednocześnie przebiegał zgodnie z wartościami i standardami społecznymi.

Rozszerzenie pojęcia wiarygodności systemu informatycznego znalazło swoje odzwierciedlenie w dokumencie „Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji" [8], będącym wynikiem pracy grupy eksperckiej powołanej przez Komisję Europejską. Dokument zawiera propozycję ram wytwarzania wiarygodnych systemów sztucznej inteligencji (ang. *artificial intelligence*, AI). Poruszono w nim zarówno kwestię technicznej solidności systemu, jak również prawne i etyczne aspekty pojawiające się w całym cyklu życia systemu. Mając na uwadze, że każdy aspekt wiarygodności jest niezbędny, ale jednocześnie niewystarczający do osiągnięcia

1 Stwierdzenie to jest pewnym uproszczeniem – należy pamiętać, że za proces zebrania i przygotowania danych odpowiedzialni są eksperci i jakość zbioru danych powinna wynikać również z ich wiedzy. Istnieją również techniki pozwalające na integrację wiedzy eksperckiej w ramach systemu AI w celu zwiększania jego wiarygodności [3]. Niemniej ilość i złożoność danych potrzebnych do wytworzenia systemu (i jego utrzymania) najczęściej uniemożliwia pełne zrozumienie i zweryfikowanie danych przez eksperta.

godnej zaufania sztucznej inteligencji, w dalszej części przedstawione zostaną trzy wybrane, mierzalne aspekty wiarygodności systemów ML, aby przybliżyć problematykę technicznej weryfikacji wiarygodności systemów.

Potencjalne źródła braku technicznej wiarygodności systemu uczącego się

Uczenie maszynowe to aktualnie najbardziej dynamiczny obszar AI, w którym rozwijane są metody pozwalające komputerom na uczenie się z danych. Dzięki ML możliwe jest wykrywanie w nich wzorców, a następnie dokonanie predykcji na podstawie dopasowania do tych wzorców. Celem procesu uczenia maszynowego jest zatem wytrenowanie modelu z wykorzystaniem danych historycznych. Wynikowy model służy do przewidywania przez system poprawnych odpowiedzi dla nowych rekordów. Zatem jakość zbiorów historycznych dostępnych w czasie uczenia ma szczególne znaczenie dla późniejszej wiarygodności systemu – zbiory danych powinny być znaczących rozmiarów, poprawne (zawierać wartości możliwie bliskie rzeczywistym wartościom), aktualne, istotne dla rozważanego problemu, a także różnorodne i reprezentatywne.

Choć dane leżą u podstaw wiedzy systemu AI, są one tylko jednym z czynników wpływających na jego finalną wiarygodność. Proces tworzenia systemu jest równie istotny. Przeprowadzony poprawnie, pozwala na uzyskanie wyższej jakości predykcyjnej niż wynikałoby to z jakości danych. Jednak z drugiej strony nawet przy wysokiej jakości danych błędy projektowe, procesowe lub programistyczne mogą skutkować stworzeniem modelu niskiej klasy. Zatem istotne jest, aby odpowiednie procedury weryfikacji szeroko rozumianej wiarygodności ML były obecne w całym procesie wytwórczym systemu. Przykładowo, zbytnie dopasowanie modelu do danych (tzw. przeuczenie modelu) lub wybór zbyt wielu cech o niskim znaczeniu może zwiększyć podatność systemu na ataki z obszaru kontradyktoryjnego uczenia maszynowego[2]. Dbałość o wiarygodność systemu dotyczy również jego fazy utrzymania. Z uwagi na możliwą

2 Więcej na ten temat w rozdziale niniejszej publikacji pt. „Zagadnienie antagonistycznego uczenia maszynowego i przykład ataku na algorytmy uczenia maszynowego nadzorowanego" autorstwa Mateusza Bursiaka.

zmianę charakterystyki danych w czasie (ang. *data drift*), jakość działania sytemu może ulegać degradacji [9]. Zatem również monitorowanie wiarygodności systemu w czasie wykracza znacząco poza zakres znany z tradycyjnych systemów informatycznych.

Sprawiedliwość

W kontekście systemów decyzyjnych, sprawiedliwość definiuje się jako brak faworyzowania oraz jakichkolwiek uprzedzeń do danej jednostki lub grupy na podstawie ich wrodzonych lub nabytych cech [4]. Ze sprawiedliwością związane jest pojęcie stronniczości[3] (ang. *bias*) systemów ML, którą z kolei definiuje się jako negatywne, niepożądane konsekwencje działania tych systemów, zwłaszcza jeśli konsekwencje te w nieproporcjonalny sposób dotykają określonych grup ludzi. W artykule [10] zidentyfikowano siedem różnych źródeł stronniczości modeli:

1. STRONNICZOŚĆ HISTORYCZNA (ang. *historical bias*) – dane treningowe zawierają uprzedzenia lub inne wady istniejące w rzeczywistym zjawisku. Przykładem mogą być rasistowskie wypowiedzi znajdujące się w korpusie służącym do wytworzenia systemu ChatGPT, przez co również (początkowo) system ten wykazywał cechy rasistowskie [11]. Warto zauważyć, że gdy niesprawiedliwość modeli wynika z uprzedzeń zawartych w danych, a te odzwierciedlają uprzedzenia obecne w rzeczywistości, tworzenie sprawiedliwych systemów ML wraz z ich wdrażaniem w miejsce istniejących procesów może przyczynić się do zmniejszenia powszechnych nierówności [8];

2. STRONNICZOŚĆ REPREZENTATYWNOŚCI (ang. *representation bias*) – w danych treningowych pewna część populacji jest niedoreprezentowana, a następnie w czasie budowy systemu nie udaje się dobrze uogólnić modelu dla tej grupy na podstawie zebranych danych [12]. Często niska reprezentacja danej grupy wynika ze sposobu zbierania danych – np. pozyskiwanie ich od użytkowników aplikacji mobilnych

3 W obszarze ML pojęcie stronniczości jest często nadużywane do opisu wszelkich niedoskonałości systemu. Jednak z uwagi na fakt, że zjawisko to jest powszechne i przyjęło się w literaturze, również naukowej, także w niniejszym opracowaniu pojęcie stronniczości będzie stosowane bardzo szeroko do określania wad systemu.

spowoduje niski udział osób starszych w danych treningowych i zagrożenie słabej jakości działania systemu dla tej grupy;

3. STRONNICZOŚĆ POMIARU (ang. *measurement bias*) – jakość pomiaru danych może być różna w zależności od badanych grup (np. kondycja zdrowotna w grupach o różnej zamożności). Dotyczy to również sytuacji, gdy dane są upraszczane w czasie zbierania (np. badanie ankietowe z pytaniami zamkniętymi), co powoduje, że część aspektów rzeczywistego zjawiska może nie znaleźć odzwierciedlenia w danych;

4. STRONNICZOŚĆ AGREGACJI (ang. *aggregation bias*) – stworzenie jednego modelu dla podgrup o różnej charakterystyce, bez uwzględnienia danych umożliwiających ich rozróżnienie, może spowodować powstanie modelu działającego poprawnie tylko dla jednej podgrupy (dominującej w danych treningowych) albo modelu o niskiej jakości w ogóle. Przykładowo, przy diagnozowaniu cukrzycy istotnym wskaźnikiem jest HbA1c, którego poziomy referencyjne różnią się w zależności od grupy etnicznej i płci. Aby uniknąć stronniczości, konieczne jest stworzenie osobnych modeli dla podgrup lub dołączenie informacji o różnicach między grupami do systemu [13];

5. STRONNICZOŚĆ UCZENIA (ang. *learning bias*) – w fazie budowania systemu konieczne jest wskazanie funkcji celu, która jest optymalizowana poprzez trening. Stronniczość może pojawić się np. w sytuacji gdy priorytet nadany ogólnej dokładności modelu jest wyższy niż ten nadany równomierności typów błędów;

6. STRONNICZOŚĆ EWALUACJI (ang. *evaluation bias*) – ten typ stronniczości powstaje, gdy ocena modelu nie jest dokonywana na zbiorze reprezentatywnym dla populacji, której system będzie służył. Poprawna ewaluacja modelu jest często utrudniona z uwagi na brak znajomości charakterystyki docelowej populacji lub kiedy istnieje wiele docelowych populacji o różnej charakterystyce;

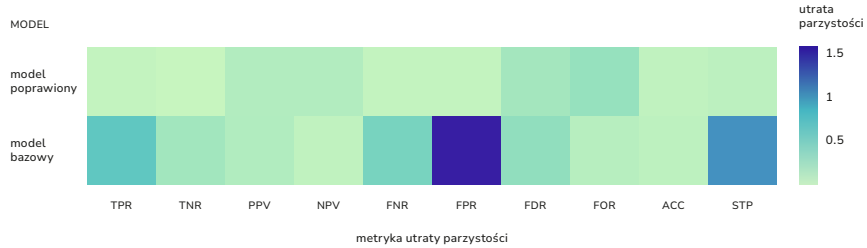
7. STRONNICZOŚĆ WDROŻENIA (ang. *deployment bias*) – faktyczne wykorzystanie modelu w inny sposób, niż był zakładany przy projektowaniu. Przykładem jest stosowanie w niektórych stanach USA systemu do oceny prawdopodobieństwa popełnienia przestępstwa, wbrew ostrzeżeniom jego twórców, do wsparcia sędziów przy określaniu długości wyroku [14].

Powyższe rodzaje stronniczości występują w różnych fazach wytwarzania systemu ML. Efektem wystąpienia stronniczości przynajmniej jednego rodzaju jest stworzenie niesprawiedliwego modelu. Aby zilustrować, jak można badać sprawiedliwość modelu przedstawiony zostanie przykład [15] zaproponowany przez twórców biblioteki Dalex [16], która służy do badania modeli ML, w tym identyfikacji wbudowanych w nich uprzedzeń.

Na rysunku 1 przedstawiono różne miary jakości dwóch modeli, których celem jest określenie, czy podejrzany jest winny – modelu bazowego i modelu poprawionego (sprawiedliwszego)[4]. Analiza sprawiedliwości została przeprowadzona w dwóch grupach, do których osoby zostały przydzielone na podstawie płci. Kolorem przedstawiono sprawiedliwość modeli pod względem różnych miar jakości modelu – im bardziej granatowy odcień, tym mniejsza sprawiedliwość według danej miary. Na podstawie analizy dokładności modelu bazowego (ACC, jasnozielony kolor) można stwierdzić, że model bazowy jest sprawiedliwy – dla obu płci prawdopodobieństwo poprawnego określenia winy podejrzanego jest bardzo zbliżone. Jednak dokładność jest powiązana z tym, jak często model się myli, bez uwzględnienia tego, w jaki sposób– w tym wypadku, czy mylnie uzna podejrzanego za przestępcę czy za niewinnego. Zatem aby stwierdzić, że model jest sprawiedliwy, należy dodatkowo zweryfikować, czy jednakowo często myli się na korzyść i niekorzyść podejrzanego niezależnie od płci. Służą do tego metryki FNR oraz FPR, czyli odpowiednio częstość błędów fałszywie negatywnych (uznania kogoś błędnie za niewinnego) oraz fałszywie pozytywnych (uznanie kogoś błędnie za winnego). Na podstawie analizy sprawiedliwości tych dwóch miar można stwierdzić, że mimo iż model ma taką samą skuteczność niezależnie w obu grupach, to typ błędu jest mocno zależny od płci. Oznacza to, że model będzie w obu grupach mylił się jednakowo często, jednak częściej będzie skazywać niewinnych mężczyzn niż niewinne kobiety oraz częściej będzie uniewinniać winne kobiety niż winnych mężczyzn (lub odwrotnie)[5].

4 Do poprawienia modelu wykorzystano narzędzie FairTorch (https://github.com/wbawakate/fairtorch)

5 W celu wskazania dyskryminowanej grupy konieczne jest przeprowadzenie dodatkowej analizy.



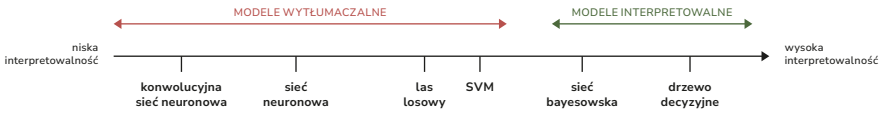
RYS. 1. Porównanie sprawiedliwości dwóch modeli – modelu niesprawiedliwego i modelu poprawionego (sprawiedliwszego) [15].

Interpretowalność i wytłumaczalność

Zagadnienie odpowiedniej komunikacji decyzji było przedmiotem rozważań już w tradycyjnych systemach eksperckich. Jak stwierdzono w pracy [17] z 1993 roku wzbogacenie decyzji pochodzących z systemów eksperckich o dodatkowe wyjaśnienie sprawia, że decyzja systemu jest bardziej akceptowalna dla ludzi. Wyjaśnialność działania systemu stała się tym bardziej istotna w kontekście pojawiających się wątpliwości związanych ze sprawiedliwością systemów ML. Tak jak w życiu codziennym w różnym stopniu jesteśmy w stanie wyjaśnić otaczające nas zjawiska (np. prawami fizycznymi, ciągiem przyczynowo-skutkowym czy jedynie zaobserwowaną korelacją), tak również decyzje systemów ML mogą charakteryzować się różnym stopniem wyjaśnialności. Choć stopień zrozumienia ma szerokie spektrum wartości, systemy ML można podzielić na dwa główne rodzaje [18]: **interpretowalne** i **wytłumaczalne**. Najprościej różnicę między tymi pojęciami można ująć następująco – modele wytłumaczalne mogą być zrozumiane przez człowieka tylko z wykorzystaniem dodatkowych technik, natomiast do zrozumienia modeli interpretowalnych wystarczy poznanie ich budowy i parametrów [19].

Znaczący wpływ na stopień wyjaśnialności modelu ma jego typ – np. drzewo decyzyjne jest modelem łatwo interpretowalnym, gdyż podjęcie decyzji na podstawie drzewa polega na odpowiedzi na szereg jawnych pytań alternatywnych, które można prześledzić i zrozumieć. Na rysunku 2 zobrazowano poziom interpretowalności wybranych typów modeli ML. Oczywiście, na interpretowalność modelu, czyli możliwość zrozumienia go przez człowieka, ma również wpływ wielkość modelu – w przypadku bardzo złożonych, sposób działania nawet tych najłatwiejszych

do interpretacji będzie trudny do uchwycenia przez człowieka. Niestety, na ogół modele łatwiejsze do zrozumienia cechują się niższą skutecznością [20].

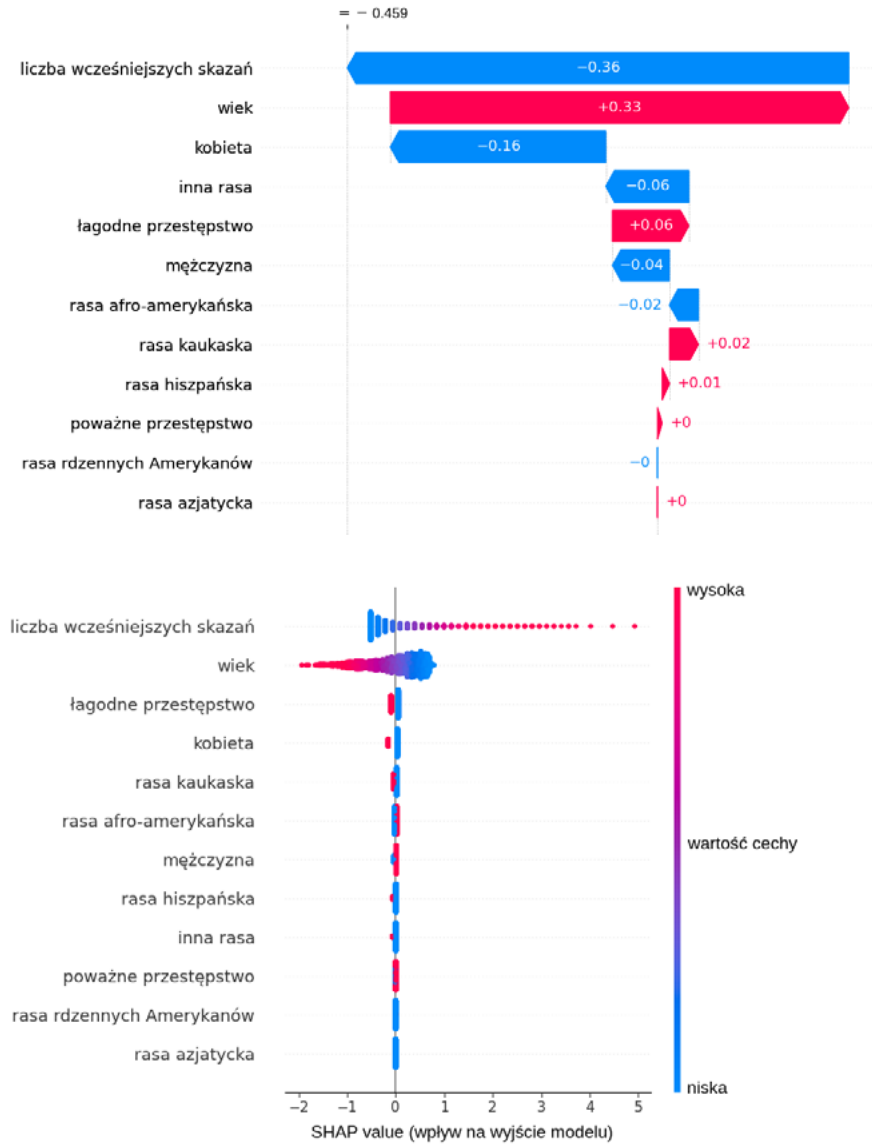


RYS. 2. Interpretowalność wybranych typów modeli ML [20]

Zatem w przypadku systemów trudniejszych w interpretacji, gdy wymagane jest zrozumienie sposobu ich działania, konieczne jest wykorzystanie dodatkowych narzędzi z obszaru XAI (ang. *explainable AI*). Standardowym podejściem jest wyjaśnialność *post hoc*, tj. po wytrenowaniu modelu [21]. Jednym z popularniejszych narzędzi do realizacji tego typu analizy jest *SHapley Additive exPlanations* (SHAP) [22]. Bada ono znaczenie każdej cechy dla decyzji podjętej przez system dla danego rekordu danych. Aby uzyskać wkład dla danej cechy, konieczne jest przejście przez przestrzeń cech i obserwowanie, jak wpływa ona na przewidywania modelu, gdy jest uwzględniona w procesie uczenia. Przykładowy wynik analizy dla modelu bazowego oceniającego winę podejrzanego przedstawiono na rysunku 3. Po lewej stronie przedstawiono analizę dla jednego podejrzanego – można zaobserwować, że wpływ na decyzję o winie podejrzanego miały w największym stopniu liczba wcześniejszych przestępstw, wiek oraz płeć. Natomiast zagregowany wynik dla wszystkich elementów w zbiorze treningowym (prawa część rysunku 3) wskazuje, że wiek był istotnym czynnikiem wpływającym na decyzję, a kobiety były faworyzowane przez model (według modelu bycie kobietą zmniejszało delikatnie szansę na popełnienie kolejnego przestępstwa).

Mimo dużej popularności, narzędzia *post hoc* spotykają się z coraz większą krytyką. W artykule [6] zwrócono uwagę, że w celu wyjaśnienia modeli, szczególnie tak skomplikowanych jak głębokie sieci neuronowe, konieczne jest uproszczenie działania modelu, co może jednak wpłynąć negatywnie na jakość działania. Nadinterpretowanie wyników narzędzia SHAP skłoniło twórców tego narzędzia do opublikowania ostrzeżenia [23], w którym zwracają uwagę, że analiza SHAP pozwala na ujawnienie korelacji w danych wykorzystywanych przez model, co nie oznacza jednak przyczynowości. Z tego powodu, wyjaśnienia są często

nieintuicyjne (a czasami wręcz sprzeczne z intuicją). Natomiast w artykule [24] pokazano, że wyjaśnienia modeli są bardzo niestabilne – dla nieznacznie różniących się elementów decyzje modelu są jednakowe, jednak wyjaśnienia tych decyzji bardzo różne. Cecha ta sprawia, że nawet mocno stronnicze systemy są potencjalnie w stanie oszukać narzędzia, takie jak SHAP, poprzez wymuszenie wyjaśnienia niewskazującego na stronniczość.



RYS. 3. Analiza SHAP *(opracowanie własne)*

Odpowiedzią na powyższe problemy analizy typu *post hoc* mogą okazać się metody z obszaru przyczynowego uczenia maszynowego (ang. *casual machine learning*) [25], które zamiast wykrywania samych wzorców korelacji, mają na celu budowanie modeli opierających się na zasadach rozumowania przyczynowego, a więc potencjalnie pozwolą na budowę modeli o znacznie wyższym poziomie interpretowalności. Modele te są jednak znacznie trudniejsze w budowie i wymagają większej wiedzy dziedzinowej w procesie wytwarzania systemu.

Bezpieczeństwo

Innym istotnym aspektem wiarygodności systemów ML jest bezpieczeństwo. Istnieje wiele typów ataków specyficznych dla systemów ML [26]. Najczęściej rozważanym w literaturze celem atakującego jest obniżenie jakości działania systemu poprzez atak na działający system (ataki unikania, ang. *evasion attacks*) lub na proces uczenia (ataki zatrucia danych, ang. *poisoning attacks*). Nie mniej istotnym aspektem bezpieczeństwa systemu jest zachowanie poufności danych trenujących, szczególnie jeśli wśród nich znajdują się dane wrażliwe. W trakcie uczenia systemu każdy element zbioru treningowego pozostawia pewien ślad w modelu decyzyjnym. Zjawisko to rodzi nową podatność na ataki – zadając systemowi odpowiednie pytania można próbować wnioskować o danych treningowych. Taki typ ataków nazwano atakiem wnioskowania (ang. *inference attack*). Wyróżnia się jego dwa rodzaje:

- atak wnioskowania o atrybutach (ang. *attribute inference attack*, AIA) [27] – atakujący posiada pewną częściową wiedzę o danych treningowych i stara się odkryć nieznane wartości atrybutów wrażliwych;
- atak wnioskowania o członkostwie (ang. *membership inference attack*, MIA) [28] – atakujący posiada rekord danych i stara się ustalić, czy należał on do zbioru treningowego.

Zwiększona podatność systemu na ten rodzaj ataków najczęściej wynika z przetrenowania modelu [28], tj. zbytniego dopasowania modelu do danych treningowych. Powoduje ono, że model wykazuje dużą pewność decyzji dla tych i tylko tych danych, co z kolei pozwala atakującemu odróżnić dane treningowe od pozostałych danych. Najprostszym sposobem,

od którego można rozpocząć szacowanie skuteczności potencjalnego ataku wnioskowania na dany model, jest określenie stopnia przetrenowania modelu np. z wykorzystaniem miary określającej lukę uogólnienia (ang. *generalization gap*) [29]. Bardziej zaawansowane techniki polegają na empirycznym badaniu zachowania modelu dla odpowiednio preparowanych danych syntetycznych [30] lub przeprowadzeniu ataku na docelowym systemie z wykorzystaniem gotowych narzędzi (np. AIJack[6]).

Podsumowanie

Badanie wiarygodności systemów ML stanowi wyzwanie, jest jednak konieczne w obliczu ich rozwoju. Mimo że istnieją narzędzia pomagające w ocenie wiarygodności gotowych systemów, należy mieć na uwadze, że wiarygodność systemu powinna wynikać z całego procesu wytwarzania. Pokusa, by najpierw wytworzyć system z pominięciem wszystkich lub wybranych aspektów wiarygodności, a wiarygodność sprawdzić na końcu procesu, może okazać się zgubna. Dokładne zweryfikowanie wszystkich czynników składających się na wiarygodność systemu (powyżej przedstawiono jedynie wybrane aspekty wymagające weryfikacji) może być bardzo trudne i kosztowne. Dodatkowo weryfikacja poprzez techniczny audyt systemu nie odpowie na wszystkie istotne pytania, np. czy dane wykorzystane do trenowania zostały zebrane zgodnie z prawem.

Zwiększenie wiarygodności systemów ML powinno nie tylko na zabezpieczyć odbiorców systemów, ale również przełożyć się na większą chęć wdrażania nowych rozwiązań, a przez to dalszy rozwój systemów uczących się. Dlatego podejście polegające na *trustworthiness by design*, połączone z cykliczną weryfikacją techniczną systemu, często zawężoną do najistotniejszych kwestii w danej domenie, wydaje się najbardziej obiecujące.

Należy jeszcze raz podkreślić, że wskazane powyżej obszary wiarygodności stanowią jedynie wąski wycinek wszystkich technicznych aspektów, które należy rozważyć przy projektowaniu, tworzeniu i wdrażaniu systemów AI. Pełniejszy obraz, uwzględniający również kwestie prawne i etyczne, został przedstawiony we wspomnianym wcześniej dokumencie „Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji" [8].

6 https://github.com/Koukyosyumei/AIJack

Bibliografia

[1] K. Saleh, G. Elshahry, „Modeling security requirements for trustworthy systems", *Encyclopedia of Information Science and Technology*, Second Edition, pp. 2657–2664, 2009.

[2] S. Paulus, N. G. Mohammadi, T. Weyer, „Trustworthy software development", *Communications and Multimedia Security*, B. De Decker, J. Dittmann, C. Kraetzer, C. Vielhauer, Eds., Springer, Berlin, Heidelberg, 2013.

[3] C. Deng, X. Ji, C. Rainey, J. Zhang, W. Lu, „Integrating machine learning with human knowledge", *iScience*, vol. 23, no. 11, p. 101656, 2020.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, „A survey on bias and fairness in machine learning", *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[5] C. E. Perez (19 listopada 2019). „Fake intuitive explanations in AI" [Online]. Dostęp: https://medium.com/intuitionmachineachieving-fake-explanations-in-ai-5e63b289a3ef.

[6] H. de Bruijn, M. Warnier, M. Janssen, „The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making", *Government Information Quarterly*, vol. 39, no. 2, p. 101666, 2022.

[7] N. Gillespie, S. Lockey, C. Curtis, J. Pool A. Akbari, "Trust in Artificial Intelligence: A global study", The University of Queensland, KPMG Australia [Online], 2023. Dostęp: https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf

[8] N. A. Smuha, „The EU approach to ethics guidelines for Trustworthy Artificial Intelligence", *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019.

[9] I. Žliobaitė, M. Pechenizkiy, J. Gama, „An overview of concept drift applications", *Studies in Big Data*, pp. 91–114, 2015.

[10] H. Suresh J. Guttag, „A framework for understanding sources of harm throughout the machine learning life cycle", w: *EAAMO'21 Equity and Access in Algorithms, Mechanisms, and Optimization*, NY, USA, 2021.

[11] S. Singh, N. Ramakrishnan (9 kwietnia 2023). „Is CHATGPT biased? A Review" [Online]. Dostęp: https://osf.io/9xkbu/

[12] Y. Li, N. Vasconcelos, „Repair: Removing representation bias by dataset resampling" w: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, USA, 2019.

[13] H. Suresh, J. V. Guttag, „A framework for understanding unintended consequences of machine learning", w: *EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization*, NY, USA, 2021.

[14] E. Collins, „Punishing risk„, *Georgetown Law Journal*, vol. 107, p. 57-108, 2018.

[15] J. Wiśniewski (12 stycznia 2021). „Visualize ML model bias with dalex!" [Online]. Dostęp: https://medium.com/responsiblemlvisualize-ml-model-bias-with-dalex-b-63f182cd649.

[16] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, P. Biecek, „Dalex: responsible machine learning with interactive explainability and fairness in python", *The Journal of Machine Learning Research* vol. 22, no. 1, pp. 9759-9765, 2021.

[17] W. R. Swartout, J. D. Moore, „Explanation in second generation expert systems„, w: *Second generation expert systems*, J-M. David, J-P. Krivine, R. Simmons (Eds.), Springer Berlin Heidelberg, pp. 543–585, 1993.

[18] R. Marcinkevičs, J. E. Vogt (3 grudnia 2020). „Interpretability and explainability: A machine learning zoo mini-tour" [Online]. Dostęp: https://arxiv.org/abs/2012.01805.

[19] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Berlin: Chrostoph Molnar, 2022.

[20] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, „Explainable AI: A brief survey on history, research areas, approaches and challenges", w: *Natural Language Processing and Chinese Computing: 8th CCF International Conference*, Dunhuang, China, pp. 563–574, 2019.

[21] D. Vale, A. El-Sharif, M. Ali, „Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law", *AI and Ethics* vol. 2, no. 1, pp. 1–12, 2022.

[22] S. M. Lundberg, S.-I. Lee, „A unified approach to interpreting model predictions", *Advances in neural information processing systems*, vol. 30, pp. 4768–4777, 2017.

[23] E. Dillon, J. LaRiviere, S. Lundberg, J. Roth, V. Syrgkanis (2018). „Be careful when interpreting predictive models in search of causal insights" [Online]. Dostęp: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%C2%A0insights.html

[24] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, „Fooling lime and shap: Adversarial attacks on post hoc explanation methods", w: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

[25] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, R. Silva (30 czerwca 2022). „Causal machine learning: A survey and open problems" [Online]. https://arxiv.org/abs/2206.15475.

[26] B. Biggio F. Roli, „Wild patterns: Ten years after the rise of adversarial machine learning", w: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2154–2156, 2018.

[27] B. Z. H. Zhao i in., „On the (in) feasibility of attribute inference attacks on machine learning models", w: *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 232–251, 2021.

[28] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, X. Zhang, „Membership inference attacks on machine learning: A survey", *ACM Computing Surveys (CSUR)*, vol. 54, no. 11, pp. 1–37, 2022.

[29] J. W. Bentley, D. Gibney, G. Hoppenworth, S. K. Jha (11 września 2020). „Quantifying membership inference vulnerability via generalization gap and other model metrics" [Online]. Dostęp: https://arxiv.org/abs/2009.05669.

[30] H. Jalalzai, E. Kadoche, R. Leluc, V. Plassier (27 lipca 2022). „Membership Inference Attacks via Adversarial Examples" [Online]. Dostęp: https://arxiv.org/abs/2207.13572.

Zagadnienie antagonistycznego uczenia maszynowego i przykład ataku na algorytmy uczenia maszynowego nadzorowanego

Mateusz Bursiak
NASK-PIB

Wstęp

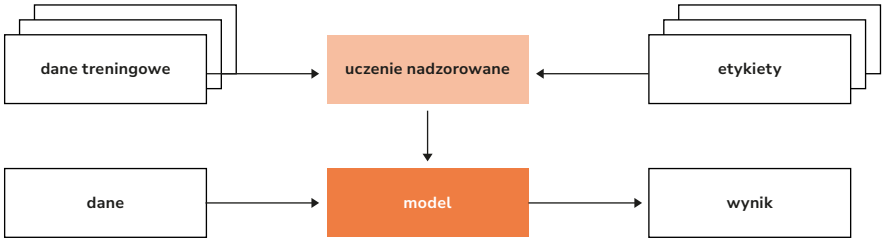
Uczenie maszynowe wykorzystywane jest we wszystkich aspektach naszego życia. W ostatnich latach algorytmy towarzyszą nam w wyborze produktów, profilowaniu, zabezpieczają nasze konta bankowe i systemy informatyczne, pomagają w podróży, czy diagnozowaniu i leczeniu pacjentów. Uczenie maszynowe jest nieodłączną częścią każdej branży i staje się równie powszechne, co strony internetowe, które tak często odwiedzamy. Dynamiczna popularyzacja nowych technologii niesie ze sobą pewne ryzyka. Analogicznie jak w przypadku pierwszych sieci internetowych i protokołów w nich wykorzystywanych, aspekt bezpieczeństwa charakteryzuje się zazwyczaj stosunkowo niższym zainteresowaniem w początkowych fazach rozwoju technologii, stając się podatnym gruntem dla złośliwych aktorów, którzy chcą wykorzystać podatności takich systemów do osiągnięcia własnych korzyści.

Przegląd klasyfikacji uczenia maszynowego

Statystyczne uczenie maszynowe zdefiniowane zostało w roku 1997 przez Toma Mitchella jako system, który uczy się zadań (decyzji, predykcji) na podstawie doświadczeń (danych). Efektywność takiego systemu mierzy się miarą wydajności, która poprawia się wraz ze wzrostem doświadczenia [1]. W literaturze uczenie maszynowe klasyfikuje się najczęściej jako uczenie nadzorowane (ang. *supervised learning*), nienadzorowane (ang. *unsupervised learning*) oraz uczenie ze wzmocnieniem (ang. *reinforcement learning*). Niekiedy połączenie dwóch pierwszych podejść traktuje się jako osobną klasę – uczenie częściowo nadzorowane (ang. *semi-supervised learning*). Odpowiedni dobór algorytmu zależy w dużej mierze od problemu, który należy rozwiązać. Na potrzeby tematu można przyjąć uproszczenie, że zadania dyskretne (klasyfikacyjne), regresyjne/predykcyjne realizuje uczenie nadzorowane.

Uczenie nadzorowane

Uczenie nadzorowane stosowane jest do rozwiązywania problemów podjęcia akcji, wykonania predykcji lub podjęcia decyzji z pewnej przestrzeni decyzji. Do oceny wydajności używana jest funkcja straty, która określa koszt konkretnej predykcji względem najlepszej dostępnej (bądź poprawnej).



RYS. 1 Schemat uczenia nadzorowanego *(opracowanie własne)*

Rysunek przedstawia schemat uczenia maszynowego. Dane wejściowe, służące wytworzeniu modelu, zawierają przykłady (wiersze), które posiadają cechy (kolumny) oraz etykiety. Z pomocą algorytmu optymalizacyjnego i funkcji straty w procesie uczenia nadzorowanego uzyskuje się model, który mapuje nowe przykłady (w oparciu o ich cechy) do etykiet.

Zbiór algorytmów optymalizacyjnych jest liczny, a dziesiątki z algorytmów dostępne są w popularnych bilbiotekach wykorzystywanych do uczenia maszynowego. Na potrzeby realizacji tematu opisana zostanie metoda gradientu prostego (ang. gradient descent). Charakterystyka algorytmów z tej rodziny używana jest powszechnie w antagonistycznym uczeniu maszynowym. Prezentacja sposobu działania tego algorytmu jest istotna do zrozumienia zdecydowanej większości ataków. Dla lepszego zobrazowania zależności od zbioru parametrów modelu $\theta \in \mathbb{R}^d$, gdzie $d=n+1$ dla n, gdzie dla będącego liczbą kardynalną zbioru cech dowolnego przykładu $x^{(i)}$, przekształca się hipotezę do postaci:

$$f(x)=g(\theta^T x)$$

Funkcja aktywująca oznaczona jest jako g. Może to być na przykład funkcja sigmoidalna $(g(z) = \frac{1}{1+e^{-z}})$ w przypadku regresji logistycznej. Na podstawie powyższego wyprowadzono wzór funkcji kosztu J, jako funkcji algebraicznej zależnej od θ opisaną poniżej.

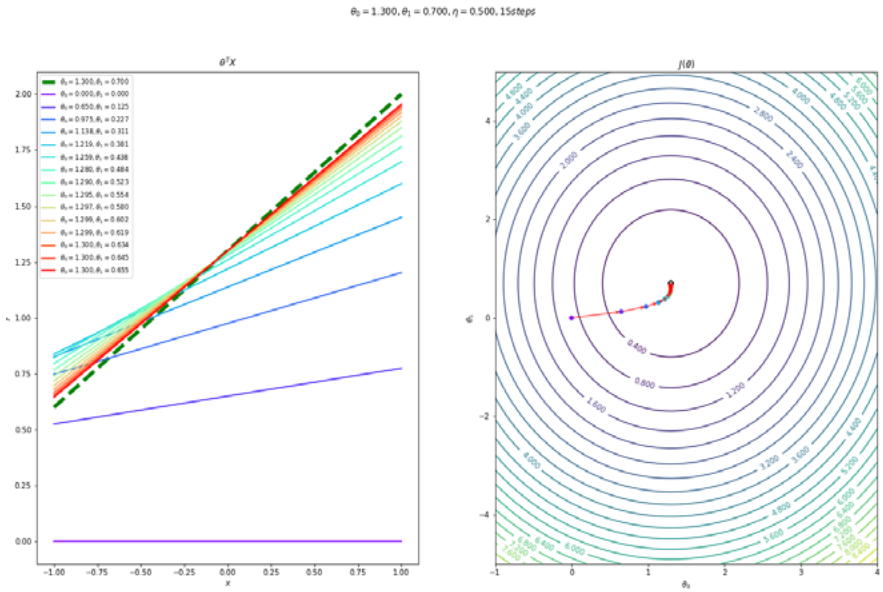
$$J(\theta) = \frac{1}{|\mathbb{D}|} \ell\left(y, g(\theta^T x)\right)$$

Przekształcenia te pozwalają zapisać ostatecznie sposób wyznaczenia θ w danej iteracji gradientu prostego, wykorzystując dodatkowo zmienną η parametryzującą wielkość kroku w kierunku lokalnego minimum.

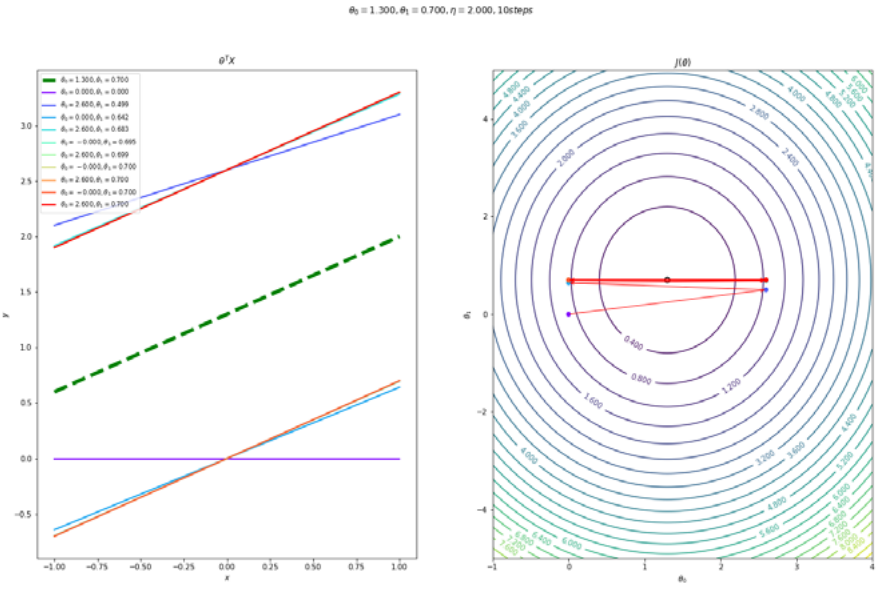
$$\theta = \theta - \eta \cdot \frac{\partial}{\partial \theta} J(\theta)$$

Metoda gradientu prostego gwarantuje (zakładając poprawnie dopasowany parametr η) konwergencję w globalnym minimum w przypadku wypukłej funkcji straty oraz w lokalnym minimum w przeciwnym wypadku.

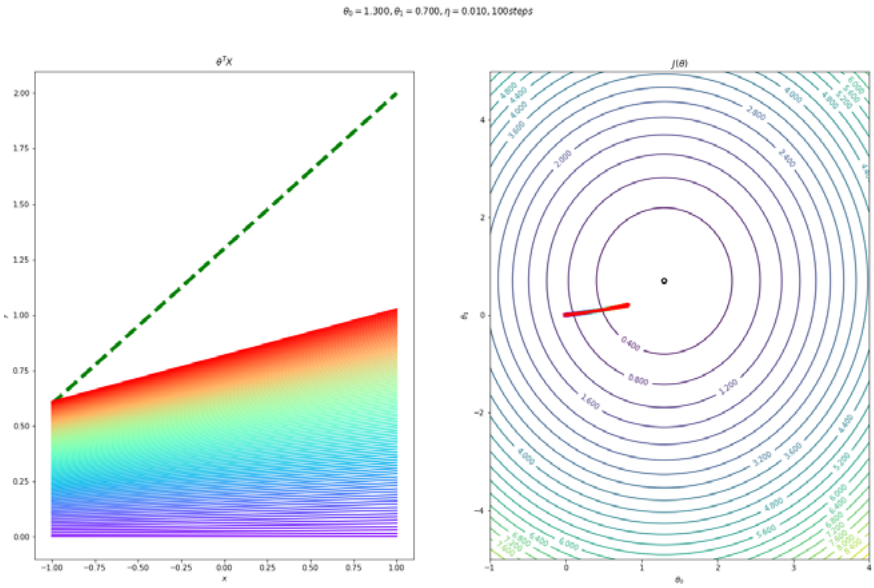
Na rysunku zaprezentowane zostało zachowanie gradientu prostego w przypadku prostej funkcji liniowej i różnych wariancji parametru η. Wariant pierwszy przedstawia funkcję z rozmiarem kroku uczenia, który gwarantuje konwergencję w stosunkowo niewielu krokach. Kolejny wariant przedstawia dobór zbyt wysokiego kroku uczenia, który powoduje, że zejście w kierunku minimum staje się bardzo trudne, a w niektórych przypadkach wręcz niemożliwe. Skutkiem tego jest nieosiągnięcie konwergencji. Ostatni przykład z bardzo niskim krokiem uczenia w relacji do skali parametrów pokazuje, że konwergencja jest możliwa, ale wymaga wielokrotnie więcej kroków do osiągnięcia tego celu.



RYS. 2. Metoda gradientu prostego dla funkcji liniowej. Wizualizacja otrzymanej funkcji oraz uzyskanego gradientu w każdej iteracji. Wykonanie dla parametrów pozwalających na poprawne wykonanie z konwergencją *(opracowanie własne)*



RYS. 3. Metoda gradientu prostego dla funkcji liniowej. Wizualizacja otrzymanej funkcji oraz uzyskanego gradientu w każdej iteracji. Wykonanie dla parametrów niepozwalających na uzyskanie konwergencji *(opracowanie własne)*



RYS. 4. Metoda gradientu prostego dla funkcji liniowej. Wizualizacja otrzymanej funkcji oraz uzyskanego gradientu w każdej iteracji. Wykonanie dla parametrów, które pozwolą na konwergencję, ale znacząco opóźnią otrzymanie wyniku *(opracowanie własne)*

Problematyka antagonistycznego uczenia maszynowego

Dynamicznie rozwijające się technologie, jak uczenie maszynowe, narażone są na asymetrię pomiędzy wzrostem świadomości o problemach związanych z bezpieczeństwem względem postępu wykonanego w ramach danej dziedziny. Pierwsze artykuły, które nawiązały do tematyki antagonistycznego uczenia maszynowego opublikowano w roku 2004 [2]. Bezpieczeństwo sieci głębokich zostało opisane dopiero na przełomie 2014 i 2015 roku [3] [4]. W ostatnich latach liczba publikacji związana z problematyką bezpieczeństwa modeli uczenia maszynowego wyraźnie wzrosła, a nowe artykuły i wariancje są już codziennością. Większość zaproponowanych ataków posiada wspólną charakterystykę i polega na podążaniu za gradientem w celu optymalizacji perturbacji. Mimo ogromnej płodności naukowców problemy znane od lat nie doczekały się skutecznych systemów obrony, co pokazuje złożoność problematyki. Obecnie znane typy ataków mogą prowadzić między innymi do utraty prywatności

lub do podjęcia przez systemy autonomiczne błędnej decyzji, która może obciążyć swoją ofiarę nie tylko finansowo, ale także narazić jej zdrowie lub życie [5].

Atak może nastąpić podczas każdej z faz procesu uczenia maszynowego, zaczynając od fazy pomiarowej i preparacji własnych danych, poprzez fazę doboru parametrów, wyboru modelu, fazy treningowej aż po fazę predykcji. Wspomnienia warta jest również metoda, która może wykraczać poza definicję samego antagonistycznego uczenia. Ataki te nazywane są atakami gąbki (ang. *sponge attack*) i wykorzystują wysoką złożoność obliczeniową procesów związanych z modelami uczenia maszynowego. Opierają się one na preparowaniu przykładów antagonistycznych, których celem nie jest błędna klasyfikacja, a nadmierne obciążenie systemów wykonujących komputacje na rzecz atakowanego podmiotu [6].

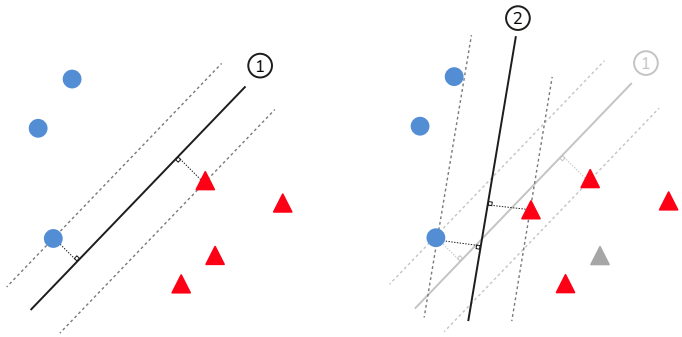
Przegląd wybranych metod ataków na algorytmy uczenia maszynowego

W ramach uczenia maszynowego spośród głównych typów ataków na modele możemy wyróżnić zatrucie (ang. *poisoning*) i uniknięcie (ang. *evasion*), które opisane zostały w pracy. W literaturze oczywiście znajdziemy i inne metody, jak na przykład ataki inżynierii wstecznej (ang. *model stealing/model reverse engineering*) czy ekstrakcji/aproksymacji modeli.

ATAKI ZATRUCIA

Ataki zatrucia to typ ataków, w których w sposób celowy dostarcza się zmanipulowane dane do zbiorów wykorzystywanych w procesie nauki modelu. Ich wynikiem jest doprowadzenie do stanu, w którym model uczy się nienaturalnych – spreparowanych – schematów atakującego. Każdy system, który pobiera dane służące wytrenowaniu modelu z niezaufanych źródeł, jest podatny na tę technikę. Szczególnie wrażliwe są systemy uczące się metodą strumieniową, starające się nadążyć za zmianami zachowania użytkowników. Funkcje regresji oraz te wykorzystujące maszynę wektorów nośnych można w sposób znaczący degradować nawet pojedynczymi przykładami antagonistycznymi. Poniżej zilustrowano jak

indywidualne, w teorii nieznaczne i niemodyfikujące pierwotnej etykiety, perturbacje mogą istotnie zniekształcić granicę decyzji modelu.

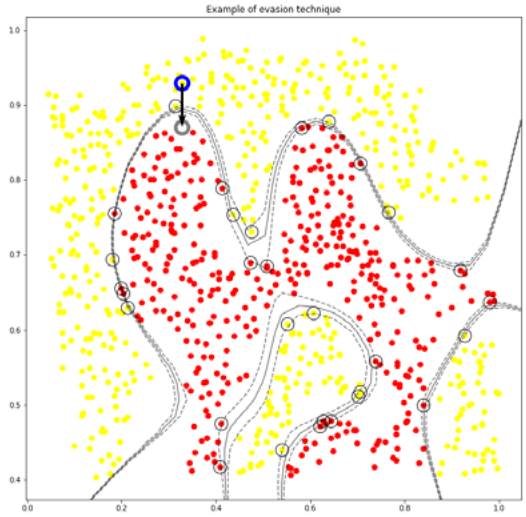


RYS. 5. Zmiana granicy modelu SVM w wyniku zmiany cech pojedynczego wpisu bez zmiany jego oryginalnej etykiety: (1) – granica przed zmianą, (2) – granica po zmianie [7]

ATAKI UNIKU

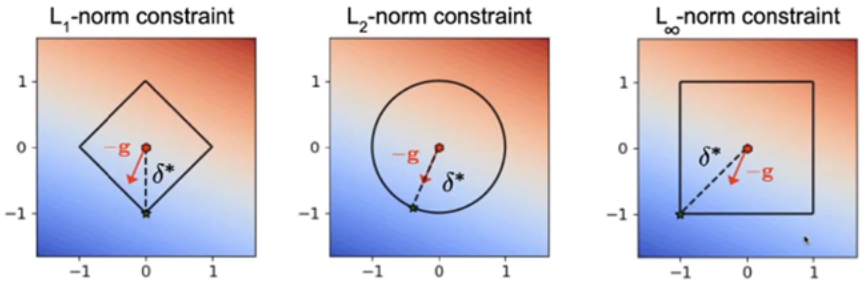
Najistotniejszą rodziną ataków są ataki uniku (ang. *evasion attacks*). Ich celem jest wpłynięcie na decyzję podejmowaną przez atakowany model. Oczekiwanym wynikiem jest uzyskanie lub uniknięcie wskazanego wyboru konkretnej klasyfikacji modelu nadzorowanego.

Ataki uniku, ze względu na znajomość modelu przez atakującego, dzielimy na *white-box* oraz *black-box*. Ataki *white-box*, czyli białej skrzynki, to ataki na modele, do których atakujący ma pełen dostęp i zna ich parametry. Ataki *black-box*, czyli czarnej skrzynki, to ataki na modele, do których badacze nie posiadają pełnego dostępu. Mogą się one charakteryzować dodatkowymi obostrzeniami, jak ograniczona dostępność informacji zwrotnych lub limitowana liczba zapytań, na które cel odpowie w danym przedziale czasowym.



RYS. 6. Przykład obrazujący wykonanie techniki uniku na modelu SVM. Punkt zaklasyfikowany pierwotnie jako żółta etykieta (niebieski pierścień) po dodaniu perturbacji zostaje zaklasyfikowany jako etykieta czerwona *(opracowanie własne)*

Ze względu na wykorzystywane techniki ataki można podzielić na gradientowe (wykorzystujące różniczkowalność modelu atakowanego) lub niegradientowe. Ataki gradientowe wykorzystują charakterystykę procesu uczenia, by poznać cechy modelu pozwalające na wykonanie uniku. W zależności od przyjętych norm w wyliczaniu perturbacji mogą być one rozproszone (ang. *sparse*) lub gęste (ang. *dense*) [8].



RYS. 7. Wizualizacja granic norm p (dla p =1, p = 2 oraz p = ∞) [3]

Ataki rozproszone dokonują zmian w ramach większej przestrzeni, a skoncentrowane – w mniejszej. W teorii oznacza to, że jedna metoda pozwala wykonać bardzo wiele małych perturbacji, a druga ograniczoną liczbę stosunkowo wysokich perturbacji do osiągnięcia tego samego celu.

Generując atak uniku wyszukiwany jest wektor o najmniejszym dystansie (ang. *minimum distance*) lub o maksymalnej pewności (ang. *maximum confidence*). Podziały te wynikają z precyzyjnie określonych ograniczeń. Ataki minimalnego dystansu minimalizują perturbację w odniesieniu do normy l_p, traktując wynik jako ograniczenie ($\min\|\delta\|_p \ni L(x+\delta,y;\theta)<t$, gdzie t to graniczna wartość wyniku). Przykładem takich metod jest *Probabilistic Jacobian-based Saliency Maps Attacks* (JSMA). Ataki maksymalnej pewności minimalizują funkcję straty i traktują jako ograniczenie perturbację ($\min L(x+\delta,y;\theta) \ni \|\delta\|_p<\epsilon$, gdzie ϵ to wartość graniczna perturbacji). Sposób ten wykorzystują między innymi ataki *Fast Gradient Sign Method* (FGSM) oraz *Projected Gradient Descent* (PGD). Technika łącząca te podejścia traktuje te ograniczenia jako parametry funkcji. Ich współczynniki podlegają zmianie w poszukiwaniu równowagi pomiędzy dystansem a wynikiem ($\min L(x+\delta,y;\theta)+ c\|\delta\|_p$). Przykładem takiego zrównoważonego podejścia jest metoda *Carlini-Wagner* [8].

Nieznajomość atakowanego modelu stanowi jeden z elementów jego bezpieczeństwa. Pewne modele używają nieróżniczkowalnych funkcji, a część nie jest dostępna dla badaczy (np. modele świadczone jako usługa), stanowiąc tym samym czarne skrzynki (ang. *black-box*). Większość metod uniku wymaga przynajmniej częściowej znajomości modelu do poprawnego wykonania i wykorzystuje zejście gradientowe jako metodę ataku. Naturalną potrzebą, z perspektywy bezpieczeństwa, stało się zatem badanie wektora ataku związanego z inżynierią wsteczną samego modelu. Prostą, lecz skuteczną metodą jest tworzenie modelu zastępczego (ang. *surrogate model*). Model taki służy wykonaniu na nim ataku uniku, a następnie transferu rezultatów na inny model (np. pierwotny). Wiele metod tworzenia modelu zastępczego wymaga wysokiej interakcji z oryginalnym modelem i prowadzi do jego aproksymacji, która może nie być wystarczająca do uzyskania oczekiwanych rezultatów. Wraz z potrzebami badaczy pojawiły się nowe modele ataku, które aproksymowały gradient modelu z pomocą próbek poddanych losowej perturbacji lub próbowały wykorzystać różniczkowalność rozkładu normalnego nakładanego na obraz wejściowy.

Ataki typu *black-box* dzielą się ze względu na rodzaj informacji, który zwraca atakowany model. Ograniczenia występują w postaci liczby zapytań do klasyfikatora (ang. *query-limited*), liczby najwyższych etykiet i ich wyników (ang. *partial-information*) lub etykiet k-najlepiej dopasowanych wyników (ang. *label-only*). W skrajnym przypadku, gdy k=1, atakujący otrzymuje jedynie zaklasyfikowaną etykietę bez dodatkowych danych [9].

Przykład ataku uniku

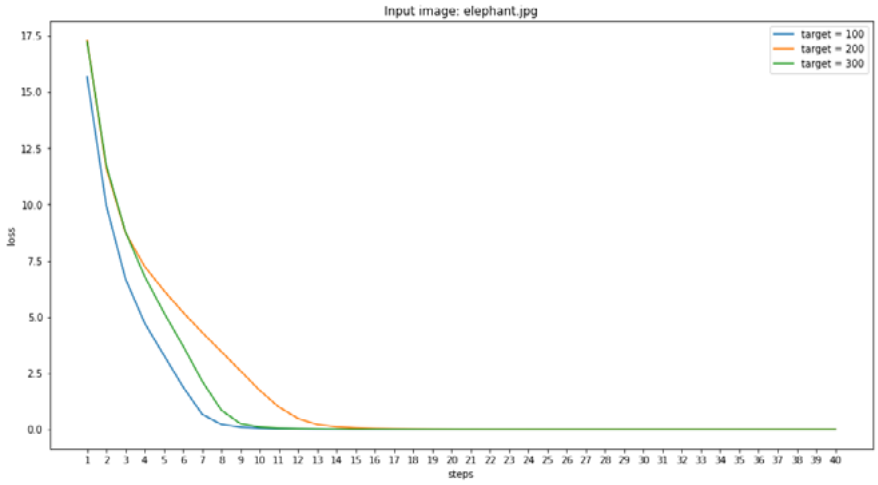
W ramach przykładu ataku uniku na obrazy wyższej rozdzielczości jako cel ataku użyto gotowego modelu ResNet50, uzyskanego w drodze uczenia nadzorowanego. Jest to rozbudowany model typu CNN (ang. *convolution neural network*) wykorzystujący, jak wskazuje nazwa, 50 warstw do przetwarzania danych wejściowych. Łączna liczba parametrów modelu przekracza 25 milionów. W ataku użyto obrazy tygrysa i słonia afrykańskiego znalezione na podstawie słów kluczowych z wykorzystaniem wyszukiwarki Google.

```
Total params: 25,636,712
Trainable params: 25,583,592
Non-trainable params: 53,120
```

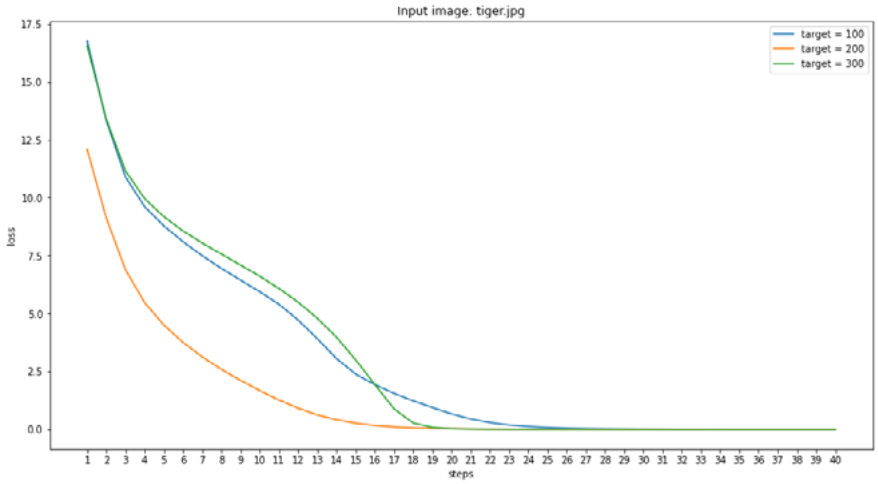
RYS. 8. Podsumowanie parametrów modelu ResNet50 użytego w przykładach ataków uniku *(opracowanie własne)*

Z założenia, ataki przeprowadzone na modelach o wysokiej liczbie parametrów, pozwalają na mniej inwazyjne dla ludzkiego oka perturbacje. Im większa przestrzeń, tym stosunkowo mniejsza perturbacja potrzebna będzie do przeprowadzenia ataku uniku, który nie zostanie wykryty bez wykonania analizy obrazu specjalistycznymi narzędziami.

Do wykonania ataku użyta została metoda I-FGM (ang. *iterative fast gradient method)*. Atak wykonany jest trzykrotnie dla każdego obrazu, za każdym razem celowany w etykiety – odpowiednio – 100 (*black_swan*), 200 (*tibetan_terrier*), 300 (*tiger_beetle*). Algorytm w każdym uruchomieniu wykonuje 40 iteracji. Na rysunku przedstawiono wgląd w funkcję straty.



RYS. 9. Relacja wyniku funkcji straty względem kolejnych iteracji dla etykiet w przykładzie ataku uniku na obraz słonia afrykańskiego w modelu ResNet50 *(opracowanie własne)*



RYS. 10. Relacja wyniku funkcji straty względem kolejnych iteracji dla etykiet w przykładzie ataku uniku na obraz tygrysa w modelu ResNet50 *(opracowanie własne)*

RYS. 11. Przykład wykonania ataku uniku na obrazach słonia afrykańskiego i tygrysa metodą I-FGM na modelu ResNet50. Kolejne pary wierszy zawierają oryginalny obraz z nałożonym szumem oraz poniżej szum powiększony przez wartość epsilon na potrzeby prezentacji graficznej *(opracowanie własne)*

Mimo niewielkich perturbacji, na poziomie prawie każdego piksela obrazu, zmiany są niedostrzegalne dla obserwatora. Przestrzeń objęta modyfikacją pozwala na zmiany, które nie wpływają na percepcję człowieka. Algorytm decyzyjny z prawdopodobieństwem graniczącym z pewnością klasyfikuje obrazy do wskazanych przez atakującego etykiet.

Przykład ten pokazuje jak w niezauważalny dla obserwatora sposób antagonistyczne przykłady mogą całkowicie wprowadzić w błąd klasyfikator. W rzeczywistości podobnie zachowują się modele tekstowe, audio czy

wideo. Sam atak może mieć również formę fizyczną i nie musi ograniczać się do przebrania, a do posiadania ze sobą odpowiednio spreparowanego przedmiotu [10].

Bibliografia

[1] T. Mitchell, *Machine Learning Textbook*. New York: McGraw Hill, 1997.

[2] N. N. Dalvi i in., „Adversarial classification", w: *Proceedings of the tenth ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*, Long Beach, CA, 2004, pp. 99-108.

[3] Ch. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, „Intriguing properties of neural networks", arXiv:1312.6199v4, 2014. Dostęp: https://arxiv.org/abs/1312.6199.

[4] I. J. Goodfellow, J. Shlens, C. Szegedy, „Explaining And Harnessing Adversarial Examples", in *ICLR*, San Diego, CA, 2015. Dostęp: https://arxiv.org/abs/1412.6572.

[5] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, „Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks", in *37th IEEE Symposium on Security and Privacy*, San Jose, 2016, pp. 582-597. Dostęp: https://arxiv.org/abs/1511.04508.

[6] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, R. Anderson, „Sponge Examples: Energy-Latency Attacks On Neural Networks", in *6th IEEE European Symposium on Security and Privacy*, Vienna, 2021. Dostęp: https://arxiv.org/abs/2006.03463.

[7] D. J. Miller, Zh. Xiang, G. Kesidis, „Adversarial Learning in Statistical Classification: A Comprehensive Review of Defenses Against Attacks", in *Proceedings of the IEEE*, vol. 108(3), pp. 402-433, 2020. Dostęp: https://arxiv.org/abs/1904.06292.

[8] N. Carlini, D. Wagner, „Towards Evaluating the Robustness of Neural Networks", in *38th IEEE Symposium on Security and Privacy*, San Jose, CA, 2017, pp. 39-57. Dostęp: https://arxiv.org/abs/1608.04644.

[9] A. Ilyas, L. Engstrom, A. Athalye, J. Lin, „Black-box Adversarial Attacks with Limited Queries and Information", in *ICML*, 2018. Dostęp: https://arxiv.org/abs/1804.08598.

[10] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, „Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition", in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 2016, pp. 1528–1540.

Kilka uwag o cyberbezpieczeństwie medycznej AI

dr Jarosław Greser

Politechnika Warszawska | Wydział Administracji i Nauk Społecznych

Wstęp

Możliwości wykorzystania sztucznej inteligencji w medycynie są ogromne. Ujawniają się one szczególnie w implementacji środków opartych na uczeniu maszynowym (*machine learning*), które stanowią znaczną część produktów wdrażanych w ostatnich latach. Jako przykłady można przywołać rozwiązania wskazujące najkorzystniejszy algorytm leczenia pacjenta na podstawie analizy dokumentacji medycznej [1], analizujące obrazy uzyskane z aparatury diagnostycznej [2], czy predykcji pogorszenia się stanu zdrowia pacjentów hospitalizowanych [3]. Skalę zjawiska pokazują środki przeznaczone na inwestycje w tym obszarze, które obecnie szacuje się na 4,4 miliarda dolarów w 2022 roku, a prognozuje się ich wzrost do 6.2 miliarda dolarów w 2023 roku [4].

W niniejszym artykule przedstawione zostaną wyzwania związane z tą technologią z perspektywy cyberzagrożeń oraz działania regulacyjne Unii Europejskiej mające na celu zwiększenie cyberbezpieczeństwa medycznej AI.

Cyberzagrożenia medycznej sztucznej inteligencji

Jak każda technologia ICT, sztuczna inteligencja jest podatna na cyberzagrożenia. Z perspektywy przedmiotowej można je podzielić na trzy grupy. Pierwszą stanowią zagrożenia, na które narażone jest każde rozwiązanie funkcjonujące w świecie cyfrowym. Drugą są podatności specyficzne dla sztucznej inteligencji, a trzecią – ryzyka związane ze środowiskiem, w jakim algorytm jest używany.

W pierwszym wypadku będzie to dotyczyło najczęściej ataków na urządzenia, na których algorytm AI jest zainstalowany albo na jego użytkowników [5]. W tym obszarze mamy do czynienia z całą gamą ataków, a ich dobór zależy od celów, które atakujący chce osiągnąć. Mogą wśród nich znaleźć się: kradzież informacji, uniemożliwienie uprawnionym użytkownikom dostępu do danych lub ich nieuprawniona zmiana [6].

Druga grupa obejmuje zagrożenia, które są właściwe dla algorytmów sztucznej inteligencji i mogą nie występować w innych rozwiązaniach lub być możliwe do zastosowania tylko w ograniczonym zakresie. Raport Agencji Unii Europejskiej ds. Cyberbezpieczeństwa wymienia kilkadziesiąt zagrożeń sklasyfikowanych w osiem głównych obszarów [7]. Dotyczą one zarówno ataków na komponenty algorytmu, jak i dane wykorzystywane do procesu jego uczenia. Zwraca się uwagę, że ataki na dane, a szczególnie atak typu *poison pill* i atak adwersarski są szczególnie groźne ze względu na możliwość wystąpienia niezależnie od architektury [7], a także problemy z ich wykryciem przez człowieka [8]. Ponadto na możliwość ich przeprowadzenia nie wpływa zaimplementowanie narzędzi wyjaśnienia, dlaczego algorytm podjął określoną decyzję [7] lub wykorzystanie do trenowania algorytmu danych syntetycznych [9]. Problem pogłębia zjawisko *automation bias*, które polega na preferowaniu przez człowieka wyników działania algorytmów, mimo że są one niepoprawne [10]. Może się to przekładać na autoryzowanie przez lekarza decyzji systemu sztucznej inteligencji, która została celowo zniekształcona.

Ostatnią grupą zagrożeń jest ekspozycja na ataki związana z zainteresowaniem atakujących służbą zdrowia. Według raportu FBI z 2021 roku sektor ten jest najczęstszym celem ataków typu *ransomware* [11]. W tym samym roku w Polsce odnotowano wzrost zgłoszeń zagrożeń cyberbezpieczeństwa o 167% oraz zarejestrowano o 172% więcej incydentów niż rok wcześniej, a służba zdrowia jest piątym pod względem ilości ataków

sektorem w Polsce [12]. Natomiast raport ENISA wskazuje, że służba zdrowia jest pod stałą presją cyberprzestępców, którzy traktują ją jako cel o wysokim priorytecie [13]. Wskazuje się również, że na poziom cyberbezpieczeństwa miała wpływ pandemia COVID-19, zarówno ze względu na przyspieszenie cyfryzacji wielu usług medycznych, m.in. upowszechnienie telemedycyny [14], jak i zwiększenie się aktywności cyberprzestępców o 600% w stosunku do roku poprzedzającego pandemię [15].

Ataki na medyczną sztuczną inteligencję mogą w skrajnym wypadku doprowadzić do zagrożenia życia lub śmierci pacjenta. Zostały odnotowane cyberataki wywołujące taki skutek [16] [17] [18], choć trzeba podkreślić, że nie były one bezpośrednio związane z atakiem na system sztucznej inteligencji. Trzeba również zwrócić uwagę, że udany atak na algorytm AI używany w medycynie mógłby mieć inne skutki w sferze jednostkowej niż w sferze społecznej. W pierwszym przypadku zafałszowanie wyników badań może być podstawą szantażu lub dyskredytacji określonej osoby. W drugim natomiast może doprowadzić do destabilizacji działania służby zdrowia, co w konsekwencji może wpłynąć na zmniejszenie zaufania do instytucji państwa. Skutek ten może być pożądanym efektem wojny hybrydowej lub operacji wywiadowczej wrogiego państwa.

Regulacja medycznej sztucznej inteligencji

Wskazuje się, że stworzenie odpowiedniego środowiska prawnego jest jednym z warunków osiągnięcia wysokiego poziomu cyberbezpieczeństwa [19]. Trzeba zauważyć, że o ile regulacja sektora medycznego w Unii Europejskiej ma ponad czterdziestoletnią tradycję, to historia horyzontalnych regulacji dotyczących cyberbezpieczeństwa jest bardzo krótka i rozpoczyna się przyjęciem w 2016 roku Dyrektywy NIS [20].

Pojęcie „medyczna sztuczna inteligencja” nie ma swojej definicji legalnej. W Unii Europejskiej podejmowane są działania aby zdefiniować ten termin na poziomie rozporządzenia [21]. Zgodnie z przyjętymi założeniami regulacja ta może również dotyczyć systemów używanych w medycynie, choć zgodnie z motywem (63) projektu Aktu o sztucznej inteligencji [21], zastosowanie jego wymogów nie powinno naruszać logiki zarządzania ryzykiem, która jest przewidziana w regulacjach dotyczących wyrobów medycznych.

Przepisy, o których mowa, opierają się głównie na rozporządzeniu MDR [22] i IVDR [23]. Pozwalają one na zakwalifikowanie sztucznej inteligencji jako wyrobu medycznego i to zarówno w sytuacji, gdy będzie ona, wbudowana w wyrób, współpracować z urządzeniem, np. analizować dane przekazywane przez urządzenie lub będzie funkcjonowała wyłącznie w postaci cyfrowej [24]. W przypadku produktów, które mogą być stosowane zarówno do celów medycznych, jak i niemedycznych, wymagane jest łączne spełnienie wymogów przewidzianych dla obu kategorii.

Jednocześnie producent nie ma obowiązku zgłaszania danego urządzenia jako wyrobu medycznego. Konsekwencją tego jest brak możliwości posługiwania się produktem w obrocie profesjonalnym, który w Polsce wynika z art. 17 ust. 1 pkt 2 i art. 18 ust. 1 pkt. 3 ustawy z dnia 15.04.2011 r. o działalności leczniczej [25]. Co nie oznacza, że dane rozwiązanie techniczne nie może być dostępne na rynku. W przypadku rozwiązań opartych na sztucznej inteligencji dotyczy to całego spektrum urządzeń typu *wearables*, w których algorytmy AI analizują dane dotyczące zachowań użytkownika czy aplikacji prozdrowotnych, w tym dotyczących tak wrażliwych obszarów jak wsparcie w chorobach psychicznych. Funkcjonują one jako produkty konsumenckie, co do których wymagania regulacyjne są duże niższe lub żadne w porównaniu z wyrobem medycznym. Część wytwórców świadomie decyduje się na taki zabieg unikając kosztów procedury certyfikacyjnej i oszczędzając czas, który trzeba na nią poświęcić. Odbywa się to kosztem przynajmniej potencjalnego zwiększenia ryzyka używania takiego rozwiązania, w tym również w sferze cyberbezpieczeństwa.

Tym samym na rynku mamy dwie grupy rozwiązań, które mogą być nazwane medyczną sztuczną inteligencją: wyroby medyczne i rozwiązania konsumenckie. Wkrótce dołączy do nich trzecia grupa rozwiązań podlegających pod regulacje AI Act. Ich rozróżnienie i prawidłowe sklasyfikowanie jest kluczowe z perspektywy określenia, jakie prawne wymogi w zakresie cyberbezpieczeństwa będą miały do nich zastosowanie.

W przypadku wyrobów medycznych, reguł dotyczących cyberbezpieczeństwa należy poszukiwać w MDR i IVDR. Obowiązki w tym zakresie możemy interpretować według ogólnej zasady, która wymaga bezpieczeństwa produktu: art. 5 ust. 1 MDR i art. 5 ust. 1 IVDR oraz regulacji dotyczących systemów informatycznych stosowanych w wyrobach medycznych tj. art. 17.2, 17.4, i 18.8 załącznika nr 1 do MDR [25]. Oba akty

zawierają również szczegółowe zasady dotyczące nadzoru po wprowadzeniu do obrotu, co obejmuje obowiązek tworzenia planów nadzoru po wprowadzeniu do obrotu, okresowych raportów aktualizacyjnych dotyczących bezpieczeństwa, analiz i zgłaszania poważnych incydentów oraz raportowania trendów. Zgodnie z wytycznymi *Medical Device Coordination Group* nadzór po wprowadzeniu do obrotu obejmuje również zagrożenia i incydenty spowodowane cyberatakami [26]. Można więc uznać, że regulacje dotyczące wyrobów medycznych odnoszą się do kwestii związanych z cyberbezpieczeństwem. Dyskusyjne jest, czy kwestie te powinny być poruszane bezpośrednio i czy wymagania są wystarczające. Niemniej jednak należy zauważyć, że stanowią one kompleksową regulację, która musi być stosowana w przypadku algorytmów sztucznej inteligencji klasyfikowanej jako wyrób medyczny.

W przypadku rozwiązań, które nie będą klasyfikowane jako wyroby medyczne, regulacji trzeba poszukiwać w aktach o charakterze horyzontalnym i sektorowym. Należą do nich przepisy dotyczące ochrony danych osobowych, w szczególności art. 5, 24 i 32 RODO [27]. W przypadku dyrektywy NIS i niedawno przyjętej dyrektywy NIS 2 [28] można zauważyć, że skupiają się one raczej na budowaniu środowiska wspierającego cyberbezpieczeństwo, niż na nakładaniu bezpośrednich obowiązków na producentów rozwiązań [28]. Pewien wpływ mogą mieć regulacje dotyczące bezpieczeństwa produktów z komponentem cyfrowym, w szczególności rozporządzenie wykonawcze Komisji Europejskiej 2022/30 [29] regulujące wymogi dotyczące cyberbezpieczeństwa w dużej części urządzeń Internetu Rzeczy oraz projektowane rozwiązania podnoszące cyberbezpieczeństwo produktów z elementami cyfrowymi i określające wymogi bezpieczeństwa produktów w ogólności [30].

Regulacja zawarta w art. 15 AI Act przewiduje, że systemy sztucznej inteligencji wysokiego ryzyka projektuje się i opracowuje się w taki sposób, aby osiągały, z uwagi na ich przeznaczenie, odpowiedni poziom dokładności, solidności i cyberbezpieczeństwa oraz działały konsekwentnie pod tymi względami w całym cyklu życia. Ponadto, zgodnie z art. 15 ust. 4 systemy takie muszą być odporne na próby nieupoważnionych osób trzecich, mające na celu zmianę ich zastosowania lub skuteczności działania poprzez wykorzystanie słabych punktów systemu, w tym w szczególności ataki adwersarskie i typu *poison pill.* Trzeba jednak zauważyć, że regulacje te mają bardzo ograniczony zakres zastosowania w przypadku systemów AI, które nie są wyrobami medycznymi. Zgodnie z art. 6,

klasyfikacja do systemów wysokiego ryzyka jest bardzo sformalizowana i opiera się o unijne prawodawstwo harmonizacyjne, które co do zasady nie będzie obejmować najpopularniejszych rozwiązań konsumenckich, takich jak aplikacje poprawiające jakość życia, rozwiązania do inteligentnych domów czy *wearables*.

Podsumowanie

Nie ma wątpliwości, że zastosowania AI w medycynie będą obejmować coraz szersze obszary. Jednocześnie wraz ze zwiększeniem ilości wdrożonych rozwiązań wzrasta ilość wektorów ataku na taki system. Regulacje prawne mogą przyczynić się do podniesienia bezpieczeństwa stosowanych rozwiązań przy założeniu, że będą one kompleksowo regulować rynek. W przypadku medycznej sztucznej inteligencji taka sytuacja nie ma miejsca. Przepisy dotyczące wyrobów medycznych nie mają wymuszającego charakteru. Oczywiście, istniejący w wielu krajach nakaz korzystania z wyrobów medycznych przez użytkowników profesjonalnych i instytucje zdrowia publicznego ogranicza takie zastosowanie. Jednocześnie są sfery, w których produkty niemające statusu wyrobu medycznego zajmują pokaźną część rynku i są powszechnie stosowane przez pacjentów do celów diagnozy lub leczenia. Przykładem mogą być wszelkiego rodzaju aplikacje zdrowotne. Z perspektywy regulacji cyberbezpieczeństwa jest to sytuacja bardzo niepożądana i *de facto* powoduje istnienie dwóch reżimów regulacyjnych do rozwiązań o takim samym lub podobnym charakterze.

W przypadku AI, która jest wyrobem medycznym, można uznać, że wymogi dotyczące cyberbezpieczeństwa są sformułowane i co do zasady użytkownicy takich rozwiązań są zabezpieczeni na podstawowym poziomie. W przypadku innych rozwiązań, wymogów dotyczących cyberbezpieczeństwa należy poszukiwać w przepisach dotyczących sfer takich jak ochrona danych osobowych lub prawa konsumenta. Sytuację może odmienić przyjęcie Aktu o sztucznej inteligencji. Trzeba natomiast zauważyć, że w wielu przypadkach będzie miał on ograniczone zastosowanie, ponieważ aplikacje nie będą wpisywać się w rozwiązania wysokiego ryzyka, które są w głównej mierze przedmiotem regulacji.

Bibliografia

[1] IBM, *Clinical Decision Support* [Online]. Dostęp: https://www.ibm.com/watson-health/solutions/clinical-decision-support

[2] D. W. Kim, H. Y. Jang, K. W. Kim, Y. Shin, S. H. Park, „Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers", *Korean Journal of Radiology* vol.3, pp. 405–410, 2019.

[3] G. Escobar, B. Turk, A. Ragins, J. Ha, B. Hoberman, S. LeVine, M. Ballesca, V. Liu, P. Kipnis, „Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals", *Journal of Hospital Medicine*, vol. 11 pp.18-24, 2016.

[4] W. Wong (29 grudnia 2022). *AI Trends in Health Care* [Online]. Dostęp: https://aibusiness.com/verticals/2023-ai-trends-in-health-care-

[5] W. Nowak, „Specyfika zagrożeń w cyberprzestrzeni", w: *Cyberbezpieczeństwo*, C. Banasiński, M. Rojszczak, Eds. Warszawa: Wolters Kluwer, 2020.

[6] M. Rojszczak, „Wybrane problemy cyberbezpieczeństwa w ochronie zdrowia", in *Jakość w opiece zdrowotnej. Zastosowanie nowoczesnych technologii w czasie pandemii.* K. Kokocińska, J. Greser, Eds. Warszawa: Wolters Kluwer, 2021.

[7] ENISA (15 grudnia 2020). „AI Cybersecurity Challenges. Threat Landscape for Artificial Intelligence" [Online]. Dostęp: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges/@@download/fullReport.

[8] I. Goodfellow, J. Shlens, C. Szegedy, (20 marca 2015). „Explaining and Harnessing Adversarial Examples" [Online]. Dostęp: https://arxiv.org/abs/1412.6572.

[9] J. Ive (24 października 2022). „Leveraging the potential of synthetic text for AI in mental healthcare", *Front. Digit. Health*, [Online] vol. 4. Dostęp: https://www.frontiersin.org/articles/10.3389/fdgth.2022.1010202/full

[10] I. Straw (listopad 2020), „The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future" [Online], Artif Intell Med. Dostęp: https://pubmed.ncbi.nlm.nih.gov/33250145/

[11] „FBI Internet Crime Report 2021„ [Online]. Dostęp: https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf

[12] „Raport roczny z działalności CERT Polska 2021" [Online]. Dostęp: https://cert.pl/uploads/docs/Raport_CP_2021.pdf;

[13] ENISA (3 listopada 2022). „ENISA Threat Landscape 2022" [Online]. Dostęp: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

[14] A. Jackowska (grudzień 2022). *Konsultacje online zmieniły ochronę zdrowia (dane OECD)* [Online]. Dostęp: https://cowzdrowiu.pl/aktualnosci/post/konsultacje-online-zmienily-ochrone-zdrowia-dane-oecd

[15] Cyberdefence (25 maja 2020). *ONZ: podczas pandemii liczba złośliwych e-maili wzrosła o 600 proc.* [Online].

[16] M. Eichelberg, K. Kleber, M. Kämmerer (2020). „Cybersecurity Challenges for PACS and Medical Imaging„, *Academic Radiology* [Online]. vol. 8. Dostęp: https://pubmed.ncbi.nlm.nih.gov/32418786/

[17] S. Gliwa (18 września 2020). *Pierwsza ofiara śmiertelna ataku ransomware. Zarzut nieumyślnego spowodowania śmierci* [Online]. Dostęp: https://cyberdefence24.pl/polityka-i-prawo/pierwsza-ofiara-smiertelna-ataku-ransomware-zarzut-nieumyslnego-spowodowania-smierci

[18] A. Gryszczyńska, „Cyberprzestępczość podczas pandemii", w: *Internet. Cyberpandemia. Cyberpandemic,* A. Gryszczyńska, G. Szpor Eds. Warszawa: CH Beck 2020.

[19] I. Priyadarshini, Ch. Cotton, *Cybersecurity: Ethics, Legal, Risks, and Policies*, Palm Bay: Apple Academic Press 2022.

[20] Dyrektywa Parlamentu Europejskiego i Rady (UE) 2016/1148 z dnia 6 lipca 2016 r. w sprawie środków na rzecz wysokiego wspólnego poziomu bezpieczeństwa sieci i systemów informatycznych na terytorium Unii.

[21] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts, COM(2021) 206 final, podejście ogólne Rady przyjęte 6 grudnia 2022 [Online]. Dostęp: https://data.consilium.europa.eu/doc/document/ST-15698-2022-INIT/en/pdf

[22] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2017/745 z dnia 5 kwietnia 2017 r. w sprawie wyrobów medycznych, zmiany dyrektywy 2001/83/WE, rozporządzenia (WE) nr 178/2002 i rozporządzenia (WE) nr 1223/2009 oraz uchylenia dyrektyw Rady 90/385/EWG i 93/42/EWG.

[23] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2017/746 z dnia 5 kwietnia 2017 r. w sprawie wyrobów medycznych do diagnostyki in vitro oraz uchylenia dyrektywy 98/79/WE i decyzji Komisji 2010/227/UE.

[24] J. Greser (2020), „Cyberbezpieczeństwo wyrobów medycznych w świetle rozporządzenia 2017/745„ *internetowy Kwartalnik Antymonopolowy i Regulacyjny* [Online]. Vol. 9, issue 2 Dostęp: https://www.researchgate.net/publication/344929652_Cyberbezpieczenstwo_wyrobow_medycznych_w_swietle_rozporzadzenia_2017745

[25] Ustawa z dnia 11.04.2011 r. o działalności leczniczej, Dz. U. z 2018 r. poz. 2190 z późn. zm.

[26] Medical Device Coordination Group (2019). *Guidance on Cybersecurity for medical devices* [Online]. Dostęp: https://health.ec.europa.eu/system/files/202201/md_cybersecurity_en.pdf

[27] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych).

[28] Dyrektywa Parlamentu Europejskiego i Rady (UE) 2022/2555 z dnia 14 grudnia 2022 r. w sprawie środków na rzecz wysokiego wspólnego poziomu cyberbezpieczeństwa na terytorium Unii, zmieniająca rozporządzenie (UE) nr 910/2014 i dyrektywę (UE) 2018/1972 oraz uchylająca dyrektywę (UE) 2016/1148 (dyrektywa NIS 2).

[29] Rozporządzenie Delegowane Komisji (UE) 2022/30 z dnia 29 października 2021 r. uzupełniające dyrektywę Parlamentu Europejskiego i Rady 2014/53/UE w odniesieniu do stosowania zasadniczych wymagań, o których mowa w art. 3 ust. 3 lit. d), e) i f) tej dyrektywy.

[30] Proposal for a Regulation of the European Parliament and of the Council on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council, COM/2021/346 final.

Wykorzystanie sztucznej inteligencji jako zagrożenie dla klientów rynku finansowego

Krzysztof Zieliński, Agata Ślusarek[1]

Urząd Komisji Nadzoru Finansowego | Departament Cyberbezpieczeństwa

Wprowadzenie

Celem artykułu jest wskazanie ryzyk, z jakimi mierzą się instytucje rynku finansowego oraz ich klienci w związku z wykorzystywaniem narzędzi i modeli sztucznej inteligencji (AI) przez zorganizowane grupy cyberprzestępcze.

Implementacja mechanizmów AI do procesów biznesowych organizacji pozwala zwiększyć przewagę rynkową i konkurencyjność. Wykorzystanie rozwiązań AI przez prywatnych użytkowników może przyspieszyć, zoptymalizować i usprawnić działania, skracając czas ich wykonania oraz dostarczając (m.in. za pomocą wbudowanych algorytmów przetwarzania i uczenia danych) precyzyjne, syntetyczne treści np. konkretne informacje z internetu czy opracowania na dany temat. W przypadku grup

1 Poglądy wyrażone w artykule są poglądami osobistymi autorów i nie wyrażają oficjalnego stanowiska instytucji, w której są zatrudnieni.

cyberprzestępczych, w ocenie autorów wykorzystanie AI zwiększa skuteczność wrogich działań, głównie w odniesieniu do klientów indywidualnych rynku finansowego, prowadząc do powiększenia wolumenu strat finansowych.

W ostatnim czasie koncepcja dostępu do narzędzi wykorzystujących uczenie maszynowe [1], powszechnie nazywane sztuczną inteligencją, stała się faktem. Nie jest to nadal „sztuczna inteligencja" w rozumieniu projektowanych przepisów prawa [2], gdyż powinna ona posiadać m.in. takie ceny jak autonomiczność, samodzielność, zdolność do adaptacji i podejmowania decyzji, jednakże na potrzeby niniejszego opracowania autorzy używają tego określenia w szerszym kontekście.

Rozwiązania oparte na uczeniu maszynowym, do niedawna dostępne jedynie dla badaczy uniwersyteckich oraz firm technologicznych posiadających własne działy R&D, stały się udziałem statystycznego użytkownika internetu. Implementacje modeli AI do powszechnie używanych narzędzi, takich jak Github Copilot [3], który w znakomity sposób wspiera pisanie oprogramowania, generujący doskonałej jakości grafiki Midjourney [4], czy też najbardziej rozpowszechniony (w chwili pisania tego artykułu) ChatGPT [5], to dopiero początek zmian cywilizacyjnych i społecznych, jakie czekają nas w przyszłości. Wspomniane narzędzia już teraz pozwalają na optymalizację i zwiększenie wydajności codziennych czynności specjalistów wielu branż, w tym IT – programistów, grafików, analityków czy badaczy bezpieczeństwa.

Każda nowa technologia, tworzona i używana w służbie ludzkości, niesie jednak za sobą, oprócz niewątpliwych korzyści, cały szereg zagrożeń, mogących wpłynąć na sposoby postrzegania i wykorzystywania przestrzeni cyfrowej. Pozostawiając bez komentarza pesymistyczne wizje przejęcia władzy nad światem przez sztuczną inteligencję i stopniowej (lub gwałtownej – według uznania) eliminacji ludzkiej cywilizacji [6] należy zwrócić uwagę, że wspomniana już powszechna dostępność AI została dostrzeżona i zaczyna być wykorzystywana nie tylko jako narzędzie usprawniające prace, ale również jako wsparcie dla działań przestępczych zarówno w obszarze dezinformacji, jak i cyberprzestępczości.

Zagrożenia dla klientów i podmiotów rynku finansowego

Rynek finansowy jest nierozerwalnie związany z technologią i od niej zależny. Można zaryzykować stwierdzenie, że obszar finansów nie może już funkcjonować bez technologicznego zaplecza, gdyż praktycznie wszystkie kluczowe procesy instytucji tego sektora zostały przeniesione do domeny cyfrowej. Pozostałe jeszcze „analogowe" odpowiedniki, umożliwiające realizację powyższych procesów, funkcjonują jedynie w obszarze planów BCP[2], uruchamianych przez instytucje finansowe na wypadek awarii lub niedostępności systemów i technologii cyfrowych.

Filarem omawianego rynku jest klient, zarówno indywidualny, jak i instytucjonalny, któremu zapewnia się dostęp do produktów i usług umożliwiających zarządzanie posiadanymi aktywami. Klienci i instytucje tego rynku od zawsze są celem działań przestępców. Wraz z ewolucją i cyfryzacją usług finansowych, rozwinęła się również przestępczość finansowa w obszarze cyfrowym, początkowo nazywana przestępczością internetową, obecnie określana głównie mianem cyberprzestępczości [7].

Można przyjąć, że najistotniejszym zagrożeniem dla klientów rynku finansowego jest utrata środków lub innych walorów przez nich posiadanych. Może to nastąpić w wyniku działania wielu czynników – ekonomicznych, społecznych czy geopolitycznych. Może to mieć miejsce również w wyniku działalności przestępczej, prowadzącej do zaciągania w imieniu klienta zobowiązań finansowych (kredytów, pożyczek), manipulacji aktywami i walorami finansowymi, czyli działań, które w ostatecznym rozrachunku prowadzą do straty finansowej.

Najpowszechniejszą techniką, wykorzystywaną przez cyberprzestępców do kradzieży środków finansowych, nie są wbrew pozorom zaawansowane technologie, a socjotechnika i manipulacja [8], używane w różnych odmianach phishingu oraz oszustwach na tzw. fałszywe inwestycje [9]. Wynika to z prostej kalkulacji cyberprzestępców prowadzących analizę BCR (*Benefit Cost Ratio*), czyli analizę kosztów i korzyści. Socjotechnika jest skutecznym narzędziem, nie wymaga znaczących nakładów finansowych ani – co najważniejsze – zaawansowanej infrastruktury technicznej, a pozwala osiągnąć założone cele ekonomiczne.

2 *Business Continuity Planning* – plany ciągłości działania (tłum. aut.)

Socjotechnika jest jednak często elementem bardziej zaawansowanych ataków wykorzystujących rozwiązania technologiczne. Dla przykładu *malware* (złośliwe oprogramowanie) na urządzenia mobilne, wykorzystywane m.in. do kradzieży poświadczeń do bankowości elektronicznej (docelowo w celu kradzieży środków finansowych), często dystrybuowane jest właśnie z wykorzystaniem elementów socjotechniki. Kolejna odmiana złośliwego oprogramowania, działająca głównie na komputerach stacjonarnych – tzw. *stealer*, – dystrybuowana jest poprzez maile phishingowe, w których przestępcy podszywają się pod legalnie działające firmy, osoby lub organizacje. Inną metodą jej rozpowszechniania są fałszywe strony internetowe hostujące złośliwą zawartość – udostępniane w internecie zmodyfikowane wersje powszechnie znanych i wykorzystywanych przez użytkowników aplikacji [10]. Również określana jako plaga ostatnich lat działalność cyberprzestępcza z wykorzystaniem *ransomware*, czyli złośliwego oprogramowania, które w przypadku uruchomienia na komputerze ofiary szyfruje dane (zwykle incydent dotyka całej organizacji), jako jedną z metod dystrybucji również używa socjotechniki w postaci np. fałszywych maili ze złośliwym załącznikiem.

Osobną grupą zagrożeń, ukierunkowaną na podmioty rynku finansowego, są działania grup APT (*Advanced Persistent Threat*), czyli grup cyberprzestępczych działających w interesie i na zlecenie aktorów państwowych. TTP (*Tactics, Techniques and Procedures*) tych grup oraz ich cele nie są przedmiotem niniejszego opracowania.

Zagrożeniem, które w ocenie autorów artykułu będzie się dynamicznie rozwijało w najbliższym czasie, będzie *deepfake*, czyli „obraz lub nagranie, które zostało przekonująco zmienione i zmanipulowane" [11].

Technologia *deepfake* – definicja, narzędzia i usługi

CZYM JEST TECHNOLOGIA DEEPFAKE

Termin deepfake to połączenie dwóch sformułowań: deep learning (głębokie uczenie się) i fake (fałszywy, nieprawdziwy). Można przyjąć, że deepfake jest jednym z owoców rozwoju sztucznej inteligencji [12].

Technika ta opiera się na sieciach neuronowych, które analizują duże zestawy próbek danych, aby nauczyć się np. naśladowania mimiki twarzy, manier, głosu i fleksji obiektu. *Deepfake* polega na generowaniu manipulacyjnych treści – statycznych, dynamicznych oraz głosowych – przy wykorzystaniu uczenia maszynowego [13]. Powstanie omawianej technologii datuje się na początek 2017 roku[3], a pierwsze przypadki wykorzystywania *deepfake'ów* skupiały się na tworzeniu zmanipulowanych treści (obrazów i wideo) o treści pornograficznej, z wykorzystaniem wizerunku znanych publicznie osób [15]. Ten rodzaj *deepfake'ów*, nazywany „syntetyczną pornografią", był najpowszechniejszym rodzajem wykorzystania technologii *deepfake* do roku 2019 [16].

PRZEGLĄD NARZĘDZI I USŁUG WYKORZYSTYWANYCH DO TWORZENIA ZMANIPULOWANYCH TREŚCI

Krok milowy w technologii *deepfake* odbył się dzięki rozwojowi narzędzi wykorzystywanych do tworzenia zmanipulowanych treści. Kryje się pod tym rozwój technologiczny i niskie ceny sprzętu komputerowego, a także usług chmurowych, a co za tym idzie zwiększenie możliwości i funkcjonalności coraz bardziej zaawansowanych narzędzi, zmniejszenie progu wejścia do używania narzędzi do tworzenia *deepfake* (zarówno w zakresie potrzebnych umiejętności technicznych, jak i kosztów ich wykorzystania) oraz dostęp do szerokiego spektrum narzędzi.

Obecnie dostępnych jest kilkadziesiąt repozytoriów oprogramowania oferujących kilkaset narzędzi do tworzenia zmanipulowanych treści. Część tych narzędzi jest powielana, modyfikowana w zakresie dodawania kolejnych funkcjonalności i publikowana pod nową nazwą. Taka praktyka w rzeczywistości pozwala tworzyć wiele rozwiązań opartych na tym samym kodzie, z bardzo zbliżonymi możliwościami i powielonymi funkcjonalnościami. Nie zmienia to jednak faktu, że użytkownikom daje to ogrom możliwości w doborze narzędzi. Dużą część tych aplikacji znaleźć można

3 Przykłady narzędzi mogących służyć do tworzenia materiałów *deepfake* publikowane były już w 1997 roku, jednak nie ma dowodów na to by były wykorzystywane do rozpowszechniania zmodyfikowanej treści. Nie posługiwano się również wtedy nazewnictwem *deepfake* [14]

w źródłach ogólnodostępnych, zaś prym wiedzie zdecydowanie repozytorium GitHub[4].

DEEPFAKE WIDEO I OBRAZ

Jednym z pierwszych i najczęściej polecanych narzędzi, jest aplikacja FaceSwap, będąca oprogramowaniem typu *open source*[5], używająca *deep learning* do analizy, a następnie zamiany twarzy na obrazach lub filmach. Do tego celu wykorzystywana jest koncepcja generatywnych sieci przeciwstawnych (GAN)[6], która obejmuje dwie pary złożone z kodera i dekodera. W tej technice parametry koderów są współdzielone [20]. Jego ulepszona wersja, nazwana *FaceSwap-GAN*, jest zdolna do tworzenia realistycznych i spójnych ruchów gałek ocznych. Opiera się ona na koncepcji strat percepcyjnych *VGGFace*, które pomagają poprawić kierunek ruchu gałek ocznych, aby był on bardziej precyzyjny i zgodny z wizerunkiem twarzy źródłowej. Narzędzie to wykorzystuje technikę wykrywania twarzy MTCNN i filtr Kalmana [21], w celu stworzenia bardziej stabilnego obrazu twarzy i jego wygładzenia.

Jednym z częściej używanych narzędzi do tworzenia filmów *deepfake* jest *DeepFaceLab*. Narzędzie to pozwala na zastępowanie twarzy, ale także jej modyfikację (z wersji wejściowej), np. usuwanie oznak starzenia się oraz rzetelną manipulację ruchem warg, sprawiając wrażenie mówienia dopasowanego do treści tekstowej [22].

Częścią wspólną opisanych narzędzi, ale i większości tego typu rozwiązań jest wykorzystywanie koncepcji generatywnych sieci przeciwstawnych (GAN). Dzięki temu aplikacje mają zdolność do wydobywania głębokich

4 GitHub – hostingowy serwis internetowy przeznaczony do projektów programistycznych wykorzystujących system kontroli wersji Git [17]

5 Oprogramowanie *open-source* – rodzaj oprogramowania komputerowego, w którym kod źródłowy jest wydawany na podstawie licencji, a właściciel praw autorskich pozwala użytkownikom na modyfikację i rozpowszechnianie oprogramowania [18]

6 Generatywne sieci przeciwstawne (GAN) – dwie niezależne sieci neuronowe, wykorzystywane do generowania danych podobnych do tych, jakie były trenowane (na pierwszych etapach uczenia maszynowego). Idea GAN polega na tym, że obie sieci ze sobą rywalizują, jednak pojedynek wygrywa jedna z nich [19]

informacji z obrazu. Na etapie kodowania wyodrębnia się mimikę twarzy, następnie analizuje ją i zapamiętuje. Kolejnym krokiem jest dekodowanie zapisanych cech oraz analizowanie możliwości modyfikacji na podstawie zdefiniowanych potrzeb użytkownika. Na tym etapie ujawnia się również działanie dyskryminatora, który służy do podejmowania decyzji o autentyczności cech każdej ze zmodyfikowanych danych. Kopiowanie cech źródłowych do cech docelowych jest najbardziej żmudnym zadaniem w zamianie twarzy i jest wykonywane przez autoenkoder po odpowiednim treningu. Niejawna warstwa wewnątrz kodera jest trenowana do generowania kodu reprezentującego dane wejściowe. Automatyczny koder składa się z dwóch warstw: kodera, który reprezentuje dane wejściowe, i dekodera, który generuje rekonstrukcję. Głównym celem autoenkodera jest radzenie sobie ze stratą MAE (strata rekonstrukcji), stratą przeciwnika (wykonywaną przez dyskryminator) i strat percepcyjnych (które optymalizują podobieństwo między obrazem źródłowym a obrazem docelowym) [23] [24] [25].

DEEPFAKE GŁOSOWY

Istnieje również wiele narzędzi do generowania realistycznej mowy syntetycznej, opartej na klonowaniu głosu. Ich celem jest zmiana tekstu na mowę z wykorzystaniem cech głosu ze zdefiniowanej próbki. Krok ten polega na konwersji odpowiedzi uzyskanej z zapytania do bazy danych i przetworzenia jej na mowę. Wyznacznikiem użyteczności tego typu narzędzi jest uzyskanie maksymalnej neutralności, zrozumiałości i podobieństwa do klonowanej próbki głosu.

Rozwiązania te najczęściej oparte są na oprogramowaniu SV2TTS, o otwartym kodzie źródłowym [26]. Mogą funkcjonować samodzielnie lub być zaimplementowane do innego narzędzia (najczęściej modyfikacji wideo). Jego ideą jest klonowanie głosu z kilku sekund oryginalnej mowy, zapisanie do bazy i analiza, bez konieczności ponownego trenowania modelu. Obsługa danych jest wydajniejsza, szybsza i mniej kosztowna obliczeniowo niż trenowanie oddzielnego modelu dla każdej próbki głosu [27]. Kompletna struktura SV2TTS to trzystopniowa warstwa, która składa się z kodera mówcy, syntezatora i vokodera (tj. dekodera głosu). Pierwszy element zasilany jest referencyjnym źródłem dźwięku mówcy do sklonowania i generuje osadzenie (niskowymiarową i średnią reprezentację głosu mówcy). Drugi pobiera tekst jako dane wejściowe

i wyprowadza spektrogram log-mel[7]. Z kolei vokoder generuje kształt fali mowy. Jakość materiału wyjściowego może być tylko tak dobra, jak referencyjny dźwięk. Celem tego kroku jest odtworzenie odpowiedzi z poprzedniego kroku w głosie odpowiedniej próbki. Tak więc, zgodnie z ramami SV2TTS, wyjściem będzie dźwięk zawierający zdanie wejściowe odtworzone za pomocą sklonowanego głosu dźwięku referencyjnego [28].

USŁUGI W MODELU CRIME-AS-A-SERVICE

W zakresie oferowanych usług w modelu *Crime-as-a-service*[8] przeprowadzony został autorski przegląd forów i komunikatorów wykorzystywanych przez przestępców (analiza *Cyber Threat Intelligence*). Opisy i wnioski wyciągnięto na podstawie wyselekcjonowania przykładowych rozmów oraz wpisów przestępców, zainteresowanych technologią *deepfake*. W internecie dostępne są oferty stworzenia treści przy wykorzystaniu omawianych technologii, a osoba decydująca się skorzystać z usługi wskazuje, czyj wizerunek lub głos zostaną wykorzystane do przygotowania zmanipulowanego materiału oraz jaka treść zostanie przedstawiona jako wynik. Korzystać można również z botów do tworzenia nagrań *deepfake* i samodzielnie konfigurować nagranie, mając jednocześnie wsparcie „przestępczego supportu". Przedstawiane i oferowane są również narzędzia z możliwością wykupienia dostępu offline, zainstalowania oprogramowania w sieci lokalnej i wdrożenia oraz przystosowania narzędzia do swoich potrzeb. Całość pracy, przygotowanie środowiska oraz dalsze zabezpieczenie użytego rozwiązania, oferowane jest przez usługodawcę (czyli przestępcę). Można również pozyskać dostęp do tzw. wersji testowej, czyli materiału demonstracyjnego, w którym można wykonać kilka podstawowych funkcji i sprawdzić jakość oraz łatwość korzystania z narzędzia.

7 Spektrogram log-mel – spektrogram, w którym częstotliwości są przeliczane na skalę mel, gdzie mel to skala wysokości dźwięku mierzona metodą akustyki określającej subiektywny odbiór poziomu dźwięku przez ucho ludzkie, względem obiektywnej skali pomiaru częstotliwości dźwięku w hercach [28].

8 *Crime-as-a-service* (CaaS) – model, w którym doświadczony cyberprzestępca opracowuje zaawansowane narzędzia lub usługi, które są wystawiane na sprzedaż lub do wynajęcia innym, często mniej doświadczonym cyberprzestępcom. W rezultacie nawet osoby o ograniczonej wiedzy i doświadczeniu są w stanie przeprowadzać ataki ze względną łatwością [29].

Ryzyko związane z wykorzystaniem sztucznej inteligencji, w tym przede wszystkim technologii *deepfake*, do ataków na rynku finansowym

Nowym trendem w działalności cyberprzestępców jest wykorzystywanie algorytmów sztucznej inteligencji do optymalizacji i usprawnienia działań przestępczych, głównie poprzez automatyzację oraz generowanie przekonujących treści tekstowych, głosowych oraz wizualnych, które zostały opisane powyżej. Działania te stanowią potencjalnie duże zagrożenie dla bezpieczeństwa środków finansowych klientów. Wprawdzie aktualne badania [30] pokazują, że jako społeczeństwo mamy świadomość zagrożeń związanych z korzystaniem z internetu, ale realne straty polskich obywateli wynikające z działań cyberprzestępczych liczone są w setkach milionów złotych [31].

Wykorzystanie *deepfake* do kradzieży środków finansowych jako stosunkowo nowy trend, stanowi w ocenie autorów istotne ryzyko, wynikające ze stałego rozwoju i powszechnej dostępności narzędzi umożliwiających generowanie treści oszukańczych, które coraz trudniej odróżnić od oryginału. Zarówno klienci, jak i instytucje rynku finansowego, zmuszeni będą do kolejnych zmian zachowania oraz poznania sposobów przeciwdziałania tym zagrożeniom.

Analiza zagrożeń dla klientów polskiego rynku finansowego

Głównymi ryzykami materializującymi się w przypadku działań cyberprzestępców na rynku finansowym (zarówno w odniesieniu do klientów indywidualnych, jak i instytucjonalnych) są działania prowadzące ostatecznie do straty finansowej, a w tym oszustwa internetowe i nieautoryzowane transakcje, naruszenie prywatności oraz ryzyko reputacyjne (głównie w odniesieniu do instytucji rynku finansowego).

OSZUSTWA INTERNETOWE I NIEAUTORYZOWANE TRANSAKCJE

Obserwując rozwój scenariuszy przestępczych realizowanych z wykorzystaniem sztucznej inteligencji poza granicami Polski można domniemywać,

iż w niedługim czasie grupy cyberprzestępcze, atakujące polskich obywateli i polskie organizacje finansowe, zaczną inkorporować AI do swoich scenariuszy. Przewidywane obszary tych działań to przede wszystkim:

- wzmocnienie przekazu phishingowego poprzez wykorzystanie modeli językowych AI do generowania wiadomości oszukańczych pisanych poprawną polszczyzną. Modele sztucznej inteligencji, mimo wbudowanych mechanizmów bezpieczeństwa, są proste do zmanipulowania. Zabezpieczenia wbudowane w ChatGPT nie pozwalają na bezpośrednie wygenerowanie maila phishingowego, ale można je ominąć zmieniając narrację, informując np. że treść tego maila jest potrzebna do badań bezpieczeństwa lub też wprowadzając AI w rolę, np. analityka bezpieczeństwa, który czyta treść maila phishingowego od cyberprzestępców;

- wykorzystanie AI jako motywu przewodniego tzw. leadu do skłonienia ofiar do działań na ich niekorzyść. Tematyka sztucznej inteligencji pojawia się w przekazach medialnych i przestrzeni publicznej jako nowatorski obszar technologii, co jest okazją przestępców do serwowania oszukańczych informacji;

- generowanie zmanipulowanych treści wideo i treści głosowych, pozwalających na autoryzację w systemach weryfikacji głosowej oraz wideoweryfikacji [32];

- tworzenie fałszywych tożsamości, wykorzystywanych do zakładania kont bankowych i używanych w różnego rodzaju działalności przestępczej;

- wykorzystanie AI do skuteczniejszego prowadzenia ataków spearphishingowych[9], poprzez tworzenie treści głosowych i wideo z wykorzystaniem wizerunku osób z rodziny ofiary [34];

9 „Spearphishing to konkretny i ukierunkowany atak na jedną lub wybraną liczbę ofiar, w odróżnieniu od typowego phishingu, który ma na celu oszukanie mas ludzi" [33].

- zwiększenie skuteczności ataków spearphishingowych, tzw. BEC (*bussines email compromise*)[10], poprzez np. wykorzystanie próbki głosu lub nagrania wideo osoby zlecającej wykonanie transakcji finansowych na wysokie kwoty;

- potencjalna możliwość zmanipulowania treści stanowiących odpowiedzi sztucznej inteligencji na zadane przez użytkowników zapytania i doprowadzenie do sytuacji, w której z wyliczeń najbardziej prawdopodobnych odpowiedzi AI przekaże odbiorcy zmanipulowaną informację[11].

NARUSZENIE PRYWATNOŚCI

Wykorzystanie wizerunku nieświadomej osoby w procesie autoryzacji tam, gdzie wykorzystywana jest biometria (próba głosu, wideo), jest przykładowym ryzykiem notowanym w innych krajach. Wiele rozwiązań wykorzystujących biometrię czy to wideo, czy głosową powstawało w czasie, kiedy nie były znane metody *deepfake*. To oznacza, że mogą (ale nie muszą) być one nieodporne na zmanipulowane treści serwowane przez cyberprzestępców. Ze względu na powszechność rozwiązań umożliwiających generowanie zmanipulowanych treści, a także dostępność wystarczającej ilości danych związanych z ofiarą (zdjęcia, nagrania wideo, nagrania głosu), potrzebnych do stworzenia zmanipulowanego materiału, istnieje możliwość wykorzystania wizerunku dowolnej osoby do ataku na jej środki finansowe. Wykrycie tego typu oszustwa na etapie jego realizacji stanowi prawdziwe wyzwanie dla systemów i zespołów bezpieczeństwa instytucji finansowych.

10 *Bussines email compromise* (BEC) to wyrafinowane oszustwo, którego celem są zarówno firmy, jak i osoby fizyczne, które wykonują uzasadnione żądania transferu środków, najczęściej poprzez podszycie pod konkretną osobę [35].

11 Autorzy nie będą również podawać przykładów możliwego użycia tej formy zagrożenia, aby nie stanowić wzoru dla działań przestępczych.

RYZYKO REPUTACYJNE

Ostatnim omawianym ryzykiem, mogącym zmaterializować się poprzez przestępcze wykorzystanie AI, jest negatywny wpływ działań przestępczych na reputację podmiotów rynku finansowego. W przypadku skutecznego wykorzystania technologii *deepfake* do przeprowadzenia oszustw, np. autoryzacji wideo lub głosowej na szkodę znacznej liczby klientów konkretnej instytucji finansowej, może to wpłynąć na zaufanie klientów do tej instytucji (w konsekwencji ograniczając wolumen klientów), a w szerszej perspektywie – również na zaufanie społeczne do całego rynku finansowego. Materializacja tego ryzyka może mieć również miejsce w sytuacji, kiedy przestępcy przygotują zaawansowane, zmanipulowane materiały, wykorzystujące wizerunek organizacji lub jej pracowników, w ramach których przedstawiane zostaną treści nieakceptowalne przez klientów bądź wprowadzające ich w błąd i prowadzące np. do strat finansowych.

Przykłady zastosowania sztucznej inteligencji przez cyberprzestępców

Znane są już na świecie przykłady wykorzystania technologii deepfake do oszustw na tle finansowym. Wynika to przede wszystkim z łatwości pozyskania materiału wejściowego do przygotowania produktu, jakim jest zmodyfikowana treść (wideo, głosowa lub statyczna). Próbkę głosu czy wizerunku danej osoby można znaleźć w internecie (media społecznościowe, platformy streamingowe itd.) lub pozyskać z konferencji (naukowych, prasowych etc.). Kolejnym sposobem zebrania potrzebnego materiału jest wykonanie połączenia głosowego (jednego lub więcej) pod wymyślonym pretekstem (np. podszycie pod telemarketera) i zebranie próbki głosu z rozmowy telefonicznej. Przestępcy udowodnili już, że jest to realne zagrożenie, a poniższe przykłady stanowią obraz możliwości zdarzeń w sektorze finansowym. Należą do nich:

- oszustwa z serii BEC, a w nim zdarzenie oszukańcze na kwotę ok. 243 000 USD [36] [37], złożenie dyspozycji przelewu, wraz z poprawną autoryzacją na kwotę 35 000 000 USD [38], a także transferów pieniędzy na łączną sumę ok. 400 000 USD [39];

- wykorzystanie wizerunku Elona Muska, który rzekomo zachęca do inwestowania pieniędzy i wyłudzenie ok. 2 000 000 USD, w ciągu 6 miesięcy, od kilkudziesięciu osób [40];
- scenariusz znany pod nazwą „metoda na wnuczka" i próba oszustwa pod pretekstem potrzeby zapłaty kaucji [34];
- podszycie wideo pod przyjaciela ofiary i wyłudzenie 662 000 USD [41].

Nowatorskim przykładem wykorzystania *deepfake* jest eksperyment przeprowadzony przez badacza bezpieczeństwa w Stanach Zjednoczonych, który przy wykorzystaniu technologii *deepfake* podszył się pod samego siebie i przeprowadził operacje bankowe [42].

Podsumowanie

Problematyka wykorzystania sztucznej inteligencji do działalności cyberprzestępczej, opisana w przedmiotowym opracowaniu, będzie – w ocenie autorów – jedną z najistotniejszych prób dla bezpiecznego funkcjonowania rynku finansowego, obok ataków na łańcuchy dostaw, wyzwań związanych z zarządzaniem podatnościami i działalności grup wykorzystujących *ransomware*.

Przewidywany wzrost skali wykorzystania sztucznej inteligencji do działań przestępczych determinuje nie tyle potrzebę, co wręcz konieczność użycia rozwiązań AI do ochrony danych oraz przeciwdziałania nadużyciom i przestępstwom. Ludzka percepcja już nie wystarcza i musi być wspomagana przez technologię, żeby odróżnić treści zmanipulowane, generowane przez AI, od prawdziwych.

Powyższa analiza implikuje potrzebę rozwijania odpowiednich regulacji i mechanizmów nadzoru, które mogą efektywnie wspierać zarządzanie ryzykiem związanym ze sztuczną inteligencją, nie hamując jednocześnie innowacji.

Dostrzegając zagrożenia opisane w przedmiotowym opracowaniu, Urząd Komisji Nadzoru Finansowego [43] powołał stałą, dedykowaną grupę

roboczą. W ramach jej prac, w których uczestniczyć będą przedstawiciele sektora bankowego, adresowane będą nadchodzące wyzwania technologiczne i legislacyjne [2], związane z wykorzystaniem sztucznej inteligencji.

Podsumowując, sztuczna inteligencja ma potencjał do znacznej transformacji wszystkich gałęzi gospodarki, w tym sektora finansowego, ale musi towarzyszyć jej rozsądek i ostrożność. Tylko poprzez zrównoważone podejście, które łączy innowacje technologiczne z odpowiednim zarządzaniem ryzykiem i regulacjami formalnymi, możemy skutecznie wykorzystać korzyści płynące z innowacyjnych narzędzi i modeli opartych na uczeniu maszynowym, jednocześnie minimalizując potencjalne szkody dla klientów rynku finansowego.

Bibliografia

[1] W. J. Murdoch, Ch. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu (14 czerwca 2019). „Interpretable machine learning: definitions, methods, and applications". *The Proceedings of the National Academy of Sciences* [Online]. Dostęp: https://arxiv.org/abs/1901.04592.

[2] Wniosek – Rozporządzenie Parlamentu Europejskiego i Rady ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji (akt w sprawie sztucznej inteligencji) i zmieniające niektóre akty ustawodawcze Unii.

[3] GitHub. *Copilot* [Online]. Dostęp: https://github.com/features/copilot

[4] Midjourney. *About* [Online]. Dostęp: https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F

[5] OpenAI (30 listopada 2022). *Introducing ChatGPT* [Online]. Dostęp: https://openai.com/blog/chatgpt

[6] O. Darcy (31 maja 2023). Experts are warning AI could lead to human extinction. Are we taking it seriously enough? [Online]. Dostęp: https://edition.cnn.com/2023/05/30/media/artificial-intelligence-warning-reliable-sources/index.html

[7] Komisja Europejska (22 maja 2007). „Komunikat do Parlamentu Europejskiego, Rady oraz Komitetu Regionów, 'W kierunku ogólnej strategii zwalczania cyberprzestępczości' [Online]. Dostęp: https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX%3A52007DC0267

[8] CERT Polska (2022). „Raport roczny z działalności CERT Polska" [Online]. Dostęp: https://cert.pl/uploads/docs/Raport_CP_2022.pdf

[9] Centrum Edukacji dla Bezpieczeństwa Rynku Finansowego. *Encyklopedia Cyberbezpieczeństwa* [Online]. Dostęp: https://cebrf.knf.gov.pl/encyklopedia/hasla

[10] J. Norem (20 stycznia 2023). Hackers Buy Google Ads to Push Malware Through Searches for Popular Apps [Online]. Dostęp: https://www.extremetech.com/computing/342464-hackers-buy-google-ads-to-push-malware-through-searches-for-popular-apps

[11] „Deepfake". Merriam webster [Online]. Dostęp: https://www.merriam-webster.com/dictionary/deepfake

[12] M. Maras M., A. Alexandrou, „Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos", *International Journal of Evidence & Proof*, vol. 23(3), ss. 255–262, 2019.

[13] A. Chadha, V. Kumar, S. Kashyap, M. Gupta, „Deepfake: An Overview" w: *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, P. K. Singh, S. T. Wierzchoń, S. Tanwar, M. Ganzha, J. J. P. C. Rodrigues (eds). Singapore: Springer 2021.

[14] Y. Lin, K. Parvataneni, „Deepfake Generation, Detection, and Use Cases: A Review Paper", *International Journal of Computational and Biological Intelligent Systems*, vol. 3(2), 2021.

[15] H. Hasan, K. Salah, „Combating Deepfake Videos Using Blockchain and Smart Contracts", *IEEE Access*, vol. 7, ss. 41596–41606, 2019.

[16] J. Cox (7 października 2019). *Most Deepfakes Are Used for Creating Non-Consensual Porn, Not Fake News* [Online]. Dostęp: https://www.vice.com/en/article/7x57v9/most-deepfakes-are-porn-harassment-not-fake-news

[17] GitHub. *Let's build from here* [Online]. Dostęp: https://github.com

[18] Open Source Initiative. *Defining Open Source AI* [Online]. Dostęp: https://opensource.org

[19] M. Ołdakowski (2022), „Tworzenie obrazów abstrakcyjnych z użyciem Generatywnej Sieci Przeciwstawnej", praca inżynierska obroniona na PJATK, 2022.

[20] Github. *Deepfakes/faceswap* [Online]. Dostęp: https://github.com/deepfakes/faceswap

[21] Github. *Ipazc/mtcnn* [Online]. Dostęp: https://github.com/ipazc/mtcnn

[22] Github. *iperov/DeepFaceLab* [Online]. Dostęp: https://github.com/iperov/DeepFaceLab

[23] Y. Guo, L. Jiao, S. Wang, F. Liu, „Fuzzy Sparse Autoencoder Framework for Single Image Per Person Face Recognition", w: *IEEE Transactions on Cybernetics*, 2018.

[24] S. Lyu (29 sierpnia 2018). *Detecting 'Deepfake' Videos In The Blink Of An Eye* [Online]. Dostęp: https://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072

[25] J. Cao, Y. Hu, B. Yu, R. He, Z. Sun, „3D Aided Duet GANs for Multi-View Face Image Synthesis", w: *IEEE Transactions on Information Forensics and Security*, 2019.

[26] Github. *CorentinJ/Real-Time-Voice-Cloning* [Online]. Dostęp: https://github.com/CorentinJ/Real-Time-Voice-Cloning

[27] C. Jemine (2019), „Real-Time Voice Cloning", praca magisterska [Online]. Dostęp: https://matheo.uliege.be/handle/2268.2/6801

[28] L. Roberts (6 marca 2020). *Understanding the Mel Spectrogram* [Online]. Dostęp: https://medium.com/analytics-vidhyaunderstanding-the-mel-spectrogram-fca2afa-2ce53

[29] D. Manky, „Cybercrime as a service: a very modern business", *Computer Fraud & Security* no. 6, ss. 9-13, 2013.

[30] Google Polska (25 listopada 2022). *Jesteśmy coraz bardziej świadomi zagrożeń w sieci* [Online]. Dostęp: https://polska.googleblog.com/2022/11/jestesmy-coraz--bardziej-swiadomi.html

[31] KNF (wrzesień 2022). „Informacja na temat sytuacji sektora bankowego w 2021 roku" [Online]. Dostęp: https://www.knf.gov.pl/knf/pl/komponenty/img/Raport_roczny_2021.pdf

[32] KNF (3 marca 2022). "Stanowisko UKNF dotyczące identyfikacji klienta instytucjonalnego i weryfikacji jego tożsamości w sektorze finansowym podlegającym nadzorowi KNF w oparciu o metodę wideoweryfikacji" [Online]. Dostęp: https://www.knf.gov.pl/knf/pl/komponenty/img/Stanowisko_UKNF_dot_wideoweryfikacji_klientow_instytucjonalnych.pdf

[33] C. Halevi, N. Memon, O. Nov (4 stycznia 2015). „Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks" [Online]. Dostęp: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2544742

[34] KPP Puck. *Oszustwa metodą na wnuczka* [Online]. Dostęp: https://puck.policja.gov.pl/m11/profilaktyka-4/59382,OSZUSTWA-METODA-NA-WNUCZKA.html

[35] D. Bakari, Ch, Shukla, „Business E-mail Compromise — Techniques and Countermeasures", w: International Conference on Advance Computing and Innovative Technologies in Engineering 2021 [Online]. Dostęp: https://www.researchgate.net/publication/352398449_Business_E-mail_Compromise_-_Techniques_and_Countermeasures

[36] A. Bednarek (18 października 2021). *Oszustwo na kolegę zaraz wejdzie na wyższy poziom. Podrobili głos prezesa i obrabowali duży bank* [Online]. Dostęp: https://spidersweb.pl/2021/10/deepfake-glos-oszustwo-bank.html

[37] T. Brewster (14 października 2021). *Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find* [Online]. Dostęp: https://www.forbes.com/sites/thomasbrewster/2021/10/14huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=4558777f7559

[38] P. Verma, *They thought loved ones were calling for help. It was an AI scam* [Online]. Dostęp: https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/

[39] J. Żabnicka (23 maja 2023). *Oszustwo angażujące tzw. „deepfake" w Chinach wywołało obawy przed oszustwami opartymi na sztucznej inteligencji* [Online]. Dostęp: https://itreseller.com.pl/oszustwo-angazujace-tzw-deepfake-w-chinach--wywolalo-obawy-przed-oszustwami-opartymi-na-sztucznej-inteligencji/

[40] M. Di Salvo (25 maja 2022). *Deepfake Video of Elon Musk Promoting Crypto Scam Goes Viral* [Online]. Dostęp: https://decrypt.co/101365/deepfake-video-elon-musk-crypto-scam-goes-viral

[41] A. Petynia-Kawa (26 maja 2023). *Nowe oszustwo z wykorzystaniem nagrania video* [Online]. Dostęp: https://www.politykabezpieczenstwa.pl/pl/a/nowe-oszustwo-z-wykorzystaniem-nagrania-video

[42] J. Cox (23 lutego 2023). *How I Broke Into a Bank Account With an AI-Generated Voice* [Online]. Dostęp: https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice.

[43] KNF. *Misja, wizja i wartości UKNF* [Online]. Dostęp: https://www.knf.gov.pl/o_nas/urzad_komisji/misja_wizja

Strategia Sztucznej Inteligencji dla NATO

Piotr Słowiński

NASK-PIB | Centrum Cyberbezpieczeństwa i Infrastruktury |
Dział Strategii i Rozwoju Bezpieczeństwa Cyberprzestrzeni
Uniwersytet Warszawski | Wydział Prawa i Administracji

Wprowadzenie

W październiku 2021 roku ministrowie obrony krajów Sojuszu Północnoatlantyckiego przyjęli strategię sztucznej inteligencji dla NATO. Stanowi ona pierwszy dokument strategiczny, dotyczący wykorzystania AI w zakresie obrony i bezpieczeństwa w sposób bezpieczny i etyczny oraz zgodny zarówno z wartościami Sojuszu, jak i prawem międzynarodowym. Opracowanie wskazuje cele i zasady odpowiedzialnego korzystania z AI w NATO, a także wybrane zidentyfikowane zagrożenia dla bezpieczeństwa sojuszniczego wykorzystania sztucznej inteligencji.

Na kanwie rozważań o strategicznym podejściu do AI w Sojuszu należy także wspomnieć o Radzie Eksperckiej NATO ds. Danych i Sztucznej Inteligencji. Obok działań w wymiarze strategicznym, aktualnie NATO jest także w trakcie pilotażowego programu wykorzystania rozwiązań AI w zakresie cyberobrony, przeciwdziałania zmianom klimatycznym oraz analizy obrazowej.

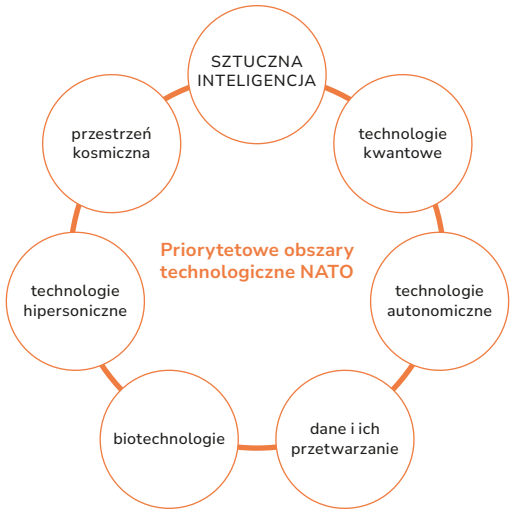
Biorąc pod uwagę fakt, że rozwój AI jest silnie związany z koncepcjami takimi jak przetwarzanie danych, w tym *big data*, technologie autonomiczne oraz biotechnologie, to jednocześnie przyjęto pierwszą politykę ramową NATO dotyczącą wykorzystania danych. Stanowi ona ramy dla wykorzystywania ich jako strategicznego zasobu w sposób

odpowiedzialny i zgodny z wartościami Sojuszu w każdej ze sfer jego działalności – wojskowej, cywilnej oraz politycznej [1].

Dla NATO rozwój rozwiązań wykorzystujących AI, które mogłyby zostać wykorzystane w różnych domenach operacyjnych, stanowi jeden z priorytetów w obszarze *emerging and disruptive technologies* (EDT) i będzie kształtował przyszłe innowacyjne działania Sojuszu na tym polu. Niezbędne wobec tego staje się stworzenie ujednoliconych oraz ustandaryzowanych założeń i ram działania oraz postępowania ze sztuczną inteligencją w ramach tak szerokiego i zróżnicowanego poziomu rozwoju technologicznego poszczególnych jego członków.

Główne założenia i cele strategii

Sztuczna inteligencja stanowi jeden z siedmiu priorytetowych dla państw NATO obszarów technologicznych, niezbędnych do zwiększenia zdolności w zakresie obrony i bezpieczeństwa.



RYS. 1. Priorytetowe obszary technologiczne NATO *(opracowanie własne)*

Zgodnie z obserwacjami poczynionymi przez autorów strategii, AI będzie stanowiła jednocześnie bezprecedensową szansę, wyzwanie, ale także zagrożenie w zakresie globalnego bezpieczeństwa. W szczególności chodzi o dynamikę jej rozwoju oraz powiązanie z innymi obszarami,

tak charakterystyczne dla współczesnych trendów w EDT. Zgodnie z przewidywaniami na poziomie NATO, AI wpłynie na pełne spektrum działań podejmowanych w trzech głównych obszarach aktywności Sojuszu: kolektywnej obronie, zarządzaniu kryzysowym oraz bezpieczeństwie kooperatywnym.

Obecne warunki rozwoju sztucznej inteligencji wymagają podjęcia określonych działań na poziomie całego Sojuszu Północnoatlantyckiego, w celu utrzymania przewagi technologicznej NATO. Państwa członkowskie zobowiązały się w strategii do pogłębienia współpracy w celu zapewnienia globalnego bezpieczeństwa wobec potencjalnych zagrożeń wynikających z rozwoju AI. Przede wszystkim wykorzystywane będą dotychczas istniejące zasoby oraz inicjatywy wdrażające określone rozwiązania – zarówno na poziomie krajowym, jak i Sojuszu.

Strategia AI dla NATO wskazuje cztery cele, dla których została ona sformułowana oraz przyjęta:

1. stworzenie podstaw dla całego Sojuszu oraz stanowienie przykładu i zachęty dla świata do rozwijania i wykorzystania sztucznej inteligencji do realizacji celów z zakresu obrony i bezpieczeństwa w sposób odpowiedzialny;

2. przyspieszenie oraz rozpowszechnienie wdrażania sztucznej inteligencji w obszarze rozwoju i kształtowania zdolności, wzmacniając tym samym interoperacyjność w NATO (m.in. poprzez propozycje scenariuszy wykorzystania AI, tworzenie nowych struktur oraz programów);

3. ochrona i monitorowanie sojuszniczych technologii AI oraz związanej z tym zdolności do wdrażania innowacyjnych rozwiązań i uwzględniania ich w tworzonych politykach bezpieczeństwa, takich jak np. zasady odpowiedzialnego użytkowania;

4. identyfikacja i ochrona przed zagrożeniami oraz szkodliwymi działaniami wykorzystującymi sztuczną inteligencję zarówno ze strony podmiotów państwowych, jak i niepaństwowych [2].

Docelowo integracja technologii wykorzystujących sztuczną inteligencję w celu wsparcia działań w trzech głównych obszarach zainteresowania

NATO ma być prowadzona w świadomy, odpowiedzialny oraz zgodny z prawem międzynarodowym sposób w całym Sojuszu, z uwzględnieniem różnych zadań oraz poziomów operacyjnych.

Niezależnie od aktywności wewnętrznych, NATO ma pozostać globalnym forum zajmującym się zastosowaniem sztucznej inteligencji w zadaniach z zakresu obrony i bezpieczeństwa, szczególnie w zakresie wykorzystania jej do szkodliwych działań ze strony podmiotów państwowych i niepaństwowych.

Zasady odpowiedzialnego wykorzystania AI według NATO

Zasadniczą częścią strategii AI dla NATO są sformułowane zasady odpowiedzialnego wykorzystania tej technologii według Sojuszu. Są one w rzeczywistości oparte na już istniejących i powszechnie znanych oraz akceptowanych standardach etycznych i prawnych. Mają jednak stanowić kolejne potwierdzenie podejścia NATO i krajów członkowskich do kwestii wykorzystania AI w określony sposób – nie tylko na podstawie wypracowanych standardów i praktyk, ale także z uwzględnieniem wartości Paktu Północnoatlantyckiego oraz przepisów prawa międzynarodowego [2]. Jednocześnie na każdym etapie podkreśla się to, że mogą istnieć już wdrożone krajowe lub międzynarodowe regulacje i procedury dotyczące sztucznej inteligencji i jej wykorzystania – założenia lub zasady formułowane przez strategię NATO mają nie wpływać ani nie zastępować już powstałych w ich wyniku obowiązków prawnych.

Kształt strategii wynika przede wszystkim z zadań NATO jako sojuszu wojskowego i zobowiązań z tego wynikających, wymagających jak największej interoperacyjności pomiędzy nie tylko krajami, ale również różnymi rodzajami wojsk. Oznacza to konieczność ewentualnego dostosowania rozwiązań do wymogów różniących się od siebie typów jednostek lub instytucji wojskowych i to już zarówno na etapie projektowania, jak i wdrażania.

Sformułowane ogólne zasady odpowiedzialnego wykorzystania (ZOW) AI według NATO mają stanowić punkt wyjścia do stworzenia bardziej szczegółowych dobrych praktyk. Mają być one również punktem odniesienia

dla krajów członkowskich w zakresie wykorzystywania sztucznej inteligencji w obszarach obrony i bezpieczeństwa. Państwa Sojuszu oraz sama organizacja na kanwie strategii sztucznej inteligencji dla NATO zobowiązują się do zapewnienia zgodności z sześcioma zasadami na każdym etapie cyklu życia aplikacji AI[1]:

1. LEGALNOŚCI – tworzenie lub wykorzystanie aplikacji AI ma być zgodne z krajowymi i międzynarodowymi regulacjami, w szczególności – z międzynarodowym prawem humanitarnym oraz zapewniać uwzględnianie praw człowieka;

2. ODPOWIEDZIALNOŚCI I ROZLICZALNOŚCI – tworzenie lub wykorzystanie aplikacji AI musi wiązać się z wdrożeniem odpowiednich poziomów ostrożności i rozwagi działania oraz określeniem odpowiedzialności człowieka w celu zagwarantowania rozliczalności;

3. WYJAŚNIALNOŚCI I IDENTYFIKOWALNOŚCI – aplikacje sztucznej inteligencji mają być przejrzyste, a ich działanie zrozumiałe, w szczególności w zakresie wykorzystywanej metodologii, źródeł oraz procedur; ma to obejmować weryfikację, ocenę oraz walidację mechanizmów na poziomie krajowym lub NATO;

4. SOLIDNOŚCI DZIAŁANIA – sytuacje, w których możliwe będzie wykorzystanie aplikacji AI, zostaną w jasny i wyraźny sposób określone, a bezpieczeństwo i niezawodność ich zdolności będą zagwarantowane i testowane przez cały cykl ich życia, łącznie z zapewnieniem zgodności z procedurami certyfikacyjnymi ustanowionymi na poziomie krajowym lub NATO;

5. ZARZĄDZALNOŚCI – aplikacje AI będą tworzone, rozwijane i wykorzystywane zgodnie z ich przeznaczeniem, a także umożliwią odpowiednią interakcję pomiędzy człowiekiem a maszyną. Będą one również zdolne do wykrywania oraz unikania niezamierzonych lub nieprzewidzianych konsekwencji, a ponadto do podejmowania działań takich

1 W oryg. *AI applications* – należy to określenie rozumieć jako wszelkiego rodzaju oprogramowanie wykorzystujące technologie lub algorytmy sztucznej inteligencji i w tym konkretnym kontekście, które mogą mieć zastosowanie w systemach lub rozwiązaniach technicznych wykorzystywanych przez NATO.

jak zaprzestanie działania lub wyłączenie systemów, jeżeli zachowywałyby się one w sposób niepożądany;

6. OGRANICZANIA UPRZEDZEŃ I STRONNICZOŚCI – podejmowane będą proaktywne działania w celu minimalizacji niezamierzonych uprzedzeń, które mogą się pojawić w rozwijanych lub wykorzystywanych aplikacjach AI lub zbiorach danych [2].

Niezależnie od zaleceń lub zasad sformułowanych przez NATO w strategii dotyczącej AI, Sojusz zobowiązał się do współpracy z międzynarodowymi podmiotami, zajmującymi się standardami związanymi ze sztuczną inteligencją w celu ułatwiania wypracowywania zgodności pomiędzy sektorami wojskowym i cywilnym w tym zakresie. W związku z tym przedstawione powyżej zasady mogą ulec zmianom lub uszczegółowieniu, szczególnie na etapie wdrażania i realizacji ich przez poszczególne kraje członkowskie Sojuszu.

Bezpieczeństwo sojuszniczej AI

Strategia NATO nie wskazuje konkretnych narzędzi, które państwa mają wykorzystywać, aby przeciwdziałać zagrożeniom dla sojuszniczej AI. Wymienione zostały potencjalne typy szkodliwych działań – zakłócanie, manipulacja lub sabotaż sztucznej inteligencji. Rozumieć przez to można wpływanie zarówno na zasady działania aplikacji AI, ale także na te oprogramowania lub sprzęt, które będą działały z nimi pomocniczo, a niekoniecznie ściśle podpadają pod definicję AI.

Postulatem autorów strategii jest stosowanie narzędzi do cyberobrony wykorzystujących sztuczną inteligencję, a także rozwijanie odpowiednich procesów certyfikacji bezpieczeństwa, takich jak ramy analizy zagrożeń i audyty bezpieczeństwa. Wdrażane środki powinny uwzględniać przetestowanie działania aplikacji AI w sytuacjach skrajnych i awaryjnych [2].

Zidentyfikowano także możliwe ryzyka, wynikające z chęci wykorzystania AI jako narzędzia do działań szkodliwych. Przykładowo, grupy inspirowane lub sponsorowane przez państwa, ale również te od nich niezależne, mogą wykorzystać sztuczną inteligencję do ataków na infrastrukturę krytyczną lub systemy odpowiedzialne za odporność sektora

cywilnego. Dodatkowo mogą jednocześnie prowadzić operacje informacyjne i kampanie dezinformacyjne ukierunkowane na podważenie zaufania publicznego do wykorzystania AI do celów militarnych i przez instytucje lub jednostki wojskowe.

Wobec powyższych zagrożeń dla bezpieczeństwa Sojuszu, jak i poszczególnych jego członków, kraje NATO mają dążyć do zapobiegania oraz przeciwdziałania tego typu przedsięwzięciom zgodnie z ZOW oraz poprzez wykorzystanie komunikacji strategicznej w miarę potrzeb [2]. Mają być w tym także wspierane przez instytucje Paktu Północnoatlantyckiego w niezbędnym zakresie i zgodnie z zapotrzebowaniem.

Rada Ekspercka NATO ds. Danych i Sztucznej Inteligencji

7 lutego 2023 roku rozpoczęła pracę Rada Ekspercka NATO ds. Danych i Sztucznej Inteligencji. Jej podstawowym zadaniem jest opracowanie standardów certyfikacyjnych dla systemów AI, zapewniających że będzie ona działała w sposób odpowiedzialny oraz przyjazny użytkownikom. Celem jest zagwarantowanie, aby podmioty z różnych branż w krajach członkowskich Sojuszu miały pewność, że nowe projekty oparte o sztuczną inteligencję lub wykorzystujące ją są zgodne z przepisami prawa międzynarodowego, a także wartościami NATO [3]. Członkami Rady będą przedstawiciele każdego kraju członkowskiego Sojuszu oraz Szwecji [4], a także eksperci [3].

Sama Rada ma przede wszystkim być głównym forum dyskusji i podejmowania przedsięwzięć dla państw członkowskich oraz instytucji Sojuszu w zakresie zarządzania odpowiedzialnym rozwojem i wykorzystaniem AI poprzez operacjonalizację ZOW określonych w strategii sztucznej inteligencji dla NATO. Do wytycznych, którymi Rada ma się kierować, należą: budowanie zaufania, odpowiedzialne wdrażanie AI oraz wspólne działanie [4].

RYS. 2. Wytyczne dla Rady Eksperckiej NATO ds. Danych i Sztucznej Inteligencji *(opracowanie własne)*

Do podstawowych celów Rady będzie należało kierowanie odpowiedzialnym wdrażaniem AI, poprzez stworzenie i wykorzystanie do tego praktycznego zestawu narzędzi. Jego kluczowym elementem będzie standard certyfikacyjny oparty o praktyczne doświadczenia instytucji NATO oraz interesariuszy spoza NATO – z sektorów publicznego i prywatnego (w tym naukowego, akademickiego, społeczeństwa obywatelskiego i środowisk międzynarodowych). W stworzenie takiego zestawu narzędzi zaangażowane ma być szerokie grono interesariuszy, reprezentujących różnorodne środowiska, nie tylko sektor obronny [4]. Pierwotnie będzie on przeznaczony dla instytucji NATO, jednak dostęp do niego uzyskają także państwa członkowskie Sojuszu.

Rada służyć będzie także jako punkt gromadzący informacje i wymieniający się nimi, jeżeli któryś z krajów wspólnoty euroatlantyckiej zechciałby wdrożyć przygotowane przez Radę rozwiązania na poziomie krajowym [4]. Udział reprezentantów wszystkich sojuszników oraz różnych środowisk (prawników, inżynierów, wojskowych oraz ekspertów z zakresu etyki) w pracach organu [3] przyczynić się może do wypracowania rozwiązań, mogących mieć zastosowanie w różnych branżach i sektorach oraz do różnorodnych rodzajów zadań. Może to również zapewnić ich dostosowanie do zróżnicowanych warunków geopolitycznych, ustrojów politycznych, porządków prawnych oraz stanu rozwoju technologicznego.

Podsumowanie i wnioski

Strategia Sztucznej Inteligencji dla NATO stanowi pierwszy tego rodzaju dokument powstały w ramach Sojuszu, a dotyczący wykorzystania AI w zakresie obrony i bezpieczeństwa w sposób bezpieczny i etyczny oraz zgodny zarówno z jego wartościami, jak i prawem międzynarodowym. Najważniejszymi jej aspektami są zasady odpowiedzialnego wykorzystania AI według NATO oraz działania Rady Eksperckiej NATO ds. Danych i Sztucznej Inteligencji w zakresie wypracowania standardów certyfikacyjnych dla AI.

Te pierwsze nie odbiegają znacząco od już sformułowanych lub przyjmowanych standardów odnośnie do odpowiedzialnego wykorzystywania sztucznej inteligencji w innych międzynarodowych wspólnotach lub w poszczególnych krajach. Jednocześnie umieszczenie ich w dokumencie strategicznym dla całego NATO świadczy o chęci rzeczywistego wdrożenia ich jako uniwersalnego standardu wśród wszystkich członków Sojuszu. Z tego względu należy uznać to za krok w dobrą stronę, mający zapewnić minimalny wzorzec, gwarantujący poszanowanie wartości i prawa oraz umożliwiający kontrolę działania systemów AI.

Szczególną uwagę zwraca skoncentrowanie się autorów strategii na przejrzystości i możliwości zrozumienia działania aplikacji AI, a także na określeniu jasnych kryteriów ich wykorzystania w ramach NATO. Mimo że może to stanowić problem, szczególnie wobec wciąż zmieniających się możliwości AI oraz potencjału jej aplikowalności, należy uznać to za kluczowe aspekty, niezbędne do zagwarantowania nie tylko odpowiedzialnego, ale i bezpiecznego dla wszystkich wykorzystania rozwiązań AI.

Istotnym ryzykiem jest także zagadnienie realności postulatu przejrzystego i zrozumiałego opisania działania aplikacji AI, szczególnie wobec ich dynamicznego rozwoju, który obserwuje się aktualnie oraz którego należy się spodziewać w przyszłości. Dodatkowo wynikająca z tego zarządzalność AI może budzić uzasadnione wątpliwości. Można sobie zadać pytanie, czy realne jest zagwarantowanie w sposób adekwatny działania systemów sztucznej inteligencji zgodnie z ich przeznaczeniem, jeżeli nie będzie możliwe zrozumienie i przejrzyste wskazanie sposobu ich działania. Ryzyko faktycznej niemożliwości w dochowaniu standardów określonych w ZOW, mimo podejmowanych starań i wysiłków, prowadzić może bezpośrednio do niewystarczającego bezpieczeństwa rozwiązań z zakresu AI.

To z kolei prowadzić może do większych lub bardziej dotkliwych zagrożeń, aniżeli tylko takich o skali lokalnej lub krajowej.

Jeżeli chodzi o działania Rady Eksperckiej NATO ds. Danych i Sztucznej Inteligencji w zakresie wypracowania standardów certyfikacyjnych dla AI, to jest to kolejny filar dbałości o, z jednej strony, przejrzystość i jasność tak wykorzystania, jak i działania aplikacji sztucznej inteligencji, ale z drugiej – pożądane dążenie do ujednolicenia i ustandaryzowania wymagań względem ich funkcjonowania i działania na terenie krajów Sojuszu. Prace Rady jako ciała interdyscyplinarnego i złożonego z wielu ekspertów o zróżnicowanym wielobranżowym i wielodziedzinowym doświadczeniu przyczynią się do powstania w istocie praktycznego i obejmującego różnorodne aspekty zestawu narzędzi. Te z kolei wyewoluują w standardy certyfikacyjne, które będzie można zastosować tak na poziomie generalnym NATO, jak i na poziomie krajowym poszczególnych państw członkowskich. Godną podkreślenia zaletą jest zebranie i wypracowanie jak najbardziej interdyscyplinarnych wniosków, które mają szansę na zwiększoną ich aplikowalność, niezależnie od stopnia zaawansowania technologicznego państw, jak i dojrzałości politycznej, prawnej i społecznej.

Ryzyko i zagrożenie stanowi w tym zakresie możliwość jednolitego wdrożenia standardów oraz ewentualna oporność poszczególnych krajów wobec tego procesu, szczególnie jeśli będzie wymagał znaczących wydatków lub dużej ingerencji w aktualne uregulowania lub procedury. W takiej sytuacji faktycznie mogą wystąpić trudności na etapie wpierw formułowania konkretnych standardów, a następnie na etapie ich wprowadzania.

Postulowane przez autorów strategii oraz przyjęte przez przedstawicieli krajów członkowskich NATO zasady: budowania zaufania, odpowiedzialnego wdrażania sztucznej inteligencji oraz działania jako wspólnota, które należy uwzględniać w wykorzystaniu aplikacji AI w ramach Sojuszu Północnoatlantyckiego, są odzwierciedleniem nie tylko wartości stojących za NATO, ale także powinny przyświecać każdemu z krajów członkowskich. Tym istotniejsze będzie kierowanie się nimi i adekwatne wdrażanie na poziomie krajowym koncepcji wypracowanych na poziomie Sojuszu, a tym samym – także monitorowanie ich wprowadzania. Należy dopilnować, aby poszczególni członkowie NATO dostosowywali swoje procedury do wymogów ogólnych, aby zapewnić interoperacyjność pomiędzy sojuszniczymi wojskami na poziomie operacyjnym oraz strategicznym oraz zagwarantować przestrzeganie wspólnych wartości, spajających kraje NATO.

Bibliografia

[1] NATO (9 grudnia 2022). *Summary of NATO's Data Exploitation Framework Policy* [Online]. Dostęp: https://www.nato.int/cps/en/natohq/official_texts_210002.htm#:~:text=The%20Data%20Exploitation%20Framework%20Policy,members%20of%20the%20NATO%20Enterprise.

[2] NATO (22 grudnia 2021). *Summary of the NATO Artificial Intelligence* Strategy [Online]. Dostęp: https://www.nato.int/cps/en/natohq/official_texts_187617.htm.

[3] NATO (7 lutego 2023). *NATO starts work on Artificial Intelligence certification standard* [Online]. Dostęp: https://www.nato.int/cps/en/natohq/news_211498.htm.

[4] NATO (17 października 2022). *NATO's data and Artificial Intelligence Review Board* [Online]. Dostęp: https://www.nato.int/cps/en/natohq/official_texts_208374.htm.

Odporność AI dla odpornej wspólnoty

r. pr. **Robert Kroplewski**

Pełnomocnik Ministra Cyfryzacji ds. społeczeństwa informacyjnego

Definicja sztucznej inteligencji

Jak dotąd międzynarodowa społeczność ekspertów, przedstawicieli państw i środowisk społecznych [1] wypracowała, jako referencyjną dla tworzenia polityki i realizacji praktyki, następującą definicję sztucznej inteligencji:

> *System AI to system oparty na maszynie, który jest w stanie wpływać na środowisko poprzez generowanie danych wyjściowych (przewidywań, zaleceń lub decyzji) dla danego zestawu celów.*
>
> *Wykorzystuje on dane wejściowe pochodzące od maszyn lub ludzi, aby:*
>
> *i. postrzegać środowiska rzeczywiste lub wirtualne;*
>
> *ii. przełożyć te spostrzeżenia na modele poprzez analizę w sposób zautomatyzowany (np. za pomocą uczenia maszynowego) lub ręcznie; oraz*
>
> *iii. użyć wnioskowania z modelu do sformułowania opcji wyników.*
>
> *Systemy sztucznej inteligencji są zaprojektowane do działania z różnymi poziomami autonomii* [2].

Konwergencja cyberbezpieczeństwa i odporności w domenie AI

Rozumiejąc sztuczną inteligencję jako zestaw technik i metod modelowania wiedzy, ujętych w system i podlegający interakcjom z otoczeniem zewnętrznym (dalej: system AI), a nie tylko jako algorytm przetwarzający dane lub polecenia, należy stwierdzić, że reguły i procedury cyberbezpieczeństwa są już niewystarczające dla budowania i utrzymania odporności samego systemu AI, a także odporności środowiska, w którym jest osadzony. Należy przez to rozumieć odporność wspólnoty społecznej, zorganizowanej czy to w formę państwa, czy organizacji międzynarodowej, opartej o poszanowanie wspólnie zdefiniowanych zasad i reguł. Dotyczy to także dostępnych tej wspólnocie zasobów egzystencjalnych, w tym naturalnych, ale przede wszystkim odporności członka tej wspólnoty – człowieka. Jego godność ludzka, integralność psychofizyczna, oraz uczestnictwo w życiu społecznym podlegają napięciom w wyniku interakcji ze środowiskami cyfrowymi, wirtualnymi lub fizycznymi, zasilanymi systemami AI komunikującymi się ze sobą w czasie rzeczywistym.

Skala wyzwania zwiększa się, w sytuacji gdy system AI jest projektowany z założenia jako dopuszczający operacje z różnym stopniem ich autonomiczności w relacji do zamierzeń twórcy, operatora, analityka, czy zarządzającego środowiskiem, w którym ten system osadzono. Nierzadko systemowi AI stawiany jest cel wyraźnego lub domniemanego zastąpienia dotychczasowej działalności człowieka. Dzisiejsze społeczeństwo, rozumiane jako wspólnota ziemi i czasu, funkcjonuje jako rozproszone jednostki, niewspółpracujące i nieczujące odpowiedzialności za jego rozwój. Proces zastąpienia działalności człowieka, przeprowadzony w nieodpowiedzialny lub bezrozumny – wyłącznie utylitarny – sposób może prowadzić do negatywnego wpływu na otoczenie, z narażeniem odbudowy sił żywotnych i zasobów wspólnoty społecznej i planety.

Aktualny stan badań i rozwoju systemów sztucznej inteligencji eksponuje problem tzw. czarnej skrzynki – niewyjaśnialności rekomendacji, jakie płyną z systemu, w regułach tradycyjnej logiki liniowej przyczyny i skutku. Naturalna jest bowiem dla systemu AI konieczność skorzystania z odmiennego podejścia, np. opartego o analizę korelacji rezultatów lub metodyk logiki rozmytej. Niezależnie od tego, stale prowadzone są badania w zakresie wyjaśnialności systemów sztucznej inteligencji. Ryzyko błędu wprowadzania danych wejściowych, błędu modelowania

reguł systemu AI, zakłóceń percepcyjnych sensorów, braku kontekstowości operacji, zakłóceń aktywatorów albo negatywnego wpływu na otoczenie są wysokie i stwarzają trudności jakościowego nadzoru człowieka nad systemem AI.

Piecza nad godną zaufania AI i cyberbezpieczeństwo

Społeczność międzynarodowa dostrzegła potrzebę wypracowania wspólnego podejścia – rekomendacji – dla systemów AI i ich wpływu na systemy ekonomiczne i polityczne. Ujęto je pierwszy raz w 2018 r. (przez ekspertów grupy AIGO[1] w OECD) w koncepcję sprawowania pieczy nad rozwojem godnej zaufania sztucznej inteligencji – *stewardship of trustworthy AI* [1], co jest nieco innym podejściem niż zarządzanie (*governance*) w tradycyjnym rozumieniu.

Analogiczne prace toczyły się w 2018 r. równolegle w grupie ekspertów wysokiego szczebla HLEG w KE [3], której skład nielicznie pokrywał się ze składem osobowym grupy AIGO. Pozwoliły one wprowadzić koncepcję *trustworthy AI* do społeczności międzynarodowej poza Unią Europejską i zuniwersalizować ją, nie tylko w wymiarze etycznym, ale także produktowym (ekonomicznym).

Historycznym momentem uniwersalizacji koncepcji *trustworthy AI* było przyjęcie przez 192 państwa członkowskie UNESCO w 2021 r. *Rekomendacji dla etyki sztucznej inteligencji* [4], na podstawie wytycznych grupy ekspertów AHEG[2] z 2019 r. W rekomendacjach tych dostrzeżono asymetrię w dostępie do infrastruktury, bibliotek algorytmów i kodów oraz danych nie tylko w relacjach centrów systemowych[3] Wschodu albo

1 OECD Working Party on Artificial Intelligence

2 Ad Hoc Expert Group (AHEG) for the Recommendation on the Ethics of Artificial Intelligence.

3 Sformułowanie „centrum systemowe" rozumiane jest w kontekście teorii systemów-światów Immanuala Wallersteina – czyli centrum generującego system-świat wewnętrzny, zależności elementów systemowych, łańcuchy wartości, warunki ekonomii i egzystencji oraz warunki konkurencji w wymiarze pozasystemowym, a także geostrategie.

Zachodu, Globalnego Południa albo Globalnej Północy, ale także tworzących się już centrów systemowych przestrzeni cyfrowej, konkurujących zarówno w wymiarze produktywności gospodarek, jak i w wymiarze mobilności talentów. Rekomendacje UNESCO budowane są w odniesieniu do definicji systemu AI przyjętej przez OECD. Unikalne i kluczowe jest dla nich ustalenie kompasu czterech wymiarów dla rozstrzygania konfliktów między zasadami i regułami etycznymi systemów AI. wymiarami tymi są: godność ludzka, dobrostan, nieczynienie szkody i autonomia człowieka [4]. Istotne jest przy tym, że zgodnie z mandatem UNESCO rekomendacje, choć są instrumentem tzw. miękkiego prawa, to jednak winny być stosowane w domenie edukacji, nauki, kultury, a także komunikacji i mediów.

Koncepcja *trustworthy AI* uwzględnia zarówno wąskie podejście do cyberbezpieczeństwa AI (w wymiarze poufności, integralności, dostępności danych), jak i szerokie, oparte na właśnie sztucznej inteligencji godnej zaufania oraz odporności systemowej (np. w wymiarze jakości czy identyfikowalności danych, interoperacyjności, dostępu do infrastruktury, łańcucha wartości, produktywności, autonomii użytkowników i suwerenności ekonomicznej oraz politycznej wspólnoty). Potrzebę szerokiego i kontekstowego podejścia do cyberbezpieczeństwa AI podkreślono także w ostatnim raporcie ENISA [5]. Wskazano w nim, że wzajemne relacje między cyberbezpieczeństwem a wiarygodnością systemu AI są oczywiste, uzupełniają się i nie można ich ignorować w celu zapewnienia prawidłowego funkcjonowania dowolnego systemu (AI, społecznego, politycznego czy środowiskowego) [5]. Niezbędne zatem jest już nie tylko ogólne monitorowanie systemu AI w złożonym środowisku, ale także wykrywanie nieprawidłowości zachowania spowodowanego cyberatakami wobec systemu AI i za pomocą systemu AI. Cechy pożądane u godnej zaufania sztucznej inteligencji, takie jak solidność, nadzór, dokładność, identyfikowalność, wyjaśnialność i przejrzystość z natury rzeczy wspierają i uzupełniają cyberbezpieczeństwo. Tym samym luka w standardach cyberbezpieczeństwa winna być uzupełniona w oparciu o test wiarygodności systemu AI. Ponadto, w praktyce powinno się uwzględniać ocenę ryzyka w całym cyklu życia tego systemu [3].

Ramy godnej zaufania AI a odporność systemowa

Koncepcja godnego zaufania systemu AI została ujęta w ramy trzech filarów: zgodności z prawem, z zasadami etycznymi dla godnej zaufania sztucznej inteligencji i solidności techniczno-organizacyjnej [3]. Filar etyczny został zdefiniowany w czterech zasadach, wypracowanych na podstawie praw i wolności podstawowych człowieka (poszanowanie ludzkiej autonomii, przeciwdziałanie szkodom, uczciwość i zrozumiałość) i w siedmiu regułach praktycznej implementacji zasad etycznych do systemu AI (przewodnia i nadzorcza rola człowieka, stabilność i bezpieczeństwo, ochrona prywatności i danych, różnorodność, niedyskryminacja i sprawiedliwość, dobrostan społeczny i środowiskowy, odpowiedzialność). Przy czym realizacja tych siedmiu reguł winna następować stale i równolegle, w całym cyklu życia systemu AI. Powinien być on także poddawany ciągłej ocenie urzeczywistniania tych reguł. Filar solidności systemu AI został uzupełniony o solidność techniczną oraz społeczną (normalizacja i standardy techniczne, zasady interoperacyjności, klasyfikacja ryzyka, ocena wpływu, korzyści implementacji).

Istotą budowania odporności systemu AI i odporności wspólnoty jest zatem całościowe i skoordynowane zarządzanie badaniami, projektowaniem, rozwojem, wprowadzeniem na rynek, a ewentualnie – podejmowaniem decyzji o wycofaniu z rynku systemu AI.

W kontekście cyklu życia systemu AI wiarygodność odnosi się nie tylko do parametrów technologii jako takiej, ale również do właściwości systemów społeczno-technicznych, w ramach których stosuje się rozwiązania z zakresu technologii sztucznej inteligencji [3]. Systemy te obejmują ludzi, państwa, przedsiębiorstwa, infrastrukturę, oprogramowanie komputerowe, protokoły, normy, mechanizmy zarządzania, obowiązujące przepisy, mechanizmy nadzoru, struktury zachęt, procedury kontroli, procedury przekazywania informacji na temat najlepszych praktyk, a także inne elementy, jak np. rejestry rozproszone dostępu do danych czy biblioteki algorytmów lub kodów. Podobnie, jak miało to miejsce w przypadku (utraty) zaufania do lotnictwa, energii jądrowej lub bezpieczeństwa żywności, zaufanie do systemu AI (lub jego brak) nie jest związane z samymi jego cechami, ale z ogólnym kontekstem, w jakim jest on wykorzystywany. W tym przypadku dotyczy to odporności wspólnoty.

Zatem dążenie do stworzenia odpornego systemu AI, opartego o ramy godnej zaufania sztucznej inteligencji (w wymiarze prawnym, etycznym i solidnościowym), wiąże się nie tylko z koniecznością zagwarantowania, że sam system będzie odporny i godny zaufania. Wymaga ono również stosowania całościowego i systemowego podejścia, obejmującego zdolność budowania i utrzymania odporności oraz wiarygodność wszystkich podmiotów i procesów tworzących kontekst społeczno-techniczny funkcjonowania systemu AI przez cały jego cykl życia.

Zasoby systemu AI

Na rozwój i skalowanie systemu AI wpływ mają głównie trzy komponenty: moce obliczeniowe, wiedza i dane oraz zasoby ludzkie w zakresie modelowania i budowania systemów AI. Każdy z nich ma decydujące znaczenie dla odporności sztucznej inteligencji oraz wspólnoty. W obszarze każdego z tych elementów toczy się gorąca rywalizacja i konkurencja. Dostęp do mocy obliczeniowych[4] – czy to ulokowanych w architekturze lokalnej, sieciach brzegowych czy usługach chmurowych lub dedykowanych mikroprocesorach *(chips)* – jest krytyczny, ale niewystarczający dla zbudowania wydajnego, wiarygodnego i solidnego systemu AI. Tym bardziej gdy dane wejściowe nie przejdą procesu wydobycia z nich jakości do tzw. *small data*, a talenty kreatorów, inżynierów, menadżerów lub analityków danych nie wydobędą z nich unikalnego, ale odpornego, inteligentnego rozwiązania (*smart solution)*. Priorytetowe jest zatem także zarządzanie wiedzą, akumulacja kapitału własnego i umiejętność skalowania. I jest to megatrend przyszłości, decydujący o produktywności gospodarek, dobrostanie społeczeństwa oraz odporności wspólnoty.

Strategicznymi wymiarami odporności systemu AI i wspólnoty jest budowa i utrzymanie zdolności (eko)systemu, aparatu skutecznych instytucji wykonawczych oraz warunków brzegowych odporności.

4 Np. *central procesor units* (CPU), *graphic procesor units* (GPU), *neural procesor units* (NPU), tensor procesor units (TPU), *quantum procesor units* (QPU)

Odporność systemowa wspólnoty

Można wyróżnić co najmniej cztery współzależne warstwy odporności systemu AI i odporności systemowej, w których mogą być one zarówno mierzone, jak i strategicznie planowane i kształtowane:

- zdolności zapewnienia zasobów, w tym mocy obliczeniowych i infrastruktury, kadr i danych wysokiej jakości – zarówno w zakresie istniejącego łańcucha dostaw oraz potrzeb wspólnoty, jak również diagnozy stosunku wykorzystania tych zasobów przez głównych aktorów w wiodących sektorach;

- skuteczności aparatu instytucji wykonawczych – zarówno w obszarze ludzkim (przewodnictwa, umiejętności, edukacji, szkoleń i różnorodności), środowiska politycznego (prawa stanowionego, regulacji, strategii, programów i planów interwencyjnych i odbudowy), innowacji (rynku pracy, laboratoriów, badań i rozwoju), jak i w obszarze dostępności technologii (kosztów, praw licencji dostępowych);

- odporności brzegowej – zarówno w zakresie bezpieczeństwa i suwerenności wspólnoty wobec lokalizacji zasobów, ich własności oraz łańcuchów dostaw, jak i jej zrównoważonego rozwoju, efektywności wykorzystania zasobów i ich wpływu na otoczenie społeczne, środowiskowe i polityczne [6];

- autonomii pojedynczego człowieka w interakcjach z systemem AI; mowa tu zarówno o wymiarze godności ludzkiej, jak i warunków dla krytycznego myślenia, wyjaśnialności, transparentności czy audytowalności systemów AI; celem jest wsparcie rozwoju osobistego, edukacji, świadomego uczestnictwa w rynku pracy, rozwoju innowacji i ich komercjalizacji, a także wspieranie uczciwej konkurencji, czy wreszcie – podmiotowego niezakłóconego uczestnictwa w życiu politycznym, społecznym i gospodarczym wspólnoty [7].

Zarządzanie i odporność systemu AI

Ramy interoperacyjności zarządzania ryzykiem AI wysokiego poziomu obejmują zarządzanie ryzykiem w całym cyklu życia godnej zaufania sztucznej inteligencji [8].

Podejścia do zarządzania ryzykiem stosowane w całym cyklu życia systemu sztucznej inteligencji mogą identyfikować, oceniać, ustalać priorytety i rozwiązywać sytuacje, które mogą niekorzystnie wpłynąć na zachowanie i wyniki działania systemu. Na podstawie ram zarządzania ryzykiem sztucznej inteligencji NIST [9], ram zarządzania ryzykiem ISO 31000 [10], wytycznych OECD dotyczących należytej staranności [11] i innych, można zidentyfikować cztery kroki zarządzania ryzykiem systemu AI przy jednoczesnym zapewnieniu poszanowania praw człowieka, wartości demokratycznych i praworządności:

- ZDEFINIOWANIE ZAKRESU, kontekstu i kryteriów, w tym odpowiednich zasad systemu AI, interesariuszy i aktorów dla każdej fazy cyklu życia systemu AI;
- OCENA RYZYKA godnej zaufania sztucznej inteligencji, przy jednoczesnej identyfikacji i analizie problemów na poziomie indywidualnymi społecznym oraz ocenie prawdopodobieństwa i poziomu szkody (np. małe ryzyko może sumować się do większego ryzyka);
- LIKWIDACJA ZAGROŻENIA w celu łagodzenia niekorzystnych skutków lub zapobiegania im, proporcjonalnie do prawdopodobieństwa i zakresu każdego z nich;
- ADMINISTROWANIE PROCESEM zarządzania ryzykiem poprzez osadzenie i podtrzymywanie kultury zarządzania ryzykiem w organizacjach; bieżące monitorowanie i przegląd procesu; dokumentowanie, komunikowanie i konsultowanie procesu i jego wyników.

Cykl życia obejmuje planowanie i projektowanie, zbieranie i procesowanie danych, budowę i użycie modeli, weryfikację i walidację systemu AI, jego rozwój, wprowadzenie na rynek, operowanie i monitorowanie, a także ewentualną utylizację.

Opracowany w ramach OECD proces zarządzania ryzykiem zachodzić powinien według zasad godnej zaufania sztucznej inteligencji, wyznaczonych w Rekomendacjach OECD dla AI [1]. Skupiać się winien na pięciu obszarach: (1) korzyściach dla ludzi i planety, (2) koncentracji na człowieku, (3) uczciwości, przejrzystości i wyjaśnialności, (4) solidności, zabezpieczeniach i bezpieczeństwie, a ponadto (5) rozliczalności i odpowiedzialności.

Ocena ryzyka winna być przeprowadzona adekwatnie do ram kwalifikacyjnych typów i sposobów zastosowań systemów sztucznej inteligencji, również opracowanych przez ekspertów OECD [12].

Sztuczna inteligencja w domenie praw człowieka, demokracji i praworządności jest obecnie przedmiotem prac legislacyjnych w ramach Rady Europy. Na podstawie wyników prac grupy ekspertów CAHAI[5], aktualnie w wymiarze politycznym reprezentantów państw członkowskich CAI[6], projektowany jest precedensowy akt prawny. Wynikiem tych prac ma być przyjęcie instrumentu tzw. twardego prawa – *binding instrument*, co pozwoliłoby nie tylko sankcjonować, ale i umocnić w systemie prawnym rekomendacje dla samoregulacji etyki sztucznej inteligencji [13] [14].

Wnioski

Budowanie odporności systemu AI na rzecz odporności wspólnoty winno wspierać autonomię pojedynczego człowieka w relacjach z maszynami, ochronę własności intelektualnej i talentów, wzajemne zasady i standardy interoperacyjności oraz znaczniki danych. Powinno uwzględniać także własność i dostęp do infrastruktury, uczciwą konkurencję i niezależność w łańcuchach budowania wartości i łańcuchach dostaw, autonomiczną suwerenność w obszarze ekonomii, bezpieczeństwo, cyberbezpieczeństwo i obronność wspólnoty.

Wszystkie te obszary pozostają ze sobą w relacji współzależności i stale wchodzą w interakcje. Z jednej strony rozszerza to kontekst budowania odporności systemów AI dla odporności wspólnoty, z drugiej jednak może

5 Ad hoc Committee on Artificial Intelligence.

6 Committee on Artificial Intelligence.

tę odporność wzmacniać przez strategiczne, skoordynowane zarządzenie wiedzą, ryzykiem i dostępem do infrastruktury, dzielenie się danymi, źródłami i sposobami finasowania, a w końcu – skalowanie w relacjach transgranicznych czy geopolitycznych.

Dla zapewnienia odporności systemu AI dla odporności wspólnoty niezbędne jest zatem zespolenie działań w zakresie prawa stanowionego, sankcjonującego ramy etyczne, uczciwą konkurencję oraz zasadę wzajemnego uznawania reguł interoperacyjności oraz oceny i walidacji ryzyk. Dotyczy to również ryzyk mających negatywny wpływ na rozwój innowacji służących podnoszeniu odporności wspólnoty. Stąd, zgodnie z *Polityką dla rozwoju sztucznej inteligencji w Polsce od roku 2020* [7], dla zapewnienia odporności wspólnoty Polski i jej sojuszników istotne jest wprowadzenie do systemu otwartych ram innowacji, szanujących wymiar prawny, etyczny, techniczno-standaryzacyjny, wolności gospodarczej oraz suwerenności politycznej, a także praw człowieka, demokracji i praworządności. Również w tym celu Polska przystąpiła w grudniu 2020 r. do *Global Partnership on Artificial Intelligence* (GPAI), którego mandatem jest wdrożenie w praktycznych rozwiązaniach politycznych wytycznych dla godnej zaufania sztucznej inteligencji według rekomendacji OECD. Na ostatnim szczycie GPAI, jaki miał miejsce w Tokio, w grudniu 2022 r., państwa członkowskie, działając poprzez swoich przedstawicieli, odnowiły treść deklaracji założycielskiej i podkreśliły potrzebę budowy odporności wspólnoty oraz wzmocnienia obywateli przy rozwoju sztucznej inteligencji [14].

Bibliografia

[1] OECD (22 maja 2019). Recommendation of Council of OECD on Artificial Intelligence [Online]. Dostęp: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, przyjęte jako OECD AI Principles [Online]. Dostęp: https://oecd.ai/en/ai-principles.

[2] Niezależna grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji (8 kwietnia 2019). *Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji* [Online]. Dostęp: https://op.europa.eu/pl/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[3] UNESCO (23 listopada 2021). *Recommendations on Ethics of AI* [Online]. Dostęp: https://unesdoc.unesco.org/ark:/48223/pf0000381137

[4] ENISA (14 marca 2023). *Cyberbezpieczeństwo AI i standaryzacja* [Online]. Dostęp: https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation

[5] OECD (28 lutego 2023). *A blueprint for building national computing capacity for artificial intelligence* [Online]. Dostęp: https://www.oecd-ilibrary.org/science-and-technology/a-blueprint-for-building-national-compute-capacity-for-artificial-intelligence_876367e3-en

[6] Uchwała nr 196 Rady Ministrów z dnia 28 grudnia 2020 r. w sprawie ustanowienia „Polityki dla rozwoju sztucznej inteligencji w Polsce od roku 2020". Dostęp:. https://www.gov.pl/web/aipolityka-dla-rozwoju-sztucznej-inteligencji-w-polsce-od-roku-2020

[7] OECD (23 lutego 2023). *Advancing accountability in AI. Governing and managing risk throughout life cycle for trustworthy AI* [Online]. Dostęp: https://www.oecd-ilibrary.org/science-and-technology/advancing-accountability-in-ai_2448f04b-en

[8] OECD (22 lutego 2022). *Framework for Classification of AI Systems* [Online]. Dostęp https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en

[9] Ad Hoc Committee On Artificial Intelligence (CAHAI) (17 grudnia 2020). *Feasability Study* [Online]. Dostęp: https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

[10] Ad hoc Committee on Artificial Intelligence (CAHAI) (3 grudnia 2021). *Possible elements of of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law* [Online]. Dostęp: https://rm.coe.int/cahai-2021-09rev-elements/1680a6d90d

[11] *Artificial Intelligence Risk Management Framework*, NIST AI 100-1, 2023.

[12] Risk management, ISO 31000, 2018.

[13] OECD (2018). *Due Diligence Guidance for Responsible Business Conduct* [Online]. Dostęp: http://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf

[14] GPAI (2022). Ministerial Declaration, GPAI Tokio GPAI Summit [Online]. Dostęp: https://www.gpai.ai/events/tokyo-2022/ministerial-declaration/

Unijne podejście do sztucznej inteligencji

Monika Stachoń

NASK-PIB | Centrum Cyberbezpieczeństwa i Infrastruktury |
Dział Strategii i Rozwoju Bezpieczeństwa Cyberprzestrzeni
Uniwersytet Warszawski | Wydział Nauk Politycznych i Studiów Międzynarodowych

Wprowadzenie

„Technologie cyfrowe, w szczególności sztuczna inteligencja, w niespotykanym dotąd tempie zmieniają świat. Zmieniły sposób, w jaki się porozumiewamy, żyjemy, pracujemy. Zmieniły nasze społeczeństwa i gospodarki" [1]. W politycznych wytycznych dla Komisji Europejskiej na lata 2019-2024 Ursula von der Leyen podkreśliła, że w związku z dokonującą się transformacją w sferze cyfrowej konieczne jest znalezienie własnej, europejskiej drogi, która pozwoli zachować równowagę między przepływem i szerokim wykorzystywaniem danych a wysokim poziomem prywatności, bezpieczeństwa i norm etycznych.

W maju 2017 roku KE opublikowała śródokresowy przegląd strategii na rzecz jednolitego rynku cyfrowego. Podkreślono w nim rolę, jaką może odegrać sztuczna inteligencja w przyszłym wzroście gospodarczym, pod warunkiem pełnego wykorzystania przez Unię Europejską jej mocnych

stron w dziedzinie nauki i przemysłu. Komisja zaznaczyła, że rozważy ewentualną potrzebę dostosowania obecnych ram prawnych do rozwoju nowych technologii, w tym AI [2]. W tym samym roku Parlament Europejski opublikował zalecenia w sprawie przepisów prawa cywilnego dotyczących robotyki, w których zawarte zostały m.in. definicja i klasyfikacja „inteligentnych robotów", zasady ich rejestracji, a także kodeks postępowania etycznego dla inżynierów robotyki [3]. W październiku 2017 roku Rada Europejska zwróciła się do Komisji z wnioskiem o przedstawienie europejskiego podejścia do problematyki sztucznej inteligencji [4].

W związku z coraz szerszą debatą społeczną w Unii Europejskiej w tej materii, Komisja Europejska uznała, że należy stworzyć solidne europejskie ramy w zakresie AI. Stała się ona bowiem jedną ze strategicznych technologii XXI wieku. Celem artykułu jest przedstawienie unijnego podejścia do sztucznej inteligencji. Omówione zostaną w nim wybrane zagadnienia zawarte przede wszystkim w unijnych dokumentach strategicznych oraz obowiązujących i proponowanych aktach prawnych.

Sztuczna inteligencja dla Europy

Pierwszym strategicznym dokumentem przyjętym przez Komisję Europejską w zakresie sztucznej inteligencji jest Komunikat Komisji zatytułowany *Sztuczna inteligencja dla Europy* [5]. Opublikowany został 25 kwietnia 2018 roku i powszechnie uznaje się go za unijną strategię w zakresie AI. To właśnie w tym dokumencie po raz pierwszy na gruncie unijnym pojawiła się definicja sztucznej inteligencji. Termin ten oznacza systemy, które „wykazują inteligentne zachowanie dzięki analizie otoczenia i podejmowaniu działań – do pewnego stopnia autonomicznie – w celu osiągnięcia konkretnych celów" [5]. Komisja podkreśliła także w dokumencie wagę, jaką mają prywatne inwestycje w systemy wykorzystujące sztuczną inteligencję, deklarując tym samym kontynuację prac nad stworzeniem środowiska sprzyjającego tego rodzaju działaniom.

W komunikacie przedstawiona została europejska inicjatywa w sprawie sztucznej inteligencji, która opiera się na trzech filarach:

RYS. 1. Filary europejskiej inicjatywy w sprawie AI *(opracowanie własne)*

W zakresie pierwszego z powyższych filarów, najważniejszym postulatem było zwiększenie poziomu inwestycji w AI, który do końca 2020 roku miał wynieść przynajmniej dwadzieścia miliardów euro. Komisja zapewniła również, że będzie wspierać technologie AI zarówno w badaniach podstawowych, jak i przemysłowych. Zobowiązała się również do wspierania i wzmacniania centrów doskonałości AI w całej Europie, a także ułatwienia dostępu do sztucznej inteligencji wszystkim potencjalnym użytkownikom. KE przewidywała w Komunikacie także wsparcie badań i eksperymentów w zakresie sztucznej inteligencji. Z kolei po 2020 roku planowała inwestować m.in. w badania i innowacje w takich dziedzinach, jak uczenie się maszyn bez nadzoru czy wydajność baz danych. Komisja zamierzała pracować również nad zwiększeniem zakresu udostępniania danych i informacji do ponownego wykorzystania, w tym do uczenia systemów AI [5].

Drugi priorytet, który znalazł się w unijnej strategii na rzecz AI, oparty był na przekonaniu, że rozwój sztucznej inteligencji doprowadzi do zmian społeczno-gospodarczych, takich jak zmiana charakteru pracy. W tym zakresie UE stoi przed trzema głównymi wyzwaniami: (1) przygotowanie całego społeczeństwa, tj. rozwijanie podstawowych umiejętności cyfrowych i takich, których technologia nie będzie w stanie zamienić; (2) pomoc pracownikom na stanowiskach, które prawdopodobnie ulegną największym przeobrażeniom; (3) przeszkolenie większej liczby specjalistów w zakresie AI. Dlatego też Komisja przyjęła podejście *leaving no one behind* („nikt nie pozostanie w tyle"), promując i uruchamiając programy

w zakresie nabywania umiejętności cyfrowych, kompetencji kluczowych, a także edukacji cyfrowej [5].

Trzecim filarem było stworzenie odpowiednich ram etycznych i prawnych dla systemów sztucznej inteligencji. Komisja zobowiązała się m.in. do opracowania projektu wytycznych dotyczących etyki AI do końca 2018 roku, zbadania, czy funkcjonujące ramy prawne dotyczące bezpieczeństwa oraz odpowiedzialności na poziomie unijnym i krajowym wymagają modernizacji, a także opublikowania w 2019 roku wytycznych w zakresie interpretacji dyrektywy w sprawie odpowiedzialności za produkt w świetle postępu technologicznego [5].

Skoordynowany plan w zakresie sztucznej inteligencji

Drugim strategicznym dokumentem w zakresie unijnego podejścia do sztucznej inteligencji jest Komunikat Komisji z dnia 7 grudnia 2018 roku, pt. *Skoordynowany plan w zakresie sztucznej inteligencji* [6]. Jego opracowanie wynikało ze zobowiązania Komisji podjętego w strategii w sprawie sztucznej inteligencji dla Europy [5]. Obejmuje on głównie działania na lata 2019 i 2020, ale KE przewidziała jego realizację również w kolejnej dekadzie. Głównymi celami planu są:

- zwiększenie inwestycji na poziomie unijnym i krajowym,
- wspieranie synergii i współpracy w całej UE, w tym w dziedzinie etyki,
- wspieranie wymiany najlepszych praktyk oraz
- wspólne określenie dalszych działań.

Idea, jaka przyświecała KE podczas opracowywaniu planu, to maksymalizacja korzyści, które sztuczna inteligencja może przynieść wszystkim Europejczykom poprzez zapewnienie wsparcia dla rozwoju niezawodnej AI, która odpowiada europejskim wartościom etycznym i aspiracjom obywateli. W dokumencie określono siedem głównych obszarów, które są kluczowe w zakresie rozwoju sztucznej inteligencji w Europie, i w których państwa członkowskie powinny podjąć działania [6].

DZIAŁANIA STRATEGICZNE I KOORDYNACJA

Po pierwsze, Skoordynowany plan tworzy ramy strategiczne dla krajowych strategii w zakresie sztucznej inteligencji, które powinny powstać w państwach członkowskich do połowy 2019 roku. Uzasadnieniem tego jest zwiększanie poziomu inwestycji, połączenie kluczowych zasobów (takich jak dane) i dostęp do nich, a także zapewnienie jednolitego otoczenia regulacyjnego w całej UE.

ZWIĘKSZANIE INWESTYCJI POPRZEZ PARTNERSTWA

Po drugie, KE kładzie nacisk na zwiększenie współpracy i partnerstw w zakresie inwestowania w sztuczną inteligencję, zarówno w sektorze publicznym, jak i prywatnym. Europa dążyć ma – jako całość – do osiągnięcia strategicznej autonomii w kontekście AI. Komisja w tym obszarze przedstawiła szereg inicjatyw, które mają służyć osiągnięciu tego celu, takich jak (1) zainicjowanie partnerstw publiczno-prywatnych; (2) udostępnienie zasobów przedsiębiorstwom typu start-up oraz innowatorom w zakresie AI i technologii blockchain; (3) zachęcanie krajowych banków rozwoju do uczestnictwa w programach wsparcia; (4) wspieranie innowacji za pośrednictwem Europejskiej Rady ds. Innowacji.

Z LABORATORIUM NA RYNEK

Trzeci element Skoordynowanego planu dotyczy stworzenia zaplecza naukowo-badawczego dla sztucznej inteligencji i obejmuje trzy kwestie:

- budowanie doskonałości naukowej,
- ustanawianie światowej rangi ośrodków badawczych,
- przyspieszanie wdrażania SI poprzez centra innowacji cyfrowych.

Komisja zobowiązała się do zwiększenia inwestycji w badania naukowe i innowacje, a także włączenia AI do wszystkich dziedzin, w odniesieniu do których można ją opracować lub wykorzystać jej korzyści. Podkreśliła również cel, który powinien przyświecać państwom członkowskim, jakim jest wspieranie współpracy między najlepszymi zespołami badawczymi w Europie. KE wskazała również, że kluczowym elementem

opracowywanej w Europie sztucznej inteligencji musi być uwzględnianie etyki oraz bezpieczeństwa na etapie projektowania.

UMIEJĘTNOŚCI I UCZENIE SIĘ PRZEZ CAŁE ŻYCIE

W tym obszarze Komisja skupiła się przede wszystkim na wyzwaniach, jakie generować będzie sztuczna inteligencja w zakresie rynku pracy. Jak podkreślono we wstępie do dokumentu, „Europa musi mieć zdolność szkolenia, przyciągania i zatrzymywania tego typu kadr (w sektorze technologicznym) oraz musi promować przedsiębiorczość oraz zwiększać w tym kontekście różnorodność i równowagę płci" [6].

Komisja przedstawiła szereg inicjatyw, które do tej pory zostały podjęte, w celu realizacji tego punktu programu, takie jak plan działania na rzecz edukacji cyfrowej czy pilotażowy program staży „Cyfrowe możliwości". Podkreśliła także konieczność zaadresowania w krajowych strategiach w zakresie AI kwestii niedoboru wykwalifikowanych w tym obszarze pracowników. Zasygnalizowała również kontynuację prac nad wzajemnym uznawaniem dyplomów i świadectw w przypadku badań dotyczących dyscyplin takich jak sztuczna inteligencja.

TWORZENIE WSPÓLNEJ EUROPEJSKIEJ PRZESTRZENI DANYCH

Wyszczególnienie tego priorytetu wynika z faktu, iż dane generowane i przechowywane przez podmioty sektora publicznego są często bardzo wysokiej jakości i stanowią ważne zasoby dla europejskich innowatorów i przedsiębiorstw. W związku z tym KE podkreśliła konieczność podjęcia działań ułatwiających wymianę danych będących w posiadaniu sektora publicznego i prywatnego poprzez stworzenie wspólnej europejskiej przestrzeni danych – jednolitego obszaru cyfrowego, który umożliwi rozwój nowych produktów i usług.

Komisja kładzie również nacisk na interoperacyjność danych, zwłaszcza poprzez wspólne formaty danych, które są otwarte, odnajdywalne, dostępne, interoperacyjne i nadające się do ponownego wykorzystania oraz odczytu maszynowego, a także znormalizowane i udokumentowane.

UWZGLĘDNIANIE ETYKI NA ETAPIE PROJEKTOWANIA I RAMY PRAWNE

Obszar ten składa się z dwóch wzajemnie zależnych od siebie elementów: etyki oraz regulacji prawnych w zakresie sztucznej inteligencji. W pierwszym z nich KE podkreśliła, że projektowana w Europie AI powinna być przewidywalna, wiarygodna, stosowana odpowiedzialnie i z poszanowaniem praw podstawowych, oraz zgodna z zasadami etycznymi. Tylko w takim przypadku społeczeństwa będą w stanie ją zaakceptować i z niej korzystać. W tym zakresie Komisja planuje opublikowanie wytycznych dotyczących etyki związanej z AI. Równocześnie kwestie etyczne ściśle wiążą się z wysokim poziomem ochrony i bezpieczeństwa, którymi charakteryzować się powinna sztuczna inteligencja. W związku z tym, KE zobowiązała się do przedstawienia sprawozdania dotyczącego ram prawnych, w tym ewentualnych luk oraz kierunku rozwoju prawodawstwa.

SZTUCZNA INTELIGENCJA DLA SEKTORA PUBLICZNEGO

Komisja Europejska, podążając za myślą przewodnią, jaką jest dążenie do poprawy jakości usług publicznych poprzez zastosowanie sztucznej inteligencji, zaproponowała również sukcesywne zwiększanie starań podejmowanych na rzecz absorpcji sztucznej inteligencji przez sektory interesu publicznego, takie jak opieka zdrowotna, transport, bezpieczeństwo czy edukacja. Obejmuje to np. określenie obszarów na potrzeby wspólnych zamówień w zakresie AI, zaoferowanie podmiotom administracji publicznej państw członkowskich usługi *eTranslation* czy przeznaczenie środków na eksperymenty z usługami wykorzystującymi sztuczną inteligencję.

Biała księga w sprawie sztucznej inteligencji

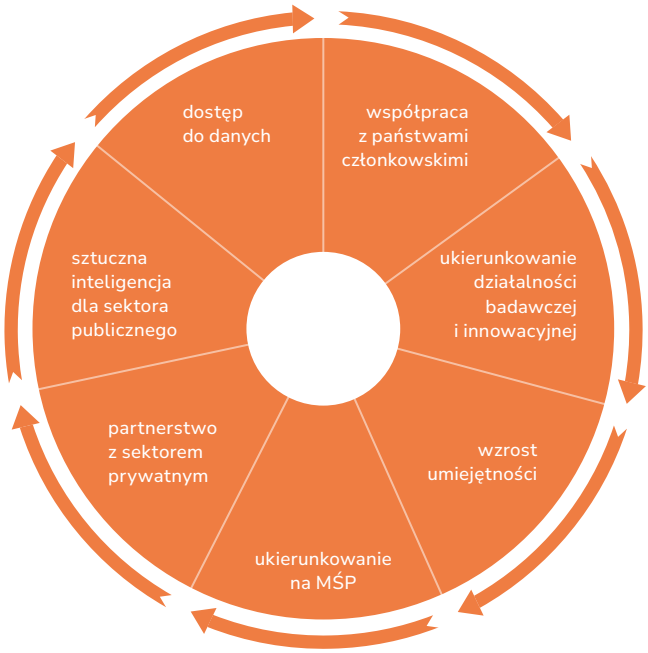
Jednym z kluczowych dokumentów, składających się na unijną strategię w zakresie sztucznej inteligencji, jest opublikowana 19 lutego 2020 roku *Biała księga w sprawie sztucznej inteligencji. Europejskie podejście do doskonałości i zaufania* [7]. Odzwierciedlone w niej zostało wypracowane wcześniej przez KE podejście regulacyjne i inwestycyjne, którego podwójnym celem jest promowanie stosowania sztucznej inteligencji

i zaadresowanie zagrożeń związanych z niektórymi jej zastosowaniami. Celem Białej księgi jest określenie wariantów strategicznych dotyczących sposobów osiągnięcia tych założeń, a główne zawarte w niej elementy to:

- RAMY POLITYCZNE określające środki służące połączeniu wysiłków na szczeblu europejskim, krajowym i regionalnym, w celu stworzenia „ekosystemu doskonałości";
- KLUCZOWE ELEMENTY PRZYSZŁYCH RAM REGULACYJNYCH dotyczących sztucznej inteligencji w Europie, które stworzą wyjątkowy „ekosystem zaufania".

EKOSYSTEM DOSKONAŁOŚCI

Zgodnie z koncepcją Komisji Europejskiej, rozwój „ekosystemu doskonałości" pomóc ma we wspieraniu rozwoju i stosowania sztucznej inteligencji w całej gospodarce UE i w administracji publicznej. Aby mogło się to dokonać, KE postuluje podjęcie działań w kilku obszarach, które zostały wypisane na rysunku poniżej (rys. 2).



RYS. 2. Obszary tworzące „ekosystem doskonałości" *(opracowanie własne)*

Warto zauważyć, że Komisja kładzie nacisk przede wszystkim na ukierunkowanie działań społeczności badawczej i innowacyjnej, tj. stworzenie większej synergii i rozwój sieci między różnymi europejskimi ośrodkami badawczymi oraz skoordynowanie ich działań. Ma to na celu poprawę doskonałości, zatrzymywanie i przyciąganie najlepszych naukowców oraz opracowywanie bezpiecznych i wysokiej jakości technologii. Aby zapobiec istniejącemu niedoborowi kompetencji, KE wspierać będzie również zwiększanie umiejętności w zakresie sztucznej inteligencji.

„Ekosystem doskonałości" tworzyć ma jak najwięcej podmiotów, dlatego w Białej księdze przewidziano także działania na rzecz zapewnienia dostępu do sztucznej inteligencji zarówno małym i średnim przedsiębiorstwom (poprzez wzmocnienie ośrodków innowacji cyfrowych i udostępnienie platformy dostępnych na żądanie usług z zakresu AI), sektorowi prywatnemu (który zachęca się do określania programu badań naukowych i innowacji oraz zapewniania niezbędnego poziomu współinwestycji w sztuczną inteligencję), jak i sektorowi publicznemu (szczególnie w obszarach administracji publicznej, opieki zdrowotnej i transportu).

Zasadniczą kwestią dla stworzenia „ekosystemu doskonałości" jest również poprawa dostępu do danych i zarządzania nimi, bez których rozwój sztucznej inteligencji i innych zastosowań cyfrowych nie jest możliwy.

EKOSYSTEM ZAUFANIA

Drugi z obszarów Białej księgi dotyczy europejskich ram regulacyjnych w zakresie godnej zaufania sztucznej inteligencji. Mają one chronić wszystkich obywateli Unii i pomóc stworzyć rynek wewnętrzny dla dalszego rozwoju i upowszechniania AI, a także wzmocnić bazę przemysłową Europy w tej dziedzinie.

W Białej księdze Komisja Europejska odniosła się do kilku aspektów przyszłych ram prawnych. Po pierwsze, określono problem, którego dotyczyć powinna regulacja AI na szczeblu europejskim, tj. minimalizację różnego rodzaju ryzyk związanych z potencjalnymi szkodami, szczególnie w zakresie praw podstawowych (w tym przepisów dotyczących ochrony danych osobowych i prywatności oraz niedyskryminacji), a także kwestii związanych z bezpieczeństwem i odpowiedzialnością sztucznej inteligencji. Po drugie, KE wskazała na konieczność ulepszenia istniejących

ram legislacyjnych, a także stworzenia odpowiednich ram prawnych w zakresie sztucznej inteligencji. Te drugie mają opierać się na analizie ryzyka i gwarantować wystarczającą skuteczność przy równoczesnym niezbyt dużym poziomie prawnych nakazów dla podmiotów podlegających temu prawu. Komisja postuluje objęcie nowymi regulacjami zastosowań AI charakteryzujących się „wysokim ryzykiem", określanych jako takie na podstawie zaproponowanych kryteriów.

Kolejnymi kwestiami, które powinny znaleźć się w unijnych aktach prawnych z zakresu sztucznej inteligencji, są wymogi dotyczące w szczególności danych szkoleniowych, przechowywania danych i prowadzenia rejestrów, solidności i dokładności oraz sprawowania nadzoru przez człowieka. KE określa również zakres przyszłych przepisów dotyczących przestrzegania i egzekwowania prawa oraz utworzenia europejskiej struktury zarządzania w zakresie sztucznej inteligencji w formie ram współpracy właściwych organów krajowych. Wskazuje również na konieczność wprowadzenia dobrowolnego etykietowania zastosowań AI, które nie kwalifikują się jako zastosowania „wysokiego ryzyka".

Pakiet dotyczący sztucznej inteligencji

Podążając za przyjmowanymi we wcześniejszych dokumentach kierunkami rozwoju sztucznej inteligencji w Europie, 21 kwietnia 2021 roku Komisja Europejska zaprezentowała pakiet dotyczący AI, w ramach którego znalazła się propozycja prawnego uregulowania kwestii sztucznej inteligencji (AI Act) wraz z oceną jego skutków [8]. Zaproponowany akt prawny prezentuje dosyć zachowawcze, wyważone i proporcjonalne horyzontalne podejście do sztucznej inteligencji. KE ograniczyła się do określenia minimalnych wymogów niezbędnych do zarządzania ryzykiem i zapobiegania problemom związanym ze sztuczną inteligencją przy jak najmniejszej ingerencji w rozwój technologiczny oraz zminimalizowaniu nadmiernych kosztów produkcji systemów AI.

W nawiązaniu do postulatów zawartych w Białej księdze, w proponowanym rozporządzeniu wprowadzono proporcjonalny system regulacyjny, który oparty jest na analizie ryzyka. Komisja określiła kategorie systemów, które uznaje się za wysokiego ryzyka, co pozwala ograniczyć interwencję regulacyjną do konkretnych sytuacji. Zapobiegnięto tym samym

nadmiernej regulacji, która mogłaby ograniczać rozwój systemów sztucznej inteligencji. Więcej informacji na temat AI Act znaleźć można w innych częściach niniejszej publikacji.

Wytyczne w sprawie etyki oraz polityki i inwestycji w zakresie sztucznej inteligencji

Na koniec warto wspomnieć, że europejskie podejście do sztucznej inteligencji nie opiera się wyłącznie na dokumentach strategicznych lub proponowanych przez Komisję Europejską aktach prawnych. Jak zostało wspomniane we wstępie, również inne instytucje unijne (w tym PE oraz Rada) zajmują się różnymi aspektami AI. Należy też zauważyć, że 1 czerwca 2018 roku w ramach KE powstała złożona z pięćdziesięciu dwóch ekspertów Grupa ekspercka ds. Sztucznej Inteligencji (*High-Level Expert Group on Artificial Intelligence*, HLEG AI), której zadaniem było opracowanie rekomendacji w zakresie rozwoju polityki sztucznej inteligencji. Grupa ta w 2019 roku przedstawiła dokumenty zawierające wytyczne dla tworzenia i funkcjonowania godnej zaufania sztucznej inteligencji w dwóch obszarach: etyki [9] oraz polityki i inwestycji [10].

W wytycznych w sprawie etyki eksperci przedstawili trzy zasadnicze cechy, którymi powinna charakteryzować się godna zaufania sztuczna inteligencja. Są to: zgodność z prawem, etyczność oraz solidność (zarówno w technicznym, jak i społecznym aspekcie). W dokumencie zawarto również cztery kluczowe zasady etyczne, które powinna uwzględniać technologia korzystająca ze sztucznej inteligencji: poszanowanie autonomii człowieka, zapobieganie szkodom, sprawiedliwość, a także możliwość wyjaśnienia decyzji podjętych przez system AI. Ostatnim elementem, który został przedstawiony w wytycznych, było siedem wymogów dla technologii korzystającej ze sztucznej inteligencji. Eksperci uznali, że tego rodzaju systemy powinny charakteryzować się solidnością techniczną i bezpieczeństwem (tj. dbać o bezpieczeństwo użytkowników i swoją niezawodność oraz radzić sobie z błędami), ochroną prywatności i danych, przejrzystością (zarówno w zakresie wiedzy o wkomponowanych algorytmach, jak i o procesie podejmowania decyzji), odpowiedzialnością oraz niedyskryminacją i sprawiedliwością. Technologie te powinny działać na rzecz dobrostanu społecznego i środowiskowego, a przewodnią i nadzorczą rolę nad nimi powinien pełnić człowiek [9].

Z kolei drugi z dokumentów, zawierający wytyczne w zakresie polityki i inwestycji, obejmuje trzydzieści trzy zalecenia skierowane do instytucji unijnych i państw członkowskich w czterech kluczowych obszarach: obywatele i społeczeństwo, sektor prywatny, sektor publiczny oraz badania i środowisko akademickie [10].

17 lipca 2020 roku Komisja Europejska przy wsparciu HLEG AI opublikowała Listę kontrolną dla godnej zaufania sztucznej inteligencji (ALTAI) [11]. Składa się ona z siedmiu sekcji odpowiadających rekomendacjom ekspertów tej grupy dla etycznej i godnej zaufania SI. Każda sekcja zawiera listę pytań o charakterze zamkniętym, a także glosariusz, który w przystępny sposób wyjaśnia stosowane pojęcia. Głównym celem listy ALTAI jest wsparcie zróżnicowanych innowacji w obszarze AI w Europie oraz uczynienie z etyki podstawowego filaru rozwoju tej technologii.

Podsumowanie

Europejskie podejście do sztucznej inteligencji ma na celu promowanie potencjału innowacyjnego Europy w tej dziedzinie, przy jednoczesnym wspieraniu rozwoju i wprowadzaniu etycznej i godnej zaufania sztucznej inteligencji w całej gospodarce. Przesłaniem płynącym z większości strategicznych dokumentów w tym obszarze jest to, że sztuczna inteligencja powinna działać na rzecz ludzi i społeczeństwa. Unijna strategia w zakresie sztucznej inteligencji promuje podejście, które w centrum rozwoju AI stawia człowieka, tj. jego dobrobyt i bezpieczeństwo, których osiągnięcie jest możliwe dzięki mocnym stronom państw europejskich w obszarach nauki i przemysłu. Równocześnie, KE zdaje sobie sprawę z tego, że świat cyfrowy wymaga zaufania budowanego w oparciu o bezpieczeństwo wykorzystywanych technologii. Dlatego w przyjmowanych dokumentach strategicznych wyraźnie widać, że ważną rolę w unijnym podejściu do AI odgrywa jej cyberbezpieczeństwo.

Komisja Europejska dąży do poprawy jakości usług publicznych, do szerokiego, międzysektorowego zaangażowania podmiotów w różnego rodzaju działania na rzecz rozwoju sztucznej inteligencji, w tym podmiotów prywatnych, które zachęca się do współinwestowania w ten rodzaj technologii. Zaproponowane przez HLEG AI wytyczne utworzyły horyzontalne ramy na potrzeby promowania i wdrażania godnej zaufania sztucznej

inteligencji w Unii Europejskiej, a także wspierania badań naukowych w tym obszarze.

Warto również dodać, że w przypadku uregulowań prawnych w zakresie AI, KE stara się ograniczyć wpływ ustawodawstwa na rozwój systemów sztucznej inteligencji. Zdaje sobie bowiem sprawę, że ramy prawne muszą obejmować elastyczne mechanizmy, dzięki którym można je dynamicznie dostosowywać wraz z rozwojem technologii i pojawiających się nowych wyzwań. Tylko w ten sposób Unia może podążać za wyzwaniami związanymi ze sztuczną inteligencją i starać się wychodzić im naprzeciw bez niepotrzebnej ingerencji w wolny rynek.

Bibliografia

[1] U. von der Leyen (16 lipca 2019). „Wytyczne polityczne dla Komisji na lata 2019–2024, ‘Unia, która mierzy wyżej’” [Online]. Dostęp: https://commission.europa.eu/system/files/2020-04/political-guidelines-next-commission_pl.pdf

[2] Komisja Europejska (10 maja 2017). „Komunikat do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów w sprawie przeglądu śródokresowego realizacji strategii jednolitego rynku cyfrowego, „Połączony jednolity rynek cyfrowydla wszystkich’” [Online]. Dostęp: https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=COM:2017:228:FIN

[3] Parlament Europejski (16 lutego 2017). „Rezolucja Parlamentu Europejskiego zawierająca zalecenia dla Komisji w sprawie przepisów prawa cywilnego dotyczących robotyki” [Online]. Dostęp: https://eur-lex.europa.eu/legal-content/PL/TXT/PDF/?uri=CELEX:52017IP0051

[4] Rada Europejska (19 października 2017). „Konkluzje po posiedzeniu Rady Europejskiej” [Online]. Dostęp: https://data.consilium.europa.eu/doc/document/ST-14-2017-INIT/pl/pdf

[5] Komisja Europejska (25 kwietnia 2018). „Komunikat do Parlamentu, Rady Europejskiego Komitetu Ekonomiczno-społecznego i Komitetu Regionów, ‘Sztuczna inteligencja dla Europy’” [Online]. Dostęp: https://eur-lex.europa.eu/legalcontent/PL/TXT/PDF/?uri=CELEX: 52018DC0237&from=EN

[6] Komisja Europejska (7 grudnia 2018). „Komunikat do PE, RE, Rady, EKES i Komitetu Regionów, 'Skoordynowany plan w sprawie sztucznej inteligencji'" [Online]. Dostęp:. https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=COM:2018:795:FIN

[7] Komisja Europejska (19 lutego 2020). „Biała księga w sprawie sztucznej inteligencji. Europejskie podejście do doskonałości i zaufania" [Online]. Dostęp: https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX%3A52020DC0065

[8] Wniosek – Rozporządzenie Parlamentu Europejskiego i Rady z 21 kwietnia 2021 ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji (Akt w sprawie sztucznej inteligencji) i zmieniające niektóre akty ustawodawcze Unii.

[9] Niezależna grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji (8 kwietnia 2019). „Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji" [Online]. Dostęp: https://op.europa.eu/pl/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[10] Niezależna grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji (26 czerwca 2019). „Policy and Investment Recommendations for Trustworthy AI" [Online]. Dostęp: https://www.europarl.europa.eu/italy/resource/static/files/import/intelligenza_artificiale_30_aprile/ai-hleg_policy-and-investment-recommendations.pdf

[11] Niezależna grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji (16 lipca 2020). „The Assessment List for Trustworthy Artificial Intelligence (ALTAI)" [Online]. Dostęp: https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1

AI Act – wyczekiwana regulacja systemów sztucznej inteligencji wysokiego ryzyka

Aleksandra Szczęsna

NASK-PIB | Centrum Cyberbezpieczeństwa i Infrastruktury | Dział Strategii i Rozwoju Bezpieczeństwa Cyberprzestrzeni

Wstęp

Od co najmniej kilku lat na forum unijnym sygnalizowana była konieczność przyjęcia przepisów dotyczących sztucznej inteligencji [1][2][3][4]. Bezpośrednią zapowiedzią przedłożenia takiej regulacji było dopiero jednak zobowiązanie polityczne przewodniczącej Komisji Europejskiej, Ursuli von der Leyen, do opracowania przepisów w sprawie skoordynowanego europejskiego podejścia do społecznych i etycznych konsekwencji AI [5].

21 kwietnia 2021 roku Komisja przedłożyła projekt AI Act (AIA), czyli rozporządzenia ustanawiającego zharmonizowane przepisy dotyczące sztucznej inteligencji. Wniosek w tej sprawie pozostaje aktualnie w toku prac legislacyjnych[1]. Akt miałby wejść w życie dwudziestego dnia po publikacji, a stosowany być co do zasady od dnia przypadającego 36 miesięcy po jego wejściu w życie[2]. W niniejszym opracowaniu ukazano stan prac

1 Na dzień 24 kwietnia 2023 r.

2 Wyjątkiem w tym zakresie mają być przepisy o organach notyfikujących i jednostkach notyfikowanych (tytuł III rozdział 4) oraz tytuł VI – zarządzanie, a także przepisy dotyczące kar (art. 71). Przepisy te ma stosować się od dnia przypadającego 12 miesięcy od daty wejścia w życie aktu.

nad projektem oraz jego treść, a także zasygnalizowano niektóre spośród wątpliwości dotyczących jego zapisów. Z uwagi na to, że w trakcie prac tekst ulega ciągłym modyfikacjom, punkt wyjścia dla niniejszego tekstu stanowi podejście ogólne Rady dotyczące wniosku w sprawie tego aktu, zatwierdzone 6 grudnia 2022 roku [6].

Motywy i stan prac

Jako cel przyjęcia AIA wskazano poprawę funkcjonowania rynku wewnętrznego poprzez ustanowienie jednolitych ram prawnych, w szczególności w zakresie rozwoju, wprowadzania do obrotu i wykorzystywania sztucznej inteligencji zgodnie z wartościami Unii. Podkreślono także konieczność realizacji takich celów jak wysoki poziom ochrony zdrowia, bezpieczeństwa i praw podstawowych. Akt ma adresować problemy z brakiem przejrzystości, wyjaśnialności i niedostatecznym dokumentowaniem działań wykonywanych przez AI [6].

Rada wprowadziła szereg poprawek do wniosku Komisji, m.in. dokonała modyfikacji w zakresie definicji systemu sztucznej inteligencji i rozszerzyła katalog zakazanych praktyk związanych z AI (w tym doprecyzowała możliwość stosowania tych systemów przez organy ścigania). Ponadto zaproponowała zmiany w zakresie klasyfikacji systemów AI jako wysokiego ryzyka oraz wymogów dotyczących takich systemów. Innym postulatem Rady było także dodanie nowych przepisów, które dotyczyłyby systemów sztucznej inteligencji ogólnego przeznaczenia (*general purpose AI* – GPAI), wykorzystywanych w systemach AI wysokiego ryzyka. Modyfikacje dotyczą także wprowadzenia większej przejrzystości wykorzystywania AI wysokiego ryzyka [7]. Przyjęcie ogólnego podejścia Rady, w tym przepisów dotyczących sztucznej inteligencji ogólnego przeznaczenia, wychodziło naprzeciw głosom wskazującym na ryzyka społeczne i ekonomiczne, związane z wykorzystywaniem takich systemów – np. dyskryminacją, szerzeniem nieprawdziwych lub manipulujących informacji czy zagrożeniami dla praw własności intelektualnej[3].

3 O czym informował Parlament Europejski w dokumencie [8].

W PE tymczasem projekt dyskutowany jest w komisjach, pod wspólnym przewodnictwem Komisji Rynku Wewnętrznego i Ochrony Konsumentów oraz Komisji Wolności Obywatelskich, Sprawiedliwości i Spraw Wewnętrznych[4]. Obserwatorium legislacyjne Parlamentu Europejskiego podaje, że planowo PE zajmie się aktem na posiedzeniu plenarnym 12 czerwca 2023 roku [8].

Treść regulacji

AIA będzie regulował m.in. obowiązki operatorów niektórych systemów typu GPAI, jak i wysokiego ryzyka oraz wymogi dla ich funkcjonowania na rynku. Poza tym w projekcie znalazły się również zakazy dotyczące określonych praktyk w zakresie sztucznej inteligencji. Osobno zostały uregulowane przepisy dotyczące przejrzystości systemów AI.

Zakres podmiotów, do jakich zastosowanie ma mieć rozporządzenie, został określony bardzo szeroko, tak aby regulacja objęła jak najszerszą grupę operatorów systemów AI[5]. Projekt zawiera jednak szereg wyłączeń – nie obejmie m.in. czysto osobistej działalności pozazawodowej osób fizycznych (z wyjątkiem obowiązków w zakresie przejrzystości).

Zaproponowana początkowo przez Komisję definicja systemu AI spotkała się z krytyką – zarzucano jej ryzyko nieobjęcia jej zakresem niektórych systemów, przez co nie musiałyby one spełniać wymogów rozporządzenia [9]. Pojęcie to doprecyzowano na etapie prac w Radzie – zgodnie z planowanym art. 3 pkt 1 AIA za system sztucznej inteligencji należało będzie uważać taki, który łącznie spełnia trzy warunki:

1. zaprojektowany został do działania w sposób częściowo autonomiczny;

2. w oparciu o dane i informacje dostarczone maszynowo lub przez człowieka, wnioskuje w jaki sposób osiągnąć zadany zestaw celów

4 Na dzień 24 kwietnia 2023 roku.

5 Zgodnie z art. 3 pkt 8 projektu rozporządzenia „operator" oznacza dostawcę, producenta produktu, użytkownika, upoważnionego przedstawiciela, importera lub dystrybutora.

– z wykorzystaniem technologii uczenia się maszyn lub metod opartych na logice i wiedzy oraz

3. generuje wyniki, takie jak treści (generatywne systemy sztucznej inteligencji), przewidywania, zalecenia lub decyzje, wpływające na środowiska, z którymi system ten wchodzi w interakcję.

Istotne dla zrozumienia przyjętych regulacji jest również rozróżnienie na systemy sztucznej inteligencji ogólnego przeznaczenia oraz wysokiego ryzyka. Systemy GPAI to takie, które w założeniu dostawcy wykonywać mają takie zadania, jak rozpoznawanie obrazów i mowy, wytwarzanie dźwięku lub treści wideo (generatywna AI), wykrywanie wzorców, udzielanie odpowiedzi na pytania, tłumaczenie lub inne (tzw. funkcje ogólnego przeznaczenia). AI Act przewiduje obowiązki głównie względem takich spośród powyższych systemów, które mogą być wykorzystane jako AI wysokiego ryzyka lub jako jej komponent. Z tego względu należy zaznaczyć, że jeśli w dalszej części niniejszego opracowania mowa będzie o systemach ogólnego przeznaczenia, oznacza to właśnie takie systemy. Warto jednak zaznaczyć, że aby zwolnić się z obowiązków przewidzianych dla tej kategorii systemów wystarczy, zgodnie z AIA, wyraźne wykluczenie (np. w instrukcji obsługi) zastosowań wysokiego ryzyka. Takiego wykluczenia będzie można dokonać na warunkach określonych w rozporządzeniu.

Drugą kategorią systemów AI, na które proponuje się nałożenie odrębnych wymogów, są systemy sztucznej inteligencji wysokiego ryzyka. Projektodawca wskazuje, że klasyfikacja systemu AI jako wysokiego ryzyka powinna ograniczać się do systemów mających wpływ na zdrowie, bezpieczeństwo i prawa podstawowe osób w Unii. Za systemy wysokiego ryzyka uznaje się:

1. systemy, które jako produkty objęte są unijnym prawodawstwem harmonizacyjnym określonym w załączniku II do aktu oraz systemy przeznaczone do wykorzystywania razem z nimi jako związane z bezpieczeństwem – jeśli muszą zostać poddane ocenie zgodności przeprowadzanej przez osobę trzecią w celu wprowadzenia ich do obrotu lub oddania do użytku zgodnie z tym prawodawstwem;

2. systemy, o których mowa w załączniku III do aktu, chyba że wynik ich działania jest wyłącznie pomocniczy w stosunku do działania lub decyzji, które należy podjąć i w związku z tym nie spowoduje istotnych zagrożeń dla zdrowia, bezpieczeństwa lub praw podstawowych.

Pierwsze z wymienionych to między innymi produkty określone w przepisach w sprawie maszyn [10], wyrobów medycznych [11] czy pojazdów [12]. Drugie z nich (załącznik III) to systemy AI wykorzystywane w niektórych celach w kategoriach takich jak biometria, infrastruktura krytyczna, zatrudnienie czy migracja[6].

Przyjęcie na stałe dziedzin, które mają decydować o kwalifikacji systemu jako wysokiego ryzyka, oceniane jest jako kontrowersyjne. Zarzuca mu się trudność w zmianie załączników (aktem delegowanym), a tym samym brak możliwości dynamicznej reakcji na zmieniające się okoliczności. Rozwiązanie tego problemu mogłoby opierać się na monitorowaniu ryzyka systemowego przez operatorów, podobnie jak ma to miejsce w przypadku DSA [13] [14].

Wymogi systemów AI

REGUŁY WSPÓLNE

W AIA zawarto wymogi dla systemów GPAI i wysokiego ryzyka, z których część jest wspólna[7]. Wśród takich obowiązków należy wymienić te określone rozdziale II tytułu III proponowanego aktu, tj.:

1. wdrożenie i utrzymanie systemu zarządzania ryzykiem w odniesieniu do systemów, w tym przyjęcie środków zarządzania ryzykiem;

2. spełnianie odpowiednich kryteriów jakości przez dane wykorzystywane do trenowania modeli;

6 Ponadto kształcenie i szkolenie zawodowe, podstawowe usługi prywatne oraz usługi i podstawowe świadczenia publiczne, ściganie przestępstw, zarządzanie migracją, azylem i kontrolą graniczną, sprawowanie wymiaru sprawiedliwości i procesy demokratyczne.

7 Warto wskazać, że wymogi dla systemów AI ogólnego przeznaczenia mają zostać dostosowane w świetle ich cech, wykonalności technicznej, specyfiki łańcucha wartości AI w świetle rozwoju rynku i rozwoju technologicznego, z uwzględnieniem stanu wiedzy technicznej. Dostosowanie to ma być dokonane przy pomocy aktów wykonawczych.

3. sporządzenie odpowiedniej dokumentacji technicznej;

4. automatyczne rejestrowanie zdarzeń w systemach sztucznej inteligencji;

5. zapewnienie przejrzystości w fazie projektowania i opracowywania, w tym zapewnienie instrukcji obsługi;

6. zapewnienie konieczności prowadzenia nadzoru przez człowieka;

7. zapewnienie odpowiedniego poziomu dokładności, solidności i cyberbezpieczeństwa przez cały cykl życia systemu.

Zarówno dostawcy systemów GPAI, jak i wysokiego ryzyka, będą musieli spełnić następujące warunki: określenie danych dostawcy systemu, poddanie systemu procedurze zgodności, rejestracja, działania naprawcze (jeśli system nie spełnia wymogów określonych w rozdziale II aktu tytułu III), odpowiednie oznakowanie, możliwość wykazania zgodności z wymogami. Będą także zobowiązani m.in. do monitorowania systemu po wprowadzeniu go do obrotu.

WYMOGI DLA AI WYSOKIEGO RYZYKA

Jak wskazano wyżej, część wymogów dla AI wysokiego ryzyka jest taka sama jak wobec systemów GPAI w nich wykorzystywanych. Poza tym, dostawcy systemów sztucznej inteligencji zaklasyfikowanej jako wysokiego ryzyka powinni: posiadać system zarządzania jakością i przechowywać rejestry zdarzeń generowane automatycznie przez te systemy. Będą oni musieli także informować odpowiednie organy krajowe oraz, w stosownych przypadkach, jednostki notyfikowane[8] o niezgodności z wymogami i o wszelkich podjętych działaniach naprawczych.

8 Zgodnie z art. 30 projektu aktu o sztucznej inteligencji, każde państwo członkowskie wyznacza lub ustanawia przynajmniej jeden organ notyfikujący odpowiedzialny za opracowanie i stosowanie procedur koniecznych do oceny, wyznaczania i notyfikowania jednostek oceniających zgodność oraz za ich monitorowanie. Jednostki notyfikowane certyfikują systemy sztucznej inteligencji.

Co ważne, jeśli dany system AI wysokiego ryzyka stwarza zagrożenie dla zdrowia i bezpieczeństwa lub praw podstawowych obywateli i jest ono znane jego dostawcy, będzie miał on obowiązek poinformować o tym odpowiednie organy oraz ściśle współpracować z nimi na zasadach określonych w akcie.

Do weryfikacji realizacji niektórych obowiązków zobligowani będą importerzy i dystrybutorzy systemów, którzy w przypadku wykrycia w nich nieprawidłowości nie będą mogli wprowadzić ich na rynek. W przypadku stwarzania przez system zagrożenia dla życia lub zdrowia, importerzy będą musieli poinformować o tym fakcie dostawcę oraz organy nadzoru rynku, a dystrybutorzy – dostawcę lub, w stosownych wypadkach, importera. Planuje się nałożenie obowiązku zgłoszenia AI stwarzającego takie ryzyko również na jego użytkowników. Obowiązkiem użytkownika systemu AI wysokiego ryzyka będzie także zapewnienie nadzoru jego działania przez człowieka i monitorowanie go w oparciu o instrukcję obsługi.

Komisja prowadzić będzie bazę danych, w której rejestracja obowiązkowa będzie w stosunku do niektórych z systemów AI wysokiego ryzyka[9] oraz ich dostawców lub upoważnionych przedstawicieli. Jednym z postulatów organizacji pozarządowych jest objęcie obowiązkiem rejestracji wszystkich użytkowników systemów wysokiego ryzyka, a także użytkowników wszystkich systemów AI w sektorze publicznym [15].

Systemy AI wysokiego ryzyka będą musiały być poddane monitorowaniu po wprowadzeniu ich do obrotu, na podstawie sporządzonego przez dostawcę planu. Ma to umożliwić dokonywanie oceny zgodności systemów z wymogami.

Ponadto, projekt rozporządzenia przewiduje odrębne obowiązki w zakresie zgłaszania incydentów w systemach AI wysokiego ryzyka. Dostawcy będą zobowiązani niezwłocznie zgłaszać incydenty poważne występujące w tych systemach organom nadzoru rynku w odpowiednim państwie członkowskim. Te z kolei poinformować mają o nim krajowe organy

9 Systemy określone w załączniku III do aktu, z wyjątkiem systemów określonych w pkt 1, 6 i 7 tego załącznika, w obszarach egzekwowania prawa, zarządzania migracją, azylem i kontrolą graniczną oraz z wyjątkiem systemów, o którym mowa w pkt 2 załącznika III.

publiczne lub organy ochrony praw podstawowych. Warto zwrócić uwagę, że obowiązki w zakresie zgłaszania incydentów nie zostały w żaden sposób odniesione do tych wynikających z innych regulacji dotyczących cyberbezpieczeństwa, jak NIS 2 [16] czy DORA [17].

Istotne jest również to, że w AIA nie zostały odrębnie uregulowane obowiązki w zależności od wielkości przedsiębiorstwa – rozporządzenie nie różnicuje w tym względzie operatorów systemów AI. Chociaż przewidziane zostały mechanizmy mające na celu ułatwienie stosowania aktu w sektorze MŚP, to jednak zakres obowiązków pozostaje ten sam.

Zakazane praktyki w zakresie sztucznej inteligencji

Planowane rozporządzenie zawierało będzie katalog zabronionych praktyk w zakresie sztucznej inteligencji. Mają one dotyczyć wszystkich dostawców systemów AI. Proponuje się zakazanie wprowadzania do obrotu, oddawania do użytku lub wykorzystywania systemu sztucznej inteligencji, który mógłby manipulować zachowaniem danej osoby, powodując u niej szkodę fizyczną lub psychiczną, a który:

A. stosuje techniki podprogowe będące poza świadomością danej osoby,

B. wykorzystuje słabości określonej grupy osób ze względu na ich wiek, niepełnosprawność lub szczególną sytuację społeczną lub ekonomiczną.

Ponadto zabroniony będzie system AI wykorzystywany na potrzeby oceny lub klasyfikacji osób fizycznych, prowadzonej na podstawie ich zachowania społecznego lub cech osobistych, jeśli prowadzi to do ich krzywdzącego lub niekorzystnego w skutkach traktowania. Obejmuje to sytuacje, kiedy ma to miejsce w takich kontekstach społecznych, które nie są związane z tymi, w których pierwotnie wygenerowano lub zgromadzono dane.

Dodatkowo rozporządzenie wprowadzi zakaz wykorzystywania w przestrzeni publicznej przez organy ścigania lub w ich imieniu systemów zdalnej identyfikacji biometrycznej w czasie rzeczywistym, chyba że jest to

niezbędne do jednego z celów związanych z konkretnym przestępstwem i za zezwoleniem. Zezwolenie będzie można uzyskać co do zasady przed wykorzystaniem systemu, jednak w niektórych przypadkach dozwolone będzie wystąpienie o nie już w trakcie stosowania systemu.

Warto zauważyć, że zakaz ten dotyczy wykorzystania zdalnej identyfikacji biometrycznej w celach egzekwowania prawa. Organizacje społeczne domagają się zaś wprowadzenia całkowitego zakazu dla wszystkich – czy to organów ścigania czy np. podmiotów sektora prywatnego [18].

Obowiązki w zakresie przejrzystości

Jednymi z najistotniejszych rozwiązań proponowanych w AI Act są zapisy dotyczące przejrzystości systemów sztucznej inteligencji, co było postulowane wcześniej przez powołaną przez Komisję grupę ekspertów wysokiego szczebla [19].

Implementacja zasady przejrzystości w AIA polegać będzie na tym, że osoby fizyczne, korzystające z systemów sztucznej inteligencji[10] (niezależnie od ich przeznaczenia), będą informowane o prowadzeniu interakcji z tego rodzaju systemem. Informacje na temat przejrzystości powinny być przekazane odbiorcom w jasny i wyraźny sposób, najpóźniej w momencie pierwszej interakcji lub kontaktu.

Wyjątkiem od zasady informowania mają być sytuacje, w których fakt korzystania z takiego systemu, biorąc pod uwagę okoliczności i kontekst, będzie dla tej osoby oczywisty, a dany odbiorca – uważny i ostrożny. Powyższy wymóg nie będzie dotyczył także systemów związanych z działalnością organów ścigania i prowadzenia postępowań przygotowawczych, jeśli zapewnione zostaną odpowiednie gwarancje zabezpieczające prawa i wolności osób trzecich (chyba że systemy te będą udostępniane na potrzeby składania zawiadomień o popełnieniu przestępstwa).

10 W przypadku systemów kategoryzacji biometrycznej i systemów rozpoznawania emocji informowane mają być także osoby, wobec których stosowany jest system.

Użytkownicy systemów generatywnej sztucznej inteligencji (w tym systemów generujących treści typu *deepfake*) również będą co do zasady zobowiązani do ujawnienia faktu skorzystania z takiego systemu.

Co więcej, dostawcy systemów AI wysokiego ryzyka (w tym również systemy GPAI, które mogą być wykorzystane jako wysokiego ryzyka) będą musieli spełnić szereg dodatkowych wymogów. Jednym z nich będzie dołączenie instrukcji obsługi, zawierającej m.in. poziom dokładności systemu czy ryzyka dla bezpieczeństwa lub praw podstawowych.

Podkreślenia wymaga, co zostało już zasygnalizowane wcześniej, że obowiązkom związanym z instrukcją obsługi podlegać będą tylko takie systemy ogólnego przeznaczenia, które mogą być wykorzystane jako systemy wysokiego ryzyka lub są ich komponentami. W celu zwolnienia się z obowiązków wynikających z AI Act, dostawca będzie musiał zastrzec, że system nie powinien być wykorzystywany w takim celu. Tym samym dostawcy systemów, które pełnią funkcje ogólnego przeznaczenia, ale wobec których zastrzeżono powyższe, nie będą musieli informować o ryzykach związanych z ich wykorzystaniem.

Systemy AI stwarzające ryzyko

Jeżeli system sztucznej inteligencji stwarza ryzyko związane z zagrożeniem dla zdrowia i bezpieczeństwa lub praw podstawowych obywateli, organ nadzoru rynku dokona jego oceny pod kątem zgodności z wymogami AIA. Jeśli ryzyko dotyczyć będzie praw podstawowych – poinformuje o tym odpowiednie krajowe organy lub organy ochrony praw podstawowych. Jeżeli podczas oceny stwierdzone zostaną uchybienia – organ nadzoru rynku zaleci podjęcie odpowiednich działań naprawczych. W razie potrzeby może on nakazać także wycofanie systemu z rynku lub od użytkowników.

Jeśli operator systemu nie dostosuje się do obowiązku wdrożenia środków naprawczych, może narazić się na zakaz lub ograniczenie jego udostępniania, nałożenie obowiązku wycofania go z rynku lub od użytkowników. Takie środki będą mogły być nałożone również wtedy, gdy system będzie stwarzał ryzyko, mimo zgodności z samym rozporządzeniem.

Pozostałe regulacje

W projekcie rozporządzenia zaproponowano wymagania wobec organów notyfikujących i jednostek notyfikowanych, a także przepisy dotyczące norm, oceny zgodności, certyfikatów i rejestracji. Ponadto, AIA wprowadzi także środki wspierające innowacyjność w zakresie sztucznej inteligencji – w tym przede wszystkim tzw. piaskownice regulacyjne[11], wymogi dla testowania systemów AI poza tymi środowiskami oraz środki wsparcia dla operatorów, w szczególności małych i średnich przedsiębiorstw (w tym typu start-up).

Co więcej, po wejściu w życie rozporządzenia ustanowiona zostanie Europejska Rada ds. Sztucznej Inteligencji, jako organ doradczy dla Komisji i państw członkowskich. Jej zadaniem będzie wsparcie tych podmiotów w celu ułatwienia spójnego i skutecznego stosowania AIA. Ponadto na państwa członkowskie nałożony zostanie obowiązek wyznaczenia właściwych organów krajowych (co najmniej jednego organu notyfikującego oraz co najmniej jednego organu nadzoru rynku). Szczególne uprawnienia przyznane zostaną organom ochrony praw podstawowych. Podmioty te będą mogły żądać dostępu do wszelkiej dokumentacji sporządzonej lub prowadzonej na podstawie rozporządzenia przez podmioty określone w załączniku III. W przypadku gdyby dokumentacja była niewystarczająca – dany organ będzie mógł wystąpić do organu nadzoru rynku z uzasadnionym wnioskiem o zorganizowanie testów systemu.

Kary

W projekcie wskazane zostały różne progi kar, w zależności od rodzaju naruszonych przepisów i wielkości przedsiębiorstwa: między 10 a 30 milionów euro lub – jeżeli naruszenia dopuści się przedsiębiorstwo – od 4 do 6 % (1-3 % w przypadku MŚP) całkowitego światowego obrotu (z poprzedniego roku obrotowego). Zastosowanie będzie mieć kwota wyższa.

11 Konkretne ramy ustanowione przez właściwy organ krajowy, umożliwiające dostawcom lub potencjalnym dostawcom systemów AI możliwość opracowywania, trenowania, walidowania i testowania – w stosownych przypadkach w warunkach rzeczywistych – innowacyjnych systemów sztucznej inteligencji, w oparciu o szczegółowy plan, w ograniczonym czasie i pod nadzorem regulacyjnym [art. 3 pkt 52].

Podsumowanie

Jak w wielu sytuacjach, tak i w przypadku AIA Unia Europejska musi wyważyć interesy przedsiębiorców, konieczność wdrożenia zasad *security and safety* oraz zagwarantowanie praw podstawowych. Jakkolwiek jeszcze wiele kwestii związanych z bezpieczeństwem, przejrzystością czy rozliczalnością AI wymaga doprecyzowania, to jednak konieczność przyjęcia regulacji dotyczącej sztucznej inteligencji jest niepodważalna. Nie ulega bowiem wątpliwości, że zagadnienia takie jak niedozwolone praktyki w zakresie AI, obowiązki w zakresie przejrzystości, czy regulacja sztucznej inteligencji wysokiego ryzyka wymagają szybkiego zaadresowania. Należy także zwrócić uwagę na to, że zmiany w zakresie rynku systemów AI następują tak dynamicznie, że po uchwaleniu aktu i upływie okresu *vacatio legis* mogą pojawić się kolejne kwestie związane z tym tematem, które wymagały będą regulacji.

Prace nad rozporządzeniem wchodzą w kluczowy etap, a tym samym pozostaje mieć nadzieję, że prawodawca unijny zdoła zrównoważyć sprzeczne interesy i zapewnić poprawę bezpieczeństwa systemów AI już w najbliższym czasie.

Bibliografia

[1] Komisja Europejska (25 kwietnia 2018). „Komunikat do Parlamentu, Rady Europejskiego Komitetu Ekonomiczno-społecznego i Komitetu Regionów, ‘Sztuczna inteligencja dla Europy’” [Online]. Dostęp: https://eur-lex.europa.eu/legalcontent/PL/TXT/PDF/?uri=CELEX:52018DC0237&from=EN.

[2] Komisja Europejska (19 lutego 2020). „Biała księga w sprawie sztucznej inteligencji. Europejskie podejście do doskonałości i zaufania„ [Online]. Dostęp: https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX%3A52020DC0065.

[3] Parlament Europejski (20 października 2020). „Rezolucja zawierająca zalecenia dla Komisji w sprawie ram aspektów etycznych sztucznej inteligencji, robotyki i powiązanych z nimi technologii” [Online]. Dostęp: https://eur-lex.europa.eu/legalcontent/PL/TXT/?uri=CELEX:52020IP0275.

[4] Rada Europejska (1-2 października 2020). „Konkluzje po nadzwyczajnym posiedzeniu", EUCO 13/20 [Online]. Dostęp: https://data.consilium.europa.eu/doc/document/ST-14-2017-INIT/pl/pdf.

[5] U. von der Leyen (16 lipca 2019). „Wytyczne polityczne dla Komisji na lata 2019–2024, 'Unia, która mierzy wyżej'" [Online]. Dostęp: https://commission.europa.eu/system/files/2020-04/political-guidelines-next-commission_pl.pdf

[6] Wniosek – Rozporządzenie Parlamentu Europejskiego i Rady ustanawiającego zharmonizowane przepisy dotyczące sztucznej inteligencji (akt w sprawie sztucznej inteligencji) i zmieniającego niektóre akty ustawodawcze Unii – Podejście ogólne (6 grudnia 2022 r.), 2021/0106(COD).

[7] Rada Unii Europejskiej (6 grudnia 2022). *Akt o sztucznej inteligencji: Rada apeluje o bezpieczną sztuczną inteligencję zgodną z prawami podstawowymi*, Komunikat prasowy [Online]. Dostęp: https://www.consilium.europa.eu/pl/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/

[8] European Parliament Legislative Observatory [Online]. Dostęp: https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2021/0106(COD)&l=en.

[9] J. J. Bryson (2 marca 2022). Europe Is in Danger of Using the Wrong Definition of AI [Online]. Dostęp: https://www.wired.com/story/artificial-intelligence-regulation-european-union/

[10] Dyrektywa 2006/42/WE Parlamentu Europejskiego i Rady z dnia 17 maja 2006 r. w sprawie maszyn, zmieniająca dyrektywę 95/16/WE, Dz.U. L 157 z 9.6.2006 [uchylona rozporządzeniem w sprawie maszyn].

[11] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2017/745 z dnia 5 kwietnia 2017 r. w sprawie wyrobów medycznych, zmiany dyrektywy 2001/83/WE, rozporządzenia (WE) nr 178/2002 i rozporządzenia (WE) nr 1223/2009 oraz uchylenia dyrektyw Rady 90/385/EWG i 93/42/EWG, Dz.U. L 117 z 5.5.2017.

[12] Rozporządzenie Parlamentu Europejskiego i Rady (UE) nr 168/2013 z dnia 15 stycznia 2013 r. w sprawie homologacji i nadzoru rynku pojazdów dwu – lub trzykołowych oraz czterokołowców, Dz.U. L 60 z 2.3.2013.

[13] N. Helberger, N. Diakopoulos (Luty 2023). „ChatGPT and the AI Act". *Internet Policy Review* [Online] vol. *12 nr* 1. Dostęp: https://doi.org/10.14763/2023.1.1682.

[14] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (akt o usługach cyfrowych).

[15] European Digital Rights i in. (19 kwietnia 2023). *European Parliament: Make sure the AI act protects peoples' rights!*, list otwarty organizacji społecznych https://panoptykon.org/sites/default/files/statement-european-parliament-make-sure-the-a-i-act-protects-peoples-rights.pdf

[16] Dyrektywa Parlamentu Europejskiego i Rady (UE) 2022/2555 z dnia 14 grudnia 2022 r. w sprawie środków na rzecz wysokiego wspólnego poziomu cyberbezpieczeństwa na terytorium Unii, zmieniająca rozporządzenie (UE) nr 910/2014 i dyrektywę (UE) 2018/1972 oraz uchylająca dyrektywę (UE) 2016/1148 (dyrektywa NIS 2).

[17] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2554 z dnia 14 grudnia 2022 r. w sprawie operacyjnej odporności cyfrowej sektora finansowego i zmieniające rozporządzenia (WE) nr 1060/2009, (UE) nr 648/2012, (UE) nr 600/2014, (UE) nr 909/2014 oraz (UE) 2016/1011.

[18] European Digital Rights i in. (Listopad 2021). „Joint civil society recommendations for an EU Artficial Intelligence Act for Fundamental Rights. Biometrics Part 1: Article 3(36) and Article 5(1)(d)" [Online]. Dostęp: https://edri.org/wp-content/uploads/2022/05/Prohibit-RBI-in-publicly-accessible-spaces-Civil-Society-Amendments-AI-Act-FINAL.pdf

[19] Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji (8 kwietnia 2019). „Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji", [Online]. Dostęp: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11-06/Ethics-guidelines-AI_PL.pdf

Wyzwania dla cyberbezpieczeństwa sztucznej inteligencji w kontekście AI Act –

wywiad z dr Gabrielą Bar

Emilia Zalewska-Czajczyńska

NASK-PIB | Centrum Cyberbezpieczeństwa i Infrastruktury | Dział Strategii i Rozwoju Bezpieczeństwa Cyberprzestrzeni

Z dużym prawdopodobieństwem w ciągu najbliższych miesięcy przyjęta zostanie finalna wersja unijnego aktu o sztucznej inteligencji. Jego rolą jest kompleksowe uregulowanie rynku i zastosowań systemów AI. Na temat projektu rozporządzenia oraz o tym, jakie wyzwania prawne i etyczne niesie ze sobą rozwój sztucznej inteligencji, rozmawiamy z dr Gabrielą Bar, partnerką zarządzającą Szostek_Bar i Partnerzy Kancelarii Prawnej, członkinią Women in AI.

Emilia Zalewska-Czajczyńska: W ostatnim czasie sporo działo się w temacie sztucznej inteligencji, dlatego na początku chciałabym zadać pytanie nawiązujące do tych wydarzeń. W dużym skrócie – najpierw, 22 marca opublikowany został list otwarty Future of Life Institute, wzywający do zaprzestania na przynajmniej sześć miesięcy prac nad tworzeniem modeli sztucznej inteligencji potężniejszych niż obecny Chat GPT-4 [1]. List ten podpisało wielu przedstawicieli świata nauki i biznesu, w tym osoby o znanych nazwiskach, takie jak Elon Musk, Steve Woźniak i Yuval Harari, co nadało dokumentowi dodatkowy rozgłos. Następnie, 24 marca pojawiła się informacja, że włoski organ ochrony danych osobowych zablokował możliwość przetwarzania danych przez Chat GPT na terenie Włoch ze względu na, między innymi, niejasną podstawę przetwarzania danych osobowych użytkowników i brak weryfikacji wieku [2]. Ponieważ obydwa te zdarzenia wiążą się, w pewnym sensie, z obawami przed nowymi rozwiązaniami z zakresu sztucznej inteligencji, chciałabym zacząć tę rozmowę od odrobinę filozoficznego pytania – czy przestajemy nadążać za rozwojem tej technologii? Czy wyprzedza nas ona na tyle, że nie jesteśmy w stanie zapewnić możliwości bezpiecznego korzystania z jej rozwiązań?

Gabriela Bar: To bardzo dobre i faktycznie trochę egzystencjalne pytanie. Jednak nie pojawia się ono prawie wcale w dyskursie prawniczym. Być może nigdy nie stworzymy ogólnej sztucznej inteligencji – Artificial General Intelligence (AGI), której poziom będzie odpowiadał zdolnościom intelektualnym człowieka (albo je przekroczy), i która jest jak „święty Graal" badań nad AI. Ale powinniśmy być przygotowani, że tak jednak się stanie. Przecież większość badaczy sztucznej inteligencji uważa, że prędzej czy później to nastąpi, a średni czas oczekiwania na pojawienie się AGI to lata 90' tego wieku. Tak więc przyszłość odległa, ale nie aż tak bardzo. W literaturze zachodniej – również prawniczej – często pojawia się temat tzw. osobliwości – singularity, czyli potencjalnej sytuacji, w której sztuczna inteligencja stanie się na tyle inteligentna, że zacznie sama tworzyć nowe sztuczne inteligencje. Te będą coraz inteligentniejsze i bardziej wszechstronne, a zatem w bardzo szybkim tempie przekroczą możliwości intelektualne człowieka, a my zupełnie stracimy nad nimi kontrolę. Kiedy kilka lat temu zaczynałam o tym mówić i pojawiały się też dyskusje na temat potencjalnej możliwości nadania AI jakiejś formy osobowości prawnej, były to tematy zupełnie niszowe. Wręcz spotykałam się z podejściem, że w ogóle szkoda o tym mówić, przynajmniej między prawnikami, ponieważ to jest zbyt daleka i nierealna przyszłość.

A ja wychodzę z założenia, że zawsze warto o takich rzeczach rozmawiać, bo wiele razy nasi przodkowie przekonali się, że coś, co wydawało się niemożliwe, jednak się wydarzyło, bo nastąpił nagły przełom w rozwoju technologii lub nauki.

W kwestii listu Future of Life Institute – co prawda nie mam zaufania do niektórych sygnatariuszy listu, na przykład Elona Muska, który jest kompletnie niespójny w swoich wypowiedziach na temat sztucznej inteligencji, to jednak podpisali go też profesorowie Max Tegmark i Stuart Russel, którzy są świetnymi specjalistami w zakresie sztucznej inteligencji i którzy już wcześniej zwracali uwagę na problem poruszony w liście. Ja też się pod tym listem podpisałam, bo uznałam, że dotyczy on kwestii na tyle ważnych – także dla mnie, jako osoby interesującej się tym tematem – że warto to zrobić. Sądzę, że sztuczna inteligencja wymyka się nam spod kontroli. Ogromne modele językowe, czyli m. in. Chat GPT, działają na zasadzie sztucznych, głębokich sieci neuronowych. To są czarne skrzynki. Nawet przedstawiciel Open AI, jeden z twórców Chatu GPT, w wypowiedzi dla prasy [3] wskazał, że ich samych przeraża tempo rozwoju stworzonej przez nich technologii, ponieważ nie do końca wiedzą, co się dzieje „w środku". Z tego względu uważam za słuszne skupienie się na uzyskaniu przejrzystości i wyjaśnialności takich modeli – co pozwoli na ich ulepszanie. Ale przede wszystkim na budowaniu zaufania do nich i usuwaniu zagrożeń, a nie na zwiększaniu ich wydajności, żeby były coraz lepsze, ale tak naprawdę niekontrolowalne.

EZC: Trudności w nadążaniu za technologią mają także regulatorzy. W wyścigu z postępem w nauce prawo właściwie zawsze zostaje w tyle. Podobnie dzieje się w przypadku sztucznej inteligencji – choć jej rozwój jest już obecnie bardzo zaawansowany, to pierwsza, kompleksowa regulacja tej kwestii na poziomie Unii Europejskiej – akt w sprawie sztucznej inteligencji – jest od 2021 r. na etapie projektu. W dodatku zmiany technologiczne, które pojawiły się w trakcie trwania prac nad tym aktem, takie jak gwałtowny rozkwit generatywnej sztucznej inteligencji, nie ułatwiają Unii zadania. Jaka jest szansa, żeby akt w sprawie sztucznej inteligencji, gdy już zostanie uchwalony, odwzorowywał faktyczne potrzeby regulacyjne w tym obszarze i nie tracił szybko na aktualności?

GB: Wydaje mi się, że tak zwane prawo nowych technologii nigdy nie nadążało za tym, co działo się w tym obszarze. Pamiętam, jak miałam

dwadzieścia kilka lat i w CBKE[1] dyskutowaliśmy nad takimi problemami, jak składanie oświadczenia woli za pośrednictwem poczty elektronicznej. To był zupełnie inny poziom złożoności, ale wtedy właśnie to stanowiło nową technologię w pracy prawnika, czy ogólnie w biznesie. Wówczas też mówiło się, że prawo nie nadąża za technologią oraz że brakuje regulacji. Ale wtedy też faktycznie takich regulacji, dotyczących chociażby składania oświadczeń woli w postaci elektronicznej, nie było.

To więc nic nowego, że prawo nie nadąża za technologią, a im szybciej ta się rozwija – a obecnie dzieje się to coraz szybciej – tym bardziej nie nadąża. I o ile ja osobiście uważam, że inicjatywa uregulowania sztucznej inteligencji, co do zasady, jest dobra, to mam wrażenie, że Unia zabrała się za to nie do końca umiejąc w ogóle określić, co jest tą sztuczną inteligencją. Bardzo przekonuje mnie pogląd, który był wyrażany w różnych raportach i opracowaniach związanych z przygotowywaniem projektu aktu o sztucznej inteligencji, że dobrze byłoby zacząć od uregulowania, czym w ogóle jest oprogramowanie. Następnie wprowadzić zasadę, według której oprogramowanie, w zależności od tego np. do czego będzie używane, powinno podlegać pewnej ocenie ryzyka i wymogom w zakresie przejrzystości i wyjaśnialności oraz różnym zaleceniom w zakresie danych. Tutaj zaczęto od razu od uregulowania sztucznej inteligencji, kiedy tak naprawdę unijni prawodawcy nadal dyskutują o tym, czego w ogóle akt ma dotyczyć.

Ta regulacja jest procedowana już od kwietnia 2021 roku, czyli od kiedy Komisja Europejska przedstawiła jej projekt, ale tak naprawdę dopiero jesienią ubiegłego roku zdano sobie sprawę, że jest coś takiego, jak generatywna sztuczna inteligencja i duże modele językowe, do których zalicza się Chat GPT. Uświadomiono sobie także, że te technologie mają ogromną moc: z jednej strony obliczeniową, ale z drugiej – także wywierania wpływu i manipulacji. Dopiero wtedy zorientowano się, że te kwestie również należałoby uregulować, co wywróciło dotychczasowe prace nad aktem o sztucznej inteligencji do góry nogami.

Obecnie sytuacja wygląda więc następująco: mamy projekt, wniosek i stanowisko Komisji Europejskiej, a także wspólne stanowisko Rady, opublikowane w grudniu ubiegłego roku. Parlament nie uzgodnił jeszcze swojego

1 Centrum Badań Problemów Prawnych i Ekonomicznych Komunikacji Elektronicznej Uniwersytetu Wrocławskiego (przyp. red.)

stanowiska. Gdy już je przyjmie, dojdzie do trilogu między Komisją, Radą i Parlamentem. Jeżeli akt będzie przyjęty, zostanie opublikowany w Dzienniku Ustaw. Ma wejść w życie po upływie 20 dni od publikacji, a zacząć obowiązywać – dwa lata później. Nawet gdyby optymistycznie na to spojrzeć, to miną jeszcze mniej więcej trzy lata, zanim akt o sztucznej inteligencji stanie się obowiązującym prawem. I obawiam się, że przy aktualnym tempie rozwoju technologii, będzie to prawo mocno nieaktualne. Mogą także pojawić się takie nowe wyzwania, które nie tyle sprawią, że akt o sztucznej inteligencji stanie się nieaktualny – jest bowiem napisany dość neutralnie technologicznie – co po prostu wykroczą poza jego obszar regulacji.

EZC: Chciałabym jeszcze dopytać o aspekt, o którym przed chwilą wspomniałaś, a mianowicie o definicję sztucznej inteligencji. Jest to jeden z elementów aktu, który zmieniał się kilkukrotnie w trakcie prac. Czy mogłabyś opowiedzieć o tym, jak wyglądały próby stworzenia tej definicji?

GB: Poprawek w akcie do definicji sztucznej inteligencji było faktycznie kilka. Natomiast ja wyróżniam trzy zasadnicze momenty w tym procesie. Pierwszy z nich to ten, kiedy Komisja zaproponowała swoją definicję, która odwoływała się do słowa „oprogramowanie". Moim zdaniem była ona bardzo dobra, ponieważ sztuczna inteligencja w dalszym ciągu to właśnie oprogramowanie, program komputerowy. W celu zachowania neutralności technologicznej, definicja uzupełniona została załącznikiem I [4], zawierającym istniejące techniki i podejścia. Wywołało to ogólną krytykę – wskazano, że w załączniku, oprócz metod uczenia maszynowego, znalazły się także np. metody statystyczne, czyli w zasadzie oparte na matematyce, co oznaczałoby, że pod pojęciem sztucznej inteligencji znalazłyby się wszystkie programy komputerowe, które umożliwiają działania na liczbach.

Drugim ważnym momentem było pojawienie się pomysłu zawężenia zakresu załącznika I, tak aby nie zawierał on wyżej wspomnianych metod statystycznych, albo w ogóle rezygnacji z niego. Zamiast tego, do definicji sztucznej inteligencji wpisano, że jest to oprogramowanie, które posiada pewien stopień autonomiczności i działa w oparciu o m. in. metody uczenia maszynowego. I to wydawało się słusznym kierunkiem. Jednak później zdecydowano, żeby do definicji wpisać, że system sztucznej inteligencji

to system, który zdolny jest do wpływania na środowisko, w którym działa, i generowania wyników (*output*). Był to błąd logiczny, ale ta definicja bardzo długo funkcjonowała. Na szczęście nie została ostatecznie uznana za właściwą i na początku marca Parlament Europejski zgodził się na wprowadzenie nowej, która jest zbieżna z definicją sztucznej inteligencji przyjętą przez OECD [4]. Nowa wersja również odwołuje do trochę niejasnego pojęcia, wskazuje bowiem, że system sztucznej inteligencji to *machine based system*. Trudno to nawet przetłumaczyć na polski, bo „system oparty na maszynie" brzmi niezbyt dobrze. Zdecydowano się więc na pozostanie przy słowie „system" zamiast „oprogramowanie". Podobno Parlament chce tę wersję definicji pozostawić, ale przyszłość pokaże, czy faktycznie to właśnie ona zostanie przyjęta. Plusem jest to, że jest to przynajmniej spójne z opracowaniem OECD [5].

To wszystko pokazuje, jak trudne jest uchwycenie w ogóle zakresu regulacji sztucznej inteligencji oraz jak bardzo ta technologia jest skomplikowana, skoro nie potrafimy nawet precyzyjnie zdefiniować, czym ona jest. A to powinno być w ogóle pierwszym krokiem do jej uregulowania.

Warto jednak dodać, że definicja legalna, czyli ta, którą mamy w akcie prawnym, nie musi być perfekcyjna. W tym sensie, że uwzględnia wszystkie aspekty danej technologii, czyli że zgodzą się z nią również matematycy, fizycy kwantowi, filozofowie itd. Dla każdego dana technologia jest trochę czymś innym ze względu na sposób, w jaki na nią patrzy. Jednocześnie jednak, definicja legalna powinna być na tyle jasna, precyzyjna i racjonalna, żeby wiadomo było, co akt reguluje. To jest główna moja obawa o definicję sztucznej inteligencji – czy rzeczywiście zrealizuje ona swoje zadanie.

EZC: W projekcie aktu o sztucznej inteligencji Komisja zdecydowała się na zaproponowanie szczegółowego uregulowania tylko jednego rodzaju dozwolonych systemów AI – systemów wysokiego ryzyka. W stosunku do innych, obowiązki nakładane przez akt są o wiele mniejsze. Dlaczego takie podejście przyjęto i jakie są tego konsekwencje?

GB: Akt o sztucznej inteligencji jest mimo wszystko wyrazem kompromisu. Z jednej strony znajduje się potrzeba zapewnienia bezpieczeństwa działania tej technologii – tego, że będzie ona godna zaufania, a więc biznes i konsumenci będą chcieli jej używać. Z drugiej strony istnieje z kolei

potrzeba wspierania innowacji i unikania ich zahamowania. Myślę więc, że przyjęcie podejścia, zgodnie z którym tylko systemy wysokiego ryzyka będą podlegały szczegółowym obowiązkom, jest właśnie takim kompromisem, który ma na celu uniknięcie ograniczenia rynku i sytuacji, w której każda sztuczna inteligencja musiałaby spełniać wszystkie wymogi z aktu. W przyjętym obecnie podejściu musi je zaś spełniać system, który rzeczywiście ma wpływ na prawa podstawowe i rodzi ryzyka związane z życiem, zdrowiem, w tym zdrowiem psychicznym, majątkiem oraz dostępem do pracy, edukacji lub wymiaru sprawiedliwości – czyli kwestiami związanymi z prawami człowieka, w tym z prawami podstawowymi.

Osobiście myślę, że to jest dobre podejście. Rzeczywiście nie można do jednego worka wkładać systemu rekomendacyjnego platform streamingowych, który na podstawie naszych preferencji sugeruje nam kolejny film do obejrzenia czy kolejny utwór do posłuchania, z systemem, który będzie proponował sędziemu rozstrzygnięcie w sprawie sądowej. Oba te zastosowania są ważne i potrzebne, ale zdecydowanie większy wpływ na życie człowieka będzie miało to drugie. Dlatego podejście oparte na szczegółowej regulacji jedynie systemów wysokiego ryzyka jest zasadne – już w tej chwili otacza nas mnóstwo różnych, drobnych algorytmów, które ułatwiają nam życie i jednocześnie nie wymagają aż tak daleko idącego uregulowania.

Natomiast problem powstał, albo raczej zaczął być komunikowany w znacznie większej skali, kiedy pojawił się Chat GPT oraz inne, generatywne systemy sztucznej inteligencji. Z jednej strony są to narzędzia, które mogą pełnić rolę po prostu czatu, na którym zadając pytania, dostanie się mniej lub bardziej mądre odpowiedzi, i tym samym prowadzona jest jakaś luźna konwersacja. Ale z drugiej strony, generatywna AI może także stać się potężnym narzędziem do manipulacji i produkcji fałszywych informacji. A także narzędziem do programowania, które jest w stanie pisać kody kolejnych algorytmów, które to z kolei mogą mieć wpływ na dalsze sfery naszego życia.

Na tym właśnie polega problem ze sztuczną inteligencją ogólnego przeznaczenia, na co zwrócili uwagę m. in. właśnie Max Tegmark i jego współpracownicy w ramach Future of Life Institute już jakiś czas temu. Dlatego też mocno udzielali się oni w dyskusjach toczących się w Parlamencie Europejskim, opowiadając się za tym, żeby w akcie o sztucznej inteligencji zająć się również tym rodzajem sztucznej inteligencji. Ponieważ

system AI ogólnego przeznaczenia, jak już wspomniałam, może zostać użyty zarówno do prowadzenia jakiejś sympatycznej rozmowy lub do oznaczenia znajomych na Facebooku, jak i do manipulacji, albo, w przypadku systemów do rozpoznawania twarzy, do identyfikacji biometrycznej. Więc znowu: rozrzut pomiędzy skutkami, jakie każdy z tych systemów może wywołać, jest ogromny.

EZC: Ze względu m.in. na wątpliwości, o których przed chwilą wspomniałaś, do pierwotnego tekstu aktu zaproponowano poprawkę dotyczącą sztucznej inteligencji ogólnego przeznaczenia. Czy udało się w niej uchwycić specyfikę systemów tego rodzaju? Spotkałam się z zarzutem, że modele generatywnej sztucznej inteligencji nadal są w stanie wymknąć się przepisom tej regulacji.

GB: To, co zaproponował Parlament Europejski, nie do końca mi się podoba. Obecnie tylko część obowiązków związanych z systemami sztucznej inteligencji wysokiego ryzyka została wprost nałożona na dostawców systemów sztucznej inteligencji ogólnego przeznaczenia, a pozostałe będą dookreślone w późniejszym czasie, na podstawie obserwacji rynku przez Komisję Europejską we współpracy z organami nadzorczymi. Komisja, Biuro ds. AI i Europejskie organizacje normalizacyjne będą miały obowiązek uwzględniania zasad etycznych opracowanych przez HLEG dla godnej zaufania sztucznej inteligencji [6]. W przypadku modeli podstawowych (*foundation models*), operatorzy systemów AI będą musieli ich przestrzegać poprzez realizację obowiązków nałożonych przez AI dla dostawców, dystrybutorów, importerów, podmiotów wdrażających lub innych stron trzecich.

EZC: Dodatkowo sztuczna inteligencja ogólnego przeznaczenia jako taka nie została dopisana do listy systemów wysokiego ryzyka, która znajduje się w załączniku III do aktu, a umieszczona w tekście rozporządzenia jako osobna kategoria.

GB: Dokładnie. Chociaż w ostatnim czasie, na posiedzeniu Komisji 14 marca, zaproponowano dodanie do listy systemów sztucznej inteligencji wysokiego ryzyka takich, które generują złożone treści tekstowe lub audio-wideo. Czyli generatywna sztuczna inteligencja znalazłaby się na tej liście, ale inne systemy AI ogólnego przeznaczenia już nie. Zostało to więc

rozdzielone. Nadal zatem sztuczna inteligencja ogólnego przeznaczenia jest objęta, przynajmniej na tym etapie, tylko niektórymi obowiązkami dotyczącymi systemów wysokiego ryzyka, z wyjątkiem jednak sztucznej inteligencji, która generuje treści tekstowe lub audio-wideo.

To wszystko jest na etapie propozycji, a właściwie nawet dyskusji w Parlamencie Europejskim w ramach dwóch komisji, które zajmują się procedowaniem aktu. Sama jestem ciekawa, do czego dojdziemy. Uważam, że dodanie do listy systemów wysokiego ryzyka generatywnej sztucznej inteligencji, która ma największą zdolność do manipulacji, to pewien krok naprzód. Inne rodzaje systemów sztucznej inteligencji ogólnego przeznaczenia – na przykład rozpoznawanie twarzy, identyfikacja biometryczna czy kategoryzacja biometryczna znajdowały się już wcześniej na liście systemów sztucznej inteligencji wysokiego ryzyka lub nawet na liście zakazanych praktyk. W ich przypadku problem był więc przynajmniej częściowo zaadresowany. Jeśli chodzi o generatory treści, to niestety nie dało się ich tak łatwo przypisać do istniejących propozycji.

EZC: Chciałabym jeszcze doprecyzować drugi istotny element rozporządzenia, czyli listę zakazanych praktyk w zakresie sztucznej inteligencji. Czy, podobnie jak w przypadku wspomnianej wcześniej listy systemów sztucznej inteligencji wysokiego ryzyka, jest to element aktu, który może wymagać szybkiej aktualizacji na skutek rozwoju technologii?

GB: Myślę, że zakazane praktyki są opisane w sposób neutralny technologicznie, dlatego ja bym nie obawiała się, że mogą się one szybko zdezaktualizować. Oczywiście, także i tu pojawiają się wątpliwości, a Chat GPT dał prawodawcom unijnym trochę do myślenia. Problemem jest znowu wyważenie z jednej strony poparcia dla rozwoju tej technologii i wsparcia innowacyjności przez Unię Europejską – przecież, zgodnie z deklaracjami Unii i Komisji, Europa ma się stać centrum rozwoju godnej zaufania sztucznej inteligencji. A z drugiej strony, skoro ta sztuczna inteligencja ma być godna zaufania, to musi uwzględniać potrzebę ochrony ludzi.

W zasadzie największe spory toczyły się wokół zdalnej identyfikacji biometrycznej w czasie rzeczywistym, ponieważ była ona co do zasady zakazana w akcie o sztucznej inteligencji, jednak w dokumencie przewidziano dużo wyjątków w tym zakresie. Pojawił się spór, czy taka identyfikacja powinna być określona jako zdalna czy nie, czy w czasie

rzeczywistym czy nie w rzeczywistym. Tak naprawdę identyfikacja biometryczna prowadzona nie w czasie rzeczywistym również może być bardzo krzywdząca czy też naruszać prawa i wolności osób, które są jej poddane. Z kolei takie cele, jak np. poszukiwanie porwanego dziecka, raczej każdego przekonują, że są przypadki, w których tej technologii powinniśmy używać. Parlament Europejski na początku marca wypracował definicję identyfikacji biometrycznej – czyli jaki jej rodzaj ma być zaliczony do zakazanych praktyk, a jaki z kolei będzie mieścił się w granicach wykorzystania wysokiego ryzyka.

Innym, ostatnio zmienionym elementem aktu jest rozszerzenie zakazu punktacji społecznej, czyli systemów typu *social scoring* na grupy, a nie tylko jednostki. W swoim artykule na temat zakazanych zastosowań sztucznej inteligencji [7] wspominam, że dyskryminacja często tak naprawdę nie dotyczy tylko jednostek, a całych grup społecznych, czy to ze względu na przekonania religijne lub polityczne, czy orientację seksualną, czy też kolor skóry itd. Tę zmianę uważam więc za dobry krok.

Aktualnie prowadzone są także dyskusje na temat manipulacji i tu wracam do inspiracji Chatem GPT. W zakazanych przez akt praktykach wymienione jest stosowanie technik podprogowych przez systemy sztucznej inteligencji, celem wpływania na decyzje osób, jeśli to wywołuje lub może wywołać u nich szkodę. Także tutaj toczyła się debata, czy zakaz ten ma dotyczyć tylko osób podatnych na tego typu manipulacje, czy też może wszystkich. We wspomnianym artykule [7] napisałam, że zakaz powinien obejmować wszystkie osoby. Manipulacja bowiem, co do zasady, jest czymś negatywnym, nie powinna więc dotykać nikogo. I teraz właśnie mówi się o tym, żeby zakazać nie tyle technik podprogowych, co manipulacji i myślę, że jest to też dobry kierunek. Bo manipulacja w żadnej odsłonie, w żadnym kształcie nie powinna być akceptowalna – jest ona sprzeczna na pewno z wartościami moimi, ale myślę, że też z wartościami Unii Europejskiej.

To są jednak takie dylematy, kwestie tak złożone, że Komisja Europejska, która bardzo dobrze przygotowała się do przedstawienia projektu aktu o sztucznej inteligencji i która zaangażowała do pracy nad nim mnóstwo mądrych ludzi, nie była w stanie przewidzieć wszystkiego. I w tym sensie, o ile uważam – jak wspominałam wcześniej – że zapisy dotyczące zakazanych praktyk są neutralne technologicznie, to mogą nie przewidywać wszystkich potencjalnych ryzyk, które wiążą się z tak szybkim

rozwojem technologii. Jak chociażby w przypadku technik manipulacyjnych – niby wiedzieliśmy, że przecież można tworzyć *fake newsy* z wykorzystaniem sztucznej inteligencji, ale nagle okazało się, że to się może wydarzyć na taką skalę i powodować tak daleko idące skutki, że stało się to ogromnym problemem społecznym na całym świecie.

EZC: Zapis dotyczący technik podprogowych także zwrócił moją uwagę, kiedy czytałam pierwszy projekt aktu. Manipulacja może bowiem polegać przecież na wielu innych działaniach – coraz więcej jest chociażby opartych na sztucznej inteligencji wirtualnych postaci, które przy pomocy różnych aplikacji wcielają się w rolę „cyfrowych przyjaciół", z którymi użytkownik może porozmawiać. I gdy zostanie z nim zbudowana relacja, pole do manipulacji lub do przemycenia w rozmowie z nim określonych treści jest bardzo duże i nie wymaga wcale korzystania z technik podprogowych. Jest to realny problem, ponieważ coraz częściej mówi się o tym, że bardzo wiele ludzi jest uzależnionych od korzystania z aplikacji umożliwiających interakcję z takimi czat botami.

GB: Oczywiście. W kontekście sztucznej inteligencji spotkałam się z pojęciami „inżynieria perswazji" i „projektowanie perswazyjne". Dotyczą one projektowania interfejsów w taki sposób, żeby użytkownik był „wciągany" w korzystanie z danej aplikacji lub serwisu, czy to usługowego, czy zakupowego. W przypadku algorytmów chodzi o to, że są tworzone w taki sposób, żeby manipulowały ludźmi, skłaniając ich do podejmowania określonych decyzji, w tym także zakupowych. Jak wiadomo, zazwyczaj chodzi o pieniądze, ale manipulacja może dotyczyć także innych sfer życia. Pokazały to afera Cambridge Analitica czy też głosowanie dotyczące Brexitu. To są bardzo poważne tematy i myślę, że do tej pory wystarczająco się nad nimi w akcie o sztucznej inteligencji nie pochylono.

EZC: Podsumowując naszą rozmowę, chciałabym zadać ostatnie pytanie: co w takim razie, poza *stricte* regulacyjnymi działaniami, które nie do końca nadążają za rozwojem technologicznym, może zrobić Unia Europejska czy też inne instytucje lub organizacje, żeby sprawić, aby sztuczna inteligencja była rozwijana w sposób bezpieczny i etyczny?

GB: Nie chciałabym zabrzmieć cynicznie w tej końcowej części, ale nachodzi mnie następująca refleksja: niestety, w tym przypadku może stać

się tak samo, jak z firmami, które opierają swoją działalność na paliwach kopalnych. Od lat wiedzieliśmy o tym, jak destrukcyjne dla środowiska jest spalanie węgla i ropy. Ostatnio przeczytałam nagłówek, który mnie poruszył – mimo że nie przywiązuję się do dokładności tych danych – że na świecie ludzie wydają w ciągu roku więcej na lody niż państwa na ochronę środowiska i zapobieganie katastrofie klimatycznej. I mam nieodparte wrażenie, że w przypadku sztucznej inteligencji może stać się to samo. Że mówimy o tym, że próbujemy uregulować, że zdajemy sobie sprawę (a przynajmniej większość rozsądnych ludzi), z ryzyk związanych ze sztuczną inteligencją, ale mimo wszystko ta siła rozpędu i pieniądze, które za tym stoją, przeważą. To jest pesymistyczny element w moim techno-optymistycznym światopoglądzie, bo ogólnie uważam, że technologia mogłaby wspaniale zmienić ten świat, m. in. zapobiec katastrofie klimatycznej i wielu problemom związanym z klimatem czy też z głodem na świecie. Tylko, jak widać, nie do końca jest ona do tego używana, a przynajmniej nie są to jej główne zastosowania.

To jest mój pesymistyczny głos, jednak oczywiście wierzę, że sens mają inicjatywy podejmowane zarówno przez instytucje naukowe, jak i przez konsorcja uniwersytetów czy instytucji naukowych z przedsiębiorstwami, które mają na celu wdrażanie rozwiązań etycznych. Powstało mnóstwo różnych wytycznych w zakresie etycznej sztucznej inteligencji, mamy np. nasze europejskie zasady dla godnej zaufania sztucznej inteligencji [6]. Uważam, że grupa ekspertów wysokiego szczebla wykonała świetną pracę, opracowali oni też zestaw do weryfikacji, czy dany system sztucznej inteligencji jest etyczny – *Assessment List for Trustworthy Artificial Intelligence* (ALTAI) [8]. To oczywiście ogólne narzędzie, ale zadaje ono bardzo konkretne pytania. Nawet więc jeśli nie ma się do końca świadomości – a często firmy jej nie mają – o co chodzi z etyką sztucznej inteligencji, to istnieją narzędzia, które mogą pomóc rzeczywiście zweryfikować, czy dane rozwiązanie jest etyczne oraz czy wzięto pod uwagę wszystkie związane z tym aspekty.

Pracuję w dwóch projektach unijnych dotyczących inteligentnego przemysłu, w których rozwijane są rozwiązania oparte o sztuczną inteligencję, mające usprawnić i zoptymalizować produkcję, poprawić jej bezpieczeństwo, efektywność i łańcuch dostaw. Bardzo często obserwuję, że ludzie, którzy pracują w tych projektach mają dość wąski cel – np. zoptymalizować zużycie materiałów lub zużycie energii. Dlatego uważam, że wytyczne i rekomendacje, opracowywane przez różne instytucje, są potrzebne,

bo dają szerszy obraz i możliwość zatrzymania się na chwilę i odpowiedzi na pytania: co my tak naprawdę robimy i po co to robimy? A przede wszystkim, jaki to będzie miało wpływ na społeczeństwo i na środowisko?

Opracowywanie wytycznych i rekomendacji to tworzenie dobrych praktyk na rynku. A im więcej ich będzie oraz im częściej będzie się o tym mówiło, tym bardziej firmy będą świadome tego, jak jest to ważne. Nie tylko ze względów PR-owych, żeby móc się pochwalić, że firma wdraża jakąś tam politykę, ale żeby mieć rzeczywiste narzędzia, z którymi będzie mogła pracować. Tak, jak nauczyliśmy się pracować z RODO, które naprawdę zmieniło podejście do ochrony prywatności i patrzenia na projekty z perspektywy ryzyka, jakie się z nimi wiążą.

Wierzę, a przynajmniej chciałabym wierzyć, że rozwój to nie tylko zarabianie większej ilości pieniędzy i tworzenie doskonalszych algorytmów maszyn, ale również bardziej dojrzałe i odpowiedzialne podejście do technologii. Myślę więc, że różnego rodzaju inicjatywy na rzecz etycznej sztucznej inteligencji, mimo że dobrowolne, są ważne. Podobnie jak edukacja i interdyscyplinarność zespołów, które zajmują się sztuczną inteligencją.

EZC: Bardzo dziękuję Ci za to podsumowanie. Pomimo zapowiedzianego pesymizmu, jest w nim jednak też odrobina nadziei. Chyba nie jesteśmy jeszcze na straconej pozycji.

GB: Też mam taką nadzieję – jest dużo inicjatyw, zarówno w ramach Unii Europejskiej, jak i poza nią. Dla przykładu, Information Commissioner's Office (ICO), czyli brytyjski organ nadzoru, stworzył zestaw narzędzi do oceny systemów sztucznej inteligencji pod kątem przetwarzania danych osobowych zgodnie z zasadami ochrony prywatności [9]. W Stanach Zjednoczonych działa IEEE, która opracowała ciekawy dokument „Ethicaly Aligned Design" [10]. Bardzo interesująca jest też inicjatywa fundacji NeuroRights, która zajmuje się tym, aby rozwiązania w zakresie ingerencji w ludzkie ciało, które są oparte na sztucznej inteligencji, zbierają ogromne ilości danych i mogą mieć wpływ bezpośrednio na ludzki mózg, były rozwijane w sposób etyczny [10]. Dużo więc dzieje się też po tej dobrej stronie mocy.

Bibliografia

[1] Future of Life Institute (22 marca 2023). *Pause Giant AI Experiments: An Open Letter* [Online]. Dostęp: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[2] Garante per la protezione dei dati personali (31 marca 2023). *Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori*, press release in English [Online]. Dostęp: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english.

[3] V. Ordonez, T. Dunn, E. Noll (16 marca 2023). *OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: 'A little bit scared of this'* [Online]. Dostęp: https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122

[4] Załącznik I do wniosku – Rozporządzenia Parlamentu Europejskiego i Rady z 21 kwietnia 2021 ustanawiającego zharmonizowane przepisy dotyczące sztucznej inteligencji (Akt w sprawie sztucznej inteligencji) i zmieniającego niektóre akty ustawodawcze Unii.

[5] OECD, *AI Principles overview* [Online]. Dostęp: https://oecd.ai/en/ai-principles

[6] Niezależna grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji (8 kwietnia 2019). „Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji" [Online]. Dostęp: https://op.europa.eu/pl/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[7] G. Bar, „Zakazane użycie sztucznej inteligencji", *ABI EXPERT* nr 3, 2021.

[8] Niezależna grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji (16 lipca 2020). „The Assessment List for Trustworthy Artificial Intelligence (ALTAI)" [Online]. Dostęp: https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1

[9] Information Commissioner's Office. *AI and data protection risk toolkit* [Online]. Dostęp: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/

[10] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* [Online]. Dostęp: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[11] The NeuroRights Foundation. Dostęp: https://neurorightsfoundation.org/

REDAKCJA | Aleksandra Szczęsna, Monika Stachoń

OPRACOWANIE GRAFICZNE I SKŁAD | Aleksandra Zaręba

ISBN | 978-83-65448-55-2

Dział Strategii i Rozwoju
Bezpieczeństwa Cyberprzestrzeni

NASK – Państwowy Instytut Badawczy

ul. Kolska 12

01-045 Warszawa

cyberpolicy@nask.pl

2023

CyberPOLICY
NASK