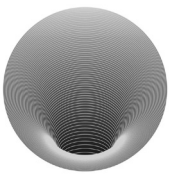
II Systemy bezpieczeństwa sztucznej inteligencji



W ROZDZIALE II:

Mikołaj Kabała
Uniwersytet Jana Kochanowskiego
w Kielcach

Inklinacja systemów sztucznej inteligencji na ataki adwersyjne w logistyce humanitarnej

Streszczenie

W niniejszym artykule skoncentrowano się na analizie inklinacji systemów sztucznej inteligencji używanych w pomocy humanitarnej, na ataki adwersyjne. Celem analizy było wskazanie istotnych słabości w systemach sztucznej inteligencji za pomocą studium przypadku i jego symulacji cyberataków. Osiągnięty został niemały progres w pojmowaniu adwersyjnych manipulacji danych i ich implikacji na ustalenia logistyczne, hipotetycznie dążąc do przekierowania wsparcia z obszarów dotkniętych katastrofą. Metodologia badań obejmuje kompilację testów bezpieczeństwa w kontekście cyberataków na sztuczne sieci neuronowe, a także interpretację pozyskanych danych poprzez wykonanie kodu źródłowego programu aplikacji w języku Python Python. Subsydiarnie wykorzystuje również dane z ogólnodostępnych i globalnych map OpenStreetMap, przy wykorzystaniu dwóch algorytmów, co uwydatnia holistyczne spojrzenie na tematykę problemu. Wyniki badań podkreślają pilną potrzebę opracowania efektywnych metod obronnych ze względu na istotną inklinację systemów sztucznej inteligencji, nie tylko ograniczając się do trwającego konfliktu w Ukrainie, ale także biorąc pod uwagę aspekt globalny. Odnosi się to do specyficznych form ataków adwersyjnych, które w ekstremalnych przypadkach mogą powodować nie tylko nieefektywną relokację zasobów, ale przede wszystkim opóźnienia w dostawach humanitarnych. Opracowane wnioski z analizy otwierają pola poszukiwań odpowiedzi na wzmocnienie odporności realnych systemów i zarazem receptę efektywnej pomocy dla ludzi potrzebujących. Niniejsze badanie wskazuje na nowe perspektywy w kontekście systemów bezpieczeństwa sztucznej inteligencji oraz pomocy humanitarnej, ukazując praktyczne eksplikacje dla wzmocnienia szeroko rozumianej odporności na wszelkie formy manipulacji, co jest istotne dla zapewnienia efektywnego wsparcia miejscowej ludności.

Słowa kluczowe: artykuł naukowy, sztuczna inteligencja, studium przypadku, kooperacyjna AI, atak adwersyjny, CIMIC

Analysis of the Inclination of Artificial Intelligence Systems Towards Adversarial Attacks in Humanitarian Logistics

Abstract

This article focuses on analyzing the inclinations of artificial intelligence systems, which are used in humanitarian aid, towards adversarial attacks. The aim of the analysis is to identify significant weaknesses in artificial intelligence systems, through a case study and its simulation of cyber-attacks. A not inconsiderable progress has been made in understanding adversarial data manipulation and its implications on logistical arrangements, hypothetically seeking to divert support from affected areas. The research methodology includes the compilation of security tests, in the context of cyber-attacks on artificial neural networks, as well as the interpretation of the acquired data through the execution of the application's source code, in Python. Subsidiarily, also uses data from publicly available and global OpenStreetMap maps, using two algorithms, which highlights a holistic view of the subject matter of the problem. The findings highlight the urgent need to develop effective defence methods due to the significant inklings

of artificial intelligence systems, not only limited to the ongoing conflict in Ukraine, but also considering the global aspect. This refers to specific forms of adversarial attacks, which, consequently, in extreme cases, may not only cause inefficient relocation of resources, but, above all, delays in humanitarian supplies. The conclusions of the analysis developed open up fields of search for answers for strengthening the resilience of viable systems and, at the same time, a recipe for effective assistance to people in need. This study points to new perspectives in the context of artificial intelligence (AI) security systems and humanitarian aid, while at the same time, demonstrating practical implications for strengthening resilience to all forms of manipulation in the broadest sense, which is important for providing effective support to the local population.

Keywords: scientific article, artificial intelligence, case study, cooperative AI, adversarial attack, CIMIC

Wstęp

Uzasadnieniem wyboru tematu niniejszego opracowania jest jego dwutorowość – z jednej strony, rewolucyjna rola sztucznej inteligencji we współczesnym świecie, z drugiej – poważne zagrożenie jej działania w logistyce humanitarnej, które wystawia ludzkość na próbę czasu w odpowiedzi na efektywność systemów inteligentnych. Kwintesencją tak określonej problematyki jest w głównej mierze jego nowoczesność. W czasach gdy globalne zagrożenie cyfrowe oraz automatyzacja procesów staje się widoczna bardziej niż kiedykolwiek, innowacyjna sztuczna inteligencja (z ang. *Artificial Intelligence*)[1] wydaje się rudymentarną potrzebą wielu sektorów, szczególnie w logistyce humanitarnej. Systemy AI, stosowane do optymalizacji i zarazem personalizacji dystrybucji pomocy, gospodarowania zasobami czy przede wszystkim prognozowania potrzeb miejscowej ludności, mają niemały wpływ na znaczny wzrost efektywności oraz skuteczności dostaw humanitarnych. Jednakże globalny wzrost zainteresowania sztuczną inteligencją nie tylko przynosi pokonanie wielu „kamieni milowych", ale także uwidacznia problemy i wyzwania, z którymi ludzkość już dziś musi, a w przyszłości będzie się prawdopodobnie musiała zmierzyć (chodzi o zagrożenia na niespotykaną dotąd skalę, a zwłaszcza w przestrzeni cyberbezpieczeństwa). Tym samym do jednych z najgroźniejszych problemów w logistyce humanitarnej można zaliczyć ataki adwersyjne (ang. *Adversarial Attacks*)[2] na aplikacje nawigujące, których zadaniem jest celowe wprowadzenie subtelnie zmodyfikowanych danych wejściowych, na bieżąco wykorzystywanych przez kierowców takiego transportu. Powoduje to błędne działanie systemów AI. Konsekwencją takich okoliczności bywa złe podejmowanie decyzji lub działań.

Nieustanny rozwój technologii sztucznej inteligencji (AI) i jej obecności w logistyce humanitarnej kreuje obietnicę „lepszego jutra", w którym wszelkie wsparcie może być dostarczane sprawniej, precyzyjniej i efektywniej. Czy to oznacza, że je-

[1] Sztuczna inteligencja (ang. *Artificial Intelligence* lub *AI*) to akcesoria i urządzenia, które uwidaczniają inteligencję w przeciwieństwie do inteligencji naturalnej. Jako autora tego terminu wskazuje się na Johna McCarthy'ego.

[2] Ataki adwersyjne (ang. *Adversarial Attacks*) to ataki, których zadaniem jest celowa manipulacja, występująca w modelach uczenia maszynowego, przy pomocy dokładnie zmodyfikowanych danych wejściowych. Więcej informacji: (Sellakumar i in., 2024).

steśmy w pełni bezpieczni? Jak można interpretować kwestię szeroko rozumianego bezpieczeństwa w kontekście pomocy humanitarnej? Na pierwszy rzut oka problem przekierowania takiej dostawy wyda się błahy. Jednak gdy przyjrzymy mu się nieco głębiej, dostrzeżemy daleko idące i bardzo poważne skutki dla regionu dotkniętego katastrofą, takie jak znaczne opóźnienia w dostawach niezbędnych zapasów medycznych i żywności, co w sytuacji kryzysowej może prowadzić do dodatkowych ofiar wśród lokalnej ludności. Ponadto błędne przekierowania pomocy mogą skutkować nieefektywnym wykorzystaniem zasobów i funduszy, co w długoterminowej perspektywie osłabia zdolność organizacji humanitarnych do reagowania na przyszłe katastrofy. Dlatego tak istotne są mocne i zaktualizowane (w czasie rzeczywistym) zabezpieczenia systemów. Co może się stać, jeśli ich zabraknie? Łatwo stwierdzić, że wówczas każda z technologii będzie podatna na ataki, które wywołają w konsekwencji trudności w funkcjonowaniu, a tym samym zwiększą ryzyko dla osób najbardziej potrzebujących. Za pomocą analizy podatności, jak również sugerowanych recept bezustannie dążymy do kształtowania systemu odpornego nie tylko na zagrożenia, ale przede wszystkim do korzystania z potencjału sztucznej inteligencji, która będzie sprawować istotną rolę, zwłaszcza w sytuacjach kryzysowych (np. ataku adwersyjnego). Świadomość skali zagrożenia oraz zrozumienia i pokonania „kamieni milowych" to zadania dla logistyki humanitarnej w świecie zdominowanym przez cyfryzację.

Wdrożenie AI w logistyce humanitarnej nie tylko otwiera nowe pola poszukiwań odpowiedzi, ale równolegle stawia zasadnicze pytanie: *Jak systemy inteligentne wykazują odporność na manipulację we współczesnym świecie?* Deliberując nad ową problematyką, należy bowiem zlokalizować wszelkie inklinacje systemów AI, a tym samym pojąć, jak my, ludzie, możemy wzmocnić te systemy w odniesieniu do domniemanych zagrożeń. Ogólnym i rudymentarnym obszarem badania jest jego wielowymiarowe podejście i uwypuklenie nowych perspektyw w obliczu zagrożeń. Rama interpretacyjna spaja wiedzę z dziedziny zarządzania kryzysowego, etyki, a także informatyki oraz bezpieczeństwa cybernetycznego.

Zgłębienie tajników wiedzy o mechanice oraz potencjalnych następstwach ataków adwersyjnych, w kontekście inteligentnych systemów i logistyki humanitarnej prowadzi do intensywnej analizy zarówno jej aspektów technologicznych, jak i obszaru użytkowania, przy założeniu prawdopodobieństwa minimalnego marginesu błędu oraz niesłychanie wysokiej stawki. Generalizując tak przedstawione treści, zakłada się, że przedmiotem badań będzie interakcja systemów inteligentnych na ich inklinację w zakresie logistyki humanitarnej. Mając na względzie powyższe uzasadnienia, założono, że celem niniejszego opracowania stanie się realizacja scenariusza studium przypadku (przy wykorzystaniu opracowanego kodu źródłowego programu aplikacji) na wpływ innowacyjnych systemów inteligentnych, skorelowanych z implikacjami użytkowania współczesnej technologii oraz propozycja horyzontów działań, które mogą się przyczynić do wzmocnienia obecnego i przyszłego bezpieczeństwa, a także efektywności wsparcia humanitarnego. W odniesieniu do tak postawionego celu oraz przedmiotu badań sformułowano reprezentacyjny i zarazem główny problem badawczy:

Jaki jest zakres podatności współczesnych systemów sztucznej inteligencji stosowanych w logistyce humanitarnej na ataki adwersyjne oraz jakie mogą wystąpić implikacje dla tych ataków na niezawodność świadczenia wsparcia humanitarnego?

Rozpoznanie, za pomocą kwerendy literatury przedmiotu oraz zakresu merytorycznych treści, umożliwiło zaproponowanie następującej hipotezy badawczej:

Przypuszczam, że wykorzystywane systemy AI w logistyce humanitarnej inklinują na ataki adwersyjne, które, w konsekwencji, mogą zasadniczo zakłócić ich funkcjonowanie oraz niezawodność. Zakładam, że przy opracowaniu odpowiednich strategii i metod obronnych przy kreowaniu obecnych i przyszłych inteligentnych systemów możliwy jest regres tej podatności, wpływający na progres w efektywności wsparcia humanitarnego.

Do wypracowania skonstruowanego głównego problemu badawczego wykorzystano teoretyczne oraz empiryczne metody badawcze: analizę, kontrastowanie, diagnozę, prognozę, indukcję, dedukcję, a także badanie eksperymentalne. W części „Metody i część eksperymentalna" przedstawiono techniki oraz metodologie użyteczne przy rozpoznaniu i analizie inklinacji współczesnych systemów sztucznej inteligencji na ataki adwersyjne. W tym celu wykorzystano opracowany kod źródłowy (w postaci pseudokodu[3]) w języku Python[4], przeprowadzając próby w środowisku CodeHS celem analizy transformacji w logistyce humanitarnej. Przeprowadzone testy pozwoliły na ocenę algorytmów wykrywania i zwalczania ataków adwersyjnych, co umożliwiło empiryczną walidację założeń teoretycznych oraz określenie potencjalnych luk w AI, ułatwiających ataki.

1. Przegląd badań nad bezpieczeństwem i obronnością w systemach sztucznej inteligencji, aplikowanych w logistyce humanitarnej

Odnosząc się do ataków adwersyjnych oraz systemów sztucznej inteligencji (AI), można stwierdzić, że ich istotą pozostaje świadomość istnienia oraz zrozumienie mechanizmów obronnych, jak również wyzwań skorelowanych z funkcjonowaniem AI. Do cennych źródeł, które pozwoliły zgłębić badaną problematykę, zaliczyć można:

- opracowanie naukowe *Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A survey*", autorstwa H. Banieckiego oraz P. Biecka (2023), koncentrujące się na holistycznym przeglądzie interpretowalnych ataków adwersyjnych. Wskazuje ono również strategie obronne w spektrum sztucznej inteligencji. Co prawda, nie skupia się dokładnie na logistyce humanitarnej, ale podkreśla w nim kluczowy element funkcjonowania takich systemów. Autorzy prezentują również

[3] „Pseudokod" to nieformalny sposób kodowania algorytmu, stanowiący uproszczoną formę języka programowania. Jego celem jest ukazanie skomplikowanej składni w prosty, zrozumiały i inspirujący dla odbiorcy sposób.

[4] „Python" to wysokopoziomowy i wieloparadygmatowy język programowania przeznaczenia ogólnego, wykorzystujący złożoność pakietów standardowych bibliotek. Jego atutami są wysoka elastyczność i prostota.

ataki, których celem są metody wyjaśnialne, np. manipulacje przykładami adwersyjnymi, bez konieczności zmian w predykcji modelu danych (Baniecki i in., 2023). Praca ta zwraca uwagę na konieczność oceny jakości wyjaśnień, głównie dla zaawansowanych modeli, rzutujących na scenariusze adwersyjne. Artykuł pozwala zrozumieć możliwe wrogie działania w kontekście manipulacji i jej wpływu w logistyce humanitarnej na interpretację modeli, jak również wskazuje rozwiązania zabezpieczeń systemów sztucznej inteligencji jako działania proaktywne w odniesieniu do ataków na te systemy (Baniecki i in., 2023);

- opracowanie naukowe *Adversarial Attacks in Cooperative AI* autorstwa T. Fujimoto oraz A.P. Pedersena (2021) skupia się na obfuskacji[5] w systemach sztucznej inteligencji, której celem jest kreowanie scenariuszy posiadających parametry trzech algorytmów: wierzeń, sekwencyjnych dylematów społecznych oraz aproksymacji średniego pola (Fujimoto i in., 2021). Taki pogląd na sztuczną inteligencję rzuca nową perspektywę na istotność bezpieczeństwa logistyki humanitarnej. Ukazuje, w jaki sposób ataki adwersyjne implikują procesy decyzyjne w wielowymiarowych środowiskach (Fujimoto i in., 2021), co podkreśla słuszność użytkowania podczas koordynacji w pomocy humanitarnej. Zaprezentowane przez autorów metody wskazują na nowe podatności w kontekście kooperacyjnej AI, dla których przedtem nie przeprowadzono badań uczenia maszynowego z oponentem. Tak przedstawiona problematyka skłania do wzmacniania strategii obronnych, skorelowanych dla współpracujących systemów sztucznej inteligencji w obszarze humanitarnym.

Przywołane i omówione w skrócie teksty są nie tylko centrum istotnych informacji w kontekście możliwych zagrożeń i wyzwań, ale także źródłem metod obronnych w obszarze logistyki humanitarnej oraz systemów AI. Jednocześnie ukazują zasadność interpretowalności oraz kooperacji w odpowiedzi na ataki adwersyjne. Implementacja odpowiednich komponentów w strategiach obronnych, jak również zrozumienie takich scenariuszy wskazują na kluczowość dla lepszych zabezpieczeń systemów sztucznej inteligencji, głównie przed manipulacją, będącą poważnym zagrożeniem w świetle krytycznych scenariuszy humanitarnych.

2. Metodologia i procedury badawcze w analizie odporności systemów sztucznej inteligencji na ataki adwersyjne w kontekście logistyki humanitarnej

W tej części pracy starannie dokonano opisu metodologii badań celem umożliwienia innym badaczom jej odwzorowanie oraz krytyczną analizę. Skupiono się na skrupulatnym zobrazowaniu technik i procedur zaimplementowanych do identyfikacji oraz analizy podatności systemów sztucznej inteligencji (AI) na ataki adwersyjne w kontekście logistyki humanitarnej. Badanie składa się z dwóch istotnych metod badawczych: studium przypadku i eksperymentu, które współgrają ze sobą, aby zro-

[5] Obfuskacja (ang. *obfuscation*) to technika zaciemniania kodu źródłowego, która powoduje, że jego analiza jest znacznie utrudniona.

zumieć i analizować podatność systemów sztucznej inteligencji (AI) na ataki adwersyjne w kontekście logistyki humanitarnej. Studium przypadku zastosowano w celu dogłębnego zrozumienia specyficznych wyzwań w implementacji systemów AI w logistyce humanitarnej, wykorzystując trasę Jarosław – Awdijiwka jako przykład realnych warunków operacyjnych. Eksperyment, realizowany za pomocą opracowanej aplikacji symulacyjnej w języku Python, służył walidacji efektywności wykrywania i neutralizacji ataków adwersyjnych w tych systemach.

- **wybór obszaru badań**

W ramach wyboru obszaru badań podjęto się analizy studium przypadku w spektrum logistyki humanitarnej. Miasta, które wzięto pod uwagę, to: Jarosław (województwo podkarpackie, Polska), położone w okolicy Pogórza Rzeszowskiego (Miasto Jarosław, 2024) i Awdijiwka (region Donieck, Ukraina)[6]. Taka decyzja była podyktowana nie tylko trwającymi konfliktami w wybranym regionie Ukrainy na dzień 10 lutego 2024 roku, ale przede wszystkim kompleksowością i znaczeniem tej trasy dla badania wpływu ataków adwersyjnych na systemy sztucznej inteligencji, asekurujących logistykę w sytuacjach kryzysowych. Za pomocą nieszablonowości przedstawionego studium przypadku łatwo skomasować jego wielowymiarowość, przy jednoczesnym uproszczeniu działań tych systemów: od zobrazowania bezprecedensowych wyzwań logistycznych aż po wystawienie technologii AI na próbę poprzez zamierzone działania adwersyjne przy określonych założeniach i ograniczeniach. Fakultatywnie jako rodzaj manipulacji oraz przy jednoczesnej zmianie algorytmu zobrazowano potencjalną formę ataków adwersyjnych na systemy sztucznej inteligencji w kontekście dostaw humanitarnych. Wnikliwe zrozumienie i świadomość funkcjonowania tych czynników pozwala na zdiagnozowanie inklinacji w takich systemach, a tym samym ukazuje propozycje metod dla ich zabezpieczeń na wypadek takiego scenariusza.

Trasa Jarosław – Awdijiwka w spektrum obszaru badań wskazuje na jego dwutorowość. Z jednej strony, to zobrazowanie specyficznych wyzwań we wdrożeniu systemów sztucznej inteligencji do obszaru logistyki humanitarnej, z drugiej – wzrost znaczenia geograficznego w weryfikacji oraz wzmacnianiu wspomnianych systemów w odpowiedzi na ataki adwersyjne wroga.

W ramach studium przypadku analiza danych obejmowała przegląd dostępnych publicznie map i danych geograficznych na trasie Jarosław – Awdijiwka. Zebrane informacje pomogły rozpoznać kluczowe wyzwania w zakresie bezpieczeństwa i wydajności systemów AI w realnych warunkach operacyjnych.

- **narzędzia, technologie oraz selekcja danych w algorytmach**

W badaniami skorelowanymi nad oddziaływaniem ataków adwersyjnych na systemy AI w dostawie humanitarnej kwintesencją jest wykorzystanie technologii i szeregu narzędzi, przy pomocy których da się wykonać wnikliwą analizę oraz symulację ataku w symulacji scenariusza studium przypadku. Istotnymi elementami, które posłużyły w badaniu, są:

[6] Awdijiwka – to miasto obwodu donieckiego, we wschodniej części Ukrainy, jedna ze starszych miejscowości.

- CodeHS – to zsynchronizowane środowisko w języku Python, w którym zaprogramowana została aplikacja do symulacji scenariusza studium przypadku;
- Python – to wysokopoziomowy i wieloparadygmatowy język programowania przeznaczenia ogólnego, wykorzystujący złożoność pakietów standardowych bibliotek. Wybór tego języka jest spowodowany jego holistycznym zastosowaniem, zwłaszcza wśród społeczności naukowej. Upraszcza się tym samym przekaz wzajemnych doświadczeń i wiedzy. Jego atutami są wysoka elastyczność i prostota;
- Tkinter – to biblioteka języka Python wykorzystana do kreowania graficznego interfejsu aplikacji, co implikowało intuicyjną prezentację rezultatów analiz z uwzględnieniem prostszej manipulacji obrazów oraz parametrów symulacji. To również tradycyjny interfejs graficzny (ang. *Graphical User Interface*, *GUI*), którego użytkowanie opierało się na stworzeniu okien, etykiet graficznych, przycisków interfejsu, jak również innych, mniej istotnych elementów interfejsu użytkownika;
- PIL – (ang. *Python Imaging Library*) – to biblioteka języka Python użyta do manipulacji kilku obrazów, a w tym ich: pobierania, przetwarzania oraz projekcji map, jak również pozostałych graficznych elementów w programie aplikacji, nieodzownych podczas analizy badań. Biblioteka ta jest niekiedy wykorzystana do niezawodnego zarządzania obrazami, niezbędnego podczas deliberacji ataków adwersyjnych. Wówczas, implikując, taka manipulacja może się stać jedną z potencjalnych technik ataku na systemy AI;
- OpenStreetMap[7] oraz algorytmy OSRM (ang. *Open Source Routing Machine*)[8] i Valhalla[9] – to trzy elementarne narzędzia będące podstawą do przeprowadzonych badań. OpenStreetMap stanowi pomoc podczas symulacji danej trasy oraz analizy możliwych ataków adwersyjnych, przy użyciu zrzutów ekranu z ogólnodostępnych i globalnych bazy danych map. Z kolei algorytmy OSRM oraz Valhalla okazują się jedynie statyczną symulacją dwóch scenariuszy podróży do zadanego celu przy zastosowaniu ich w odpowiedniej sekwencji obrazującej kroki detekcji i neutralizacji potencjalnego ataku adwersyjnego.

Narzędzia, takie jak CodeHS i Python, umożliwiły stworzenie złożonych scenariuszy symulacyjnych ataków adwersyjnych, które były następnie przeprowadzane przeciwko modelowi AI. Za pomocą biblioteki Tkinter stworzony został interfejs użytkownika, który pozwalał na intuicyjne śledzenie przebiegu uproszczonej symulacji oraz na bieżącą analizę wyników.

W ramach implementacji i symulacji projektu wykonana została aplikacja graficzna w Pythonie z uwzględnieniem biblioteki Tkinter dla graficznego interfejsu

[7] OpenStreetMap – to darmowa, ogólnodostępna baza danych kuli ziemskiej, nieustannie aktualizowana przez zarejestrowanych użytkowników. To pomocne narzędzie podczas wizualizacji danych, nawigacji tras oraz przede wszystkim wsparcia humanitarnego.

[8] OSRM – (ang. *Open Source Routing Machine*) – to ogólnodostępny algorytm i zarazem system wspierający efektywne optymalizacje tras na podstawie bazy danych OpenStreetMap.

[9] Valhalla – to ogólnodostępny algorytm w postaci oprogramowania, którego celem jest wyznaczenie złożonych tras z uwzględnieniem elastyczności i różnych środków transportu.

użytkownika (GUI), a biblioteka PIL posłużyła do eksploatacji obrazów. Głównym celem aplikacji jest zobrazowanie symulacji studium przypadku dotyczącego detekcji oraz neutralizacji potencjalnego ataku adwersyjnego na systemy AI w spektrum pomocy humanitarnej;

- **kod źródłowy i procedury badawcze**

Rozważając wykonany kod źródłowy, który posłużył jako narzędzie i zarazem aplikacja graficzna do odtworzenia symulacji studium przypadku dotyczącego detekcji i neutralizacji ataku adwersyjnego na systemy sztucznej inteligencji, można stwierdzić, że składa się on z kilku sekwencji wywoływanych chronologicznie, by ukazać kwintesencję badanego problemu[10]. Głównym celem kodu źródłowego nie jest wnikliwe jego omawianie, ale jedynie, na jego podstawie, opis najistotniejszych struktur. W ramach opracowanej aplikacji zaimplementowano dwie istotne funkcje: 'wykryj_anomalie' (Rys. 1) oraz 'neutralizuj_atak' (Rys. 2), stanowiące procedurę badawczą w procesie detekcji ataków adwersyjnych, jak również w mechanizmach obronnych ich neutralizacji.

Poniżej znajdują się obie funkcje ze wskazaniem na ich rolę w programie aplikacji:

```
# Wykrywanie anomalii, zestawiając przykładowe dane drogowe
# z przykładowymi bezpiecznymi danymi. Jeśli stany dróg są różne,
# to wówczas taki stan drogi jest uznawany za anomalię.
def wykryj_anomalie(self):
    self.anomalie = []
    for droga, stan in self.przykladowe_dane_drogowe.items():
        if stan != self.przykladowe_bezpieczne_dane.get(droga, not stan):
            self.anomalie.append(droga)
    return self.anomalie
```

Rys. 1. Funkcja 'wykryj_anomalie', przeszukująca bazę danych drogowych.

Źródło: Opracowanie własne, przy pomocy środowiska CodeHS.

```
# Symulacja neutralizacji ataku adwersyjnego - polega na dostosowaniu
# flag wszystkich dróg na zgodny z przykładowymi bezpiecznymi danymi.
# Fakultatywnie, interfejs użytkownika jest aktualizowany,
# aby finalnie wyświetlił docelową mapę po neutralizacji anomalii.
def neutralizuj_atak(self):
    for droga in self.przykladowe_dane_drogowe:
        if droga in self.przykladowe_bezpieczne_dane:
            self.przykladowe_dane_drogowe[droga] = self.przykladowe_bezpieczne_dane[droga]
    self.aktualizuj_informacje(neutralizacja=True) # Aktualizacja informacji po neutralizacji
    self.pokaz_mape('Mapa.png')
```

Rys. 2. Funkcja 'neutralizuj_atak', odtwarzająca bezpieczny stan dróg.

Źródło: Opracowanie własne, przy pomocy środowiska CodeHS.

[10] EasyPaste – to serwis i jednocześnie darmowy hosting do przesyłania plików publicznych i prywatnych; https://www.easypaste.org/file/SUy126Rf/Aplikacja.kod.zrodlowy.zip?lang=pl (dostęp 3.02.2024).

Funkcja 'wykryj_anomalie' pozwala na dokonywanie kwerendy zbioru danych drogowych, zestawiając obecny stan każdej drogi z prognozowanym bezpiecznym. Sama zaś anomalia jest rozpoznawana wówczas, gdy stan drogi pozostaje różny od bezpiecznego, co może implikować realny atak adwersyjny (np. nieprawdziwe dane techniczne na temat stanu drogi). Rozpoznane anomalie są gromadzone w liście[11], co pozwala na dalszą analizę programu w odpowiedzi na kolejne sekwencje zdarzeń.

Funkcja 'neutralizuj_atak' umożliwia odtworzenie bezpiecznego statusu dróg, odnosząc się do uprzednio zdefiniowanych, prognozowanych bezpiecznych danych. Jeśli anomalia zostanie wykryta, wówczas status drogi zmienia się do bezpiecznego, jak również wyświetlana jest prawidłowa mapa. To uproszczona wersja symulacji neutralizacji ataku adwersyjnego. Funkcja ta ukazuje proste zabezpieczenia, które mogą być znacznie bardziej złożone w istniejących już interfejsach.

Opracowany kod źródłowy, zawierający funkcje 'wykryj_anomalie' i 'neutralizuj_atak', był poddawany wielokrotnym testom, aby zapewnić jego skuteczność w realistycznych warunkach. Te testy miały na celu ocenę zdolności algorytmów do wykrywania i neutralizowania potencjalnych zagrożeń, wykorzystując różnorodne scenariusze, co umożliwiło dokładne dostosowanie metod w odpowiedzi na wykryte anomalie;

- **warunki eksperymentalne**

Eksperyment został wykonany w monitorowanym środowisku testowym w celu zagwarantowania rzetelności oraz jednolitości wyników. Takie podejście zapewnia solidną podstawę dla wiarygodnej analizy i interpretacji uzyskanych danych. Konfiguracja sprzętowa została dobrana tak, aby odzwierciedlać potencjalne warunki operacyjne, w których mogą pracować systemy AI. Jednocześnie pozwoliło to potwierdzić fakt, że wnioski będą odpowiednio aplikowalne w różnorodnych środowiskach logistycznych. Poniżej zaprezentowano szczegółowe dane techniczne warunków eksperymentalnych:

Warsztat programisty:

– *Środowisko programistyczne:* całość eksperymentu wykonana została przy wykorzystaniu środowiska CodeHS, stanowiącego narzędzie i zarazem intuicyjną platformę cyfrową w formie warsztatów programistycznych. To narzędzie umożliwia fakultatywnie zintegrowane środowisko programistyczne, oferując efektywną analizę, ocenę, diagnozę i prognozę kodu oprogramowania.
– *Język programowania:* w celu wdrożenia aplikacji użyto uniwersalnego języka wysokopoziomowego Python w wersji 3.8, wspierającego programowanie obiektowe[12]. Umożliwia on nie tylko szerokie spektrum zastosowań w algorytmach sztucznej inteligencji, ale także w szeroko rozumianej analizie danych.

[11] Lista (ang. *list*) to reprezentacyjna i sekwencyjna struktura zbiorów dynamicznych, której ideą jest usytuowanie elementów w układzie linearnym.

[12] Programowanie obiektowe – to paradygmat programowania, którego celem jest kreowanie programów, na podstawie współdziałających zdefiniowań oraz oddziaływań między obiektami. Natomiast, obiekty to nic innego, jak struktury, które spajają dane (pola), a także funkcje (metody) funkcjonujące w obrębie tych danych. Więcej informacji (Lutz, 2015).

– *Biblioteki:* w ramach języka Python posiłkowano się jedynie wymaganymi bibliotekami do zobrazowania scenariusza studium przypadku. Są to biblioteki: Tkinter oraz PIL omawiane we wcześniejszym etapie tego rozdziału. Ich wykorzystanie wiązało się ponadto z kompatybilnością wersji języka programowania Python.

Konfiguracja sprzętowa:

– *Procesor:* eksperyment został wykonany na laptopie o parametrach procesora Intel Core i7-9750H (2.6 GHz), co jest w pełni wystarczającym zapotrzebowaniem obliczeniowym.
– *Pamięć RAM:* urządzenie zostało wyposażone w 32 GB pamięci RAM, wpływając na płynne działanie aplikacji oraz przetwarzanie danych.
– *System operacyjny:* eksperyment wykonano na systemie operacyjnym Windows 11 zintegrowanym z wykorzystywanym akcesorium.

Środowisko testowe:

– *Dane testowe:* poddana symulacja ataków adwersyjnych posłużyła się uproszczonymi zbiorami danych w postaci statycznych zdjęć. Takie dane zawierały prawidłowe, jak również spreparowane dane trasy. W celu prostej obróbki map użyto ogólnodostępnych map ze źródła OpenStreetMap, jak również potencjalnego, symulowanego scenariusza studium przypadku.
– *Reprodukowalność:* eksperyment zapewnia jego reprodukowalność i zdolność repetycyjną, począwszy od doboru narzędzi oraz selekcji danych w algorytmach po procedury badawcze. Jednocześnie kod źródłowy wykonanej aplikacji oraz instrukcja obsługi i załączniki (w postaci obrazów map) zostały udostępnione w internetowym repozytorium w części poświęconej procedurom badawczym.

• **założenia metodologiczne i bariery metod**

– *Założenia związane z danymi*: przyjęto, że wszelkie informacje pochodzące ze źródła OpenStreetMap oraz algorytmów OSRM i Valhalla są aktualne i dokładne. Jednocześnie założono, że całościowość oraz jakość danych ze źródeł globalnych i ogólnodostępnych może być niejednolita, co implikuje dogłębność symulacji oraz przeprowadzonych analiz.
– *Założenia związane z algorytmami*: założono, że wdrażane algorytmy wykrywania ataków adwersyjnych są efektywne w rozpoznawaniu znanych rodzajów ataków. Natomiast te algorytmy mogą nie mieć pełnego przystosowania w odpowiedzi na nieznajome lub nowe strategie ataków, co jest zasadne w kontekście postępu technologicznego, a tym samym wciąż rozwijających się technik adwersyjnych.

Bariery metodologiczne:

– *Dane:* niepełność danych ze źródeł ogólnodostępnych (np. OpenStreetMap) może wpływać na precyzję symulacji potencjalnych ataków adwersyjnych.
– *Algorytmy:* opracowane wzorce detekcji mogą być nieefektywne lub ograniczone podczas nieszablonowych ataków w kontekście neutralizacji enigmatycznych zagrożeń.

– *Sprzęt:* specyfikacja techniczna sprzętu determinuje efektywność i wyniki algorytmów.

Wpływ na wyniki i wnioski:

– *Interpretacja:* wskazana jest ostrożność w interpretacji rezultatów ze względu na potencjalne mankamenty posiadanych danych oraz algorytmów, co podkreśla kluczowe znaczenie podczas poprawnej ekstrapolacji wniosków.
– *Spektrum wniosków:* uwarunkowania wyznaczają zakresy dla konkluzji, przede wszystkim w kontekście skuteczności rozpoznawania oraz zwalczania ataków w ewoluujących warunkach.

- **eksperyment**

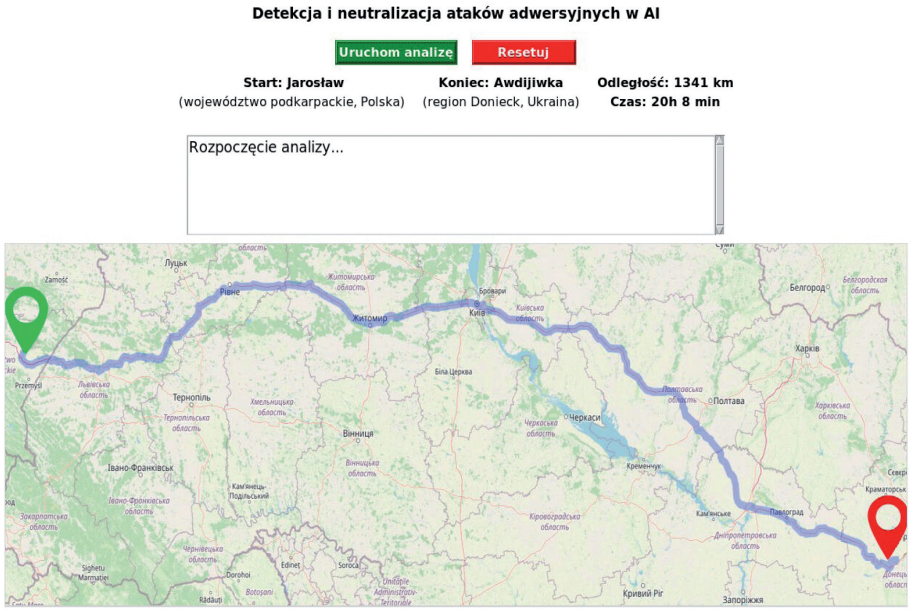
W ramach realizacji celów badawczych niniejsze badanie opierało się na dwóch głównych metodach: studium przypadku i studium eksperymentu. Studium przypadku było skoncentrowane na analizie realnych warunków operacyjnych na trasie Jarosław – Awdijiwka, które pozwoliło zrozumieć specyficzne wyzwania w implementacji systemów sztucznej inteligencji w logistyce humanitarnej. Skupiono się na zebraniu i analizie danych z publicznie dostępnych map, co pozwoliło na dogłębne zrozumienie kontekstu operacyjnego.

Eksperyment, realizowany za pomocą opracowanej aplikacji symulacyjnej w języku Python, służył do bezpośredniej walidacji efektywności systemów AI w wykrywaniu i neutralizacji ataków adwersyjnych. Użyto uproszczonego kodu źródłowego (pseudokodu), który symulował serię procedur badawczych stanowiących sekwencję poszczególnych kroków walidacji potencjalnego ataku adwersyjnego na systemy sztucznej inteligencji z wykorzystaniem specjalnie zaprojektowanych funkcji wykryj_anomalie i neutralizuj_atak. Proces eksperymentalny rozpoczynał się od uruchomienia aplikacji, co prowadziło do sekwencyjnego przeglądu kroków symulujących atak i odpowiedź systemu.

Kompilacja oraz aktywizacja aplikacji w środowisku CodeHS umożliwiła bezzwłoczną obserwację funkcjonowania systemu. Uruchomienie aplikacji powoduje, że ekran początkowy (Rys. 3) składa się z opisu przycisków: „Uruchom analizę" i „Resetuj" oraz danych wejściowych, bez uwzględnienia odległości i czasu. Kliknięcie przycisku „Uruchom analizę", tuż po wyświetleniu ekranu początkowego, powoduje zatem proces analizy danych początkowych (Rys. 4). Następnie występuje diagnozowanie potencjalnych ataków adwersyjnych (Rys. 5) oraz opóźniona o kilka sekund reakcja systemu symulująca likwidację zagrożeń (Rys. 6). Warto zwrócić szczególną uwagę na to, że parametry: „Odległość" i „Czas", a także obrazy zmieniają się w zależności od etapu sekwencji programu.

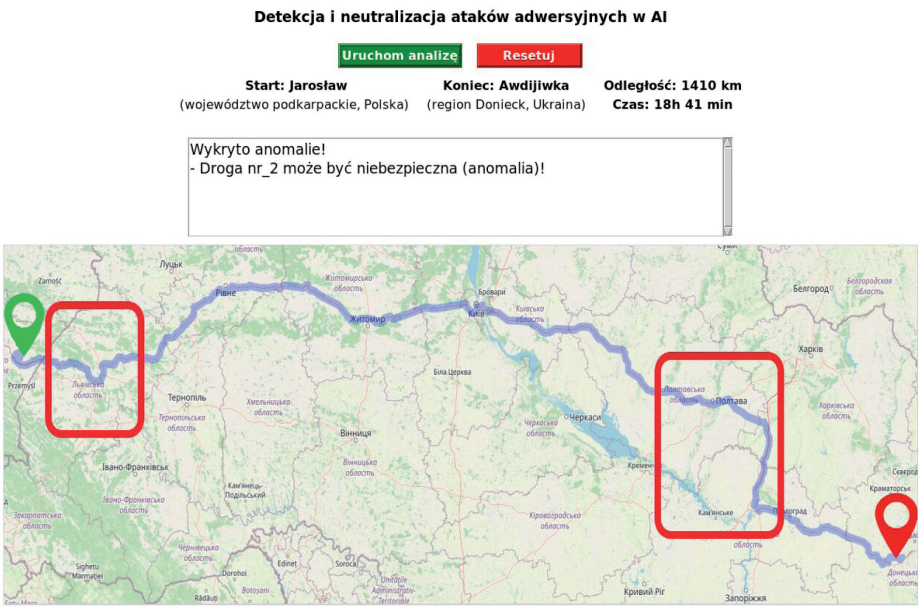
Rys. 3. Główny interfejs programu z parametrami wejściowymi.

Źródło: Opracowanie własne, przy użyciu środowiska CodeHS oraz języka Python.



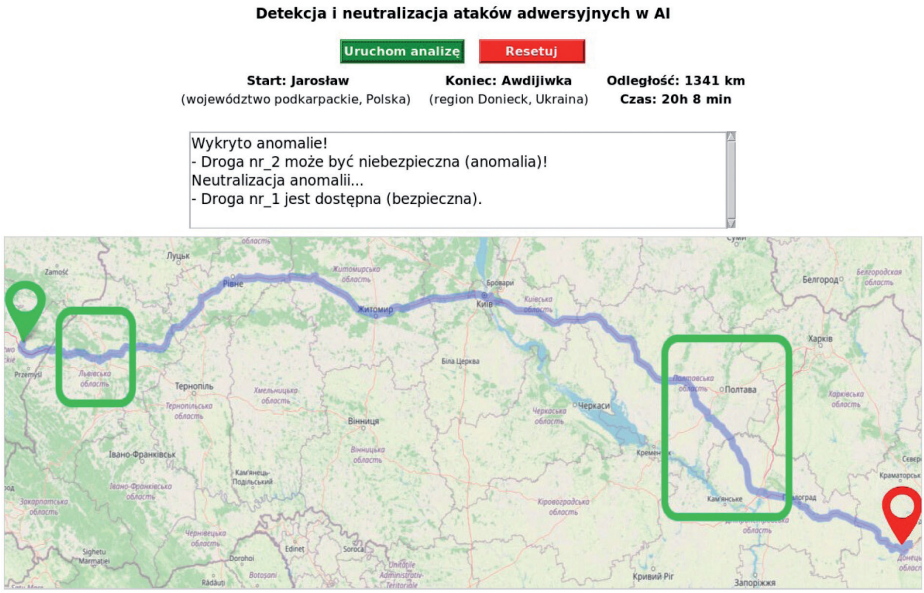
Rys. 4. Pierwsza sekwencja programu, rozpoczynająca wykrywanie potencjalnego ataku.

Źródło: Opracowanie własne, przy użyciu środowiska CodeHS oraz języka Python.



Rys. 5. Druga sekwencja programu, wykrywająca anomalię.

Źródło: Opracowanie własne, przy użyciu środowiska CodeHS oraz języka Python.



Rys. 6. Trzecia sekwencja programu, neutralizująca anomalię.

Źródło: Opracowanie własne, przy użyciu środowiska CodeHS oraz języka Python.

Eksperyment zwizualizował istotne odkrycia, aprobując założone hipotezy oraz sugerując optymalizację i ewolucję sztucznej inteligencji odpornej na wrogie ataki. Warto również zaznaczyć, że w drugiej sekwencji na obrazach zostały zaznaczone fragmenty wskazujące na potencjalną anomalię na trasie (Rys. 5). Na koniec następuje ostatnia sekwencja neutralizująca zagrożenie wraz z ukazaniem prawidłowej i zarazem początkowej mapy (Rys. 6), co wieńczy eksperyment.

3. Wyniki badań

Wyniki badań dostarczyły nowych perspektyw na pojmowanie ataków adwersyjnych w systemach sztucznej inteligencji. W odróżnieniu od tradycyjnego podejścia prezentowania tego zagadnienia podjęto próbę interaktywnej wizualizacji problematyki, które umożliwiają, w sposób prosty, zrozumienie złożonych wzorców oraz intuicyjne eksplorowanie informacji przez osoby niezaznajomione z ową tematyką. Poprzez analizę wyników wykazano, że w odniesieniu do przedstawienia danych o znacznej zmienności systemy sztucznej inteligencji mogą przedstawiać niestabilną odporność na specyficzne typy ataków adwersyjnych. Ewenement ten kwestionuje szeroko akceptowalną hipotezę, że wieloaspektowość danych automatycznie przekłada się na intensywność wystąpienia ataków.

Jako słuszność tej tezy można przedstawić opracowanie naukowe pt. „Defending against adversarial attacks on medical imaging AI system, classification or detection?", autorstwa X. Li, D. Pan, D. Zhu, w którym autorzy eksplorują obronę przed

atakami adwersyjnymi na systemy AI do obrazowania medycznego, podkreślając wyzwania związane z unikatowym charakterem danych medycznych. Ich badanie pokazuje, że mimo wieloaspektowości danych medycznych specyficzne dla tej dziedziny wyzwania, takie jak podobieństwo między klasami obrazów, mogą sprawić, że systemy te będą nadal podatne na ataki adwersyjne. To wskazuje, że sama wieloaspektowość danych nie gwarantuje automatycznej odporności na ataki, podważając hipotezę o bezpośrednim związku między złożonością danych a ich bezpieczeństwem.

Podobny wniosek można wysunąć na podstawie opracowania naukowego pt. „AI Data poisoning attack: Manipulating game AI of Go", autorstwa J. Shen, M. Xia, w którym autorzy dostarczają przykład ataku zatruwania danych AI, który skutecznie manipuluje systemem AI grającym w aplikację „Go" poprzez instalację trojańskiego wirusa. Udało się to mimo zastosowania wieloaspektowych danych w procesie uczenia, co pokazuje, że ataki mogą być przeprowadzone nawet na systemach wykorzystujących złożone i różnorodne zbiory danych. Przypadek ten rzuca światło na to, że wieloaspektowość danych sama w sobie nie wystarcza do zapewnienia pełnej odporności na złożone i celowe ataki adwersyjne, co kwestionuje przyjętą wcześniej hipotezę o automatycznym wzroście bezpieczeństwa wraz ze zwiększeniem różnorodności danych.

Akcesoryjnie eksperyment uwidocznił, że algorytmy detekcyjne bezsprzecznie mogą być wzmocnione za pomocą wdrożenia parametrów uczenia maszynowego, którego rudymentarność elementów polega na symulacjach behawioralnych, co prowadzi do innowacyjnego podejścia w spektrum kreowania systemów obronnych.

Innym ważnym wynikiem było zaobserwowanie zdolności systemów AI do identyfikacji wcześniej nieznanych ataków adwersyjnych, używających subtelnych zmian parametrów w danych. Owa zdolność do prognozowania oraz adaptacji na możliwe zagrożenia pokazała, że algorytmy potrafią kształtować rodzaj cyfrowej intuicji, rozszerzając horyzonty dla autonomicznych systemów zabezpieczeń sztucznej inteligencji. Nieoczekiwanym odkryciem było bowiem ujawnienie, że specyficzne ustawienia zmiennych mogą stanowić rolę naturalnych mechanizmów ochronnych i obronnych, redukując efektywność ataków adwersyjnych bez bezpośredniego przedsięwzięcia fakultatywnych środków bezpieczeństwa.

Niniejsze wyniki badań dostarczają wartościowych wniosków, otwierając jednocześnie nowe obszary dla dalszych badań, które mogą się przyczynić do rozwoju bardziej zaawansowanych i odpornych systemów sztucznej inteligencji. Przeprowadzone analizy wskazują na znaczący potencjał systemów AI w opracowywaniu nowych strategii obronnych przeciwko atakom adwersyjnym, co podkreśla możliwość tworzenia systemów zdolnych do proaktywnej adaptacji do nowych wyzwań. Te odkrycia zachęcają do dalszego eksplorowania możliwości tzw. intuicji cyfrowej jako kluczowego elementu w przyszłych systemach zabezpieczeń.

Generalizując, należy stwierdzić, że niniejsze badania wskazują na znaczny potencjał systemów AI do rozwoju niekonwencjonalnych strategii obronnych w obliczu ataków adwersyjnych. Rzutują na realne możliwości rozwoju systemów AI zdolnych

do adaptacji, które nie tylko reaktywnie odpierają znane zagrożenia, ale również proaktywnie adaptują się do nowych, nierozpoznanych wyzwań. Te wyniki sugerują wartość dalszego badania roli ‚intuicji cyfrowej' w kontekście przyszłych systemów zabezpieczeń.

4. Dyskusja

Przedstawione wyniki badań łączą dwa kluczowe obszary: eksplikację AI oraz kooperacyjną sztuczną inteligencję, które zostały omówione w opracowaniach H. Banieckiego i P. Biecka, a także T. Fujimoto i A.P. Pedersena. Wyniki badań wskazują na możliwość, że systemy AI mogą identyfikować enigmatyczne ataki adwersyjne, co może sugerować zdolność do adaptacyjnego rozpoznawania nowych zagrożeń. W kontraście do wspomnianych prac, które skupiały się na kompleksowym przeglądzie oraz zaciemnianiu kodu źródłowego, badanie autora rozszerza horyzonty w analizie tych dwóch prac, badając obronę i ochronę przed atakami adwersyjnymi, jak również przejrzystość AI. Dowodzi, że AI potrafi „intuicyjnie" rozpoznawać enigmatyczne niebezpieczeństwa, co sugeruje jej kompetencję do samodoskonalenia w dynamicznych środowiskach.

Wykracza to poza zakres konwencjonalnych technik obronnych, sygnalizując, że przyszłe systemy sztucznej inteligencji prawdopodobnie będą w stanie nie tylko działać reaktywnie, ale i proaktywnie, adaptując się do nowoczesnych wyzwań.

Od strony prawnej i etycznej wszelkie przedstawione dane miały jedynie charakter poglądowy z ogólnodostępnych map, a symulacje przeprowadzonych ataków adwersyjnych odbyły się w środowisku monitorowanym. Owa dedykacja dla zintegrowanej nauki z odpowiedzialnością zaznacza, jak ważne jest stosowanie się do zasad etycznych oraz prawnych w badaniach nad sztuczną inteligencją, aby technologie te mogły się przyczynić do pozytywnych celów i ochrony społeczeństwa. Wspomniane odkrycia otwierają nowe pola poszukiwań odpowiedzi dla zrozumienia oraz konstruowania systemów AI, które nie tylko są klarowne, ale także w naturalny sposób odpowiadają na ataki adwersyjne. Warto kontynuować badania nad systemami sztucznej inteligencji, zdolnymi do samoadaptacji oraz samodzielnej ewolucji w obliczu nowoczesnych wyzwań, zachowując jednocześnie zakres ustalonych ram prawnych i etycznych.

Konkludując, niniejsze badanie podkreśla potrzebę dalszych badań nad etycznymi i bezpiecznymi aspektami ataków adwersyjnych na AI, zwracając uwagę na znaczenie odpowiedzialnego projektowania systemów.

Podsumowanie

Omówione w tym opracowaniu badanie przedstawia nową perspektywę na złożoność oraz ewolucję obrony i ochrony w odpowiedzi na ataki adwersyjne w systemach sztucznej inteligencji, kładąc akcent na ich wykorzystanie w obszarze logistyki

humanitarnej. Dokonano analizy, w jaki sposób „intuicja cyfrowa" oraz elastyczność AI mogą implikować nieoczekiwane niebezpieczeństwa, dając odpowiedź na tak sformułowaną hipotezę. Owo odkrycie wskazuje, że zaawansowane systemy AI zaopatrzone w funkcjonalności samouczenia nie tylko oferują skuteczną obronę w odpowiedzi na ataki adwersyjne, ale również prezentują nowe horyzonty dla wzrostu bezpieczeństwa oraz wydajności w krytycznych obszarach, do których należy m.in. logistyka humanitarna. Uwydatniono także istotę aspektów prawnych i etycznych w spektrum badań nad AI, akcentując potrzebę bezustannej dyskusji i kooperacji międzydyscyplinarnej dla kreowania przyszłych technologii w sposób odpowiedzialny, etyczny i bezpieczny.

Przedstawione studium przypadku podkreśla podwaliny pod przyszłe badania, sugerując, że następne eksperymenty powinny zwrócić szczególną uwagę na rozwijanie zaawansowanych strategii detekcji, które staną na wysokości zadania w kontekście wciąż ewoluujących ataków adwersyjnych, utrzymując jednocześnie zgodność z przyjętymi standardami etyki. Owa pionierska praca motywuje do otwierania kolejnych „drzwi" w spektrum wykorzystania AI w obszarze humanitarnym, ze szczególnym naciskiem na kreowanie systemów autonomicznych oraz umiejętność adaptacji do zmieniających się warunków. Nadchodzące eksploracje mogą się także zająć synergią AI do dodatkowych technologii (np. blockchain, czyli cyfrowy łańcuch bloków, znacząco zintensyfikuje transparentność i bezpieczeństwo w kluczowych aplikacjach).

Na podstawie dokonanego badania eksperymentalnego, zawartych metod i analiz można stwierdzić, że odpowiedź na założoną hipotezę jest afirmatywna. Eksperyment dowiódł, że systemy AI stosowane w logistyce humanitarnej pozostają wrażliwe na wrogie ataki adwersyjne, które w konsekwencji mogą poważnie wpływać na ich wiarygodność i funkcjonowanie. Eksperyment ujawnił także, że za sprawą odpowiedniej implementacji metod i środków obronnych możliwa staje się znaczna redukcja tej podatności. Implementacja nowoczesnych systemów wykrywania i eliminacji ataków adwersyjnych, a zwłaszcza zastosowanie „intuicji cyfrowej" do wykrywania nieznanych zagrożeń uwidacznia, że redukcja inklinacji systemów AI na tego typu ataki bywa osiągalna. W obliczu niniejszych odkryć, niczym latarnia nawigacyjna rozjaśniająca nieznane morza, praca stanowi drogowskaz do drogi „bezpieczniejszego jutra" w spektrum logistyki humanitarnej, w których systemy sztucznej inteligencji, kształtowane przez innowacyjność i erudycję, odgrywają rolę nieugiętych strażników w obronie ludzkości.

Badania pokazały, że systemy AI mają potencjał do zarządzania ryzykiem wykraczającym poza obecne modele. Mogą teoretycznie reagować na kompleksowe zagrożenia w logistyce, w tym na te niewidoczne na pierwszy rzut oka. Eksperyment wyróżnia również możliwość zastosowania AI do monitorowania komunikacji w celu szybkiego identyfikowania kryzysów. Owe wnioski nakreślają przyszłość, w której AI nie tylko reaguje, ale także przewiduje i kształtuje odpowiedź na humanitarne potrzeby, stając się proaktywnym elementem zarządzania kryzysowego. W ten sposób, w pełni uświadamiając sobie potencjał tkwiący w sztucznej inteligencji, stajemy na

progu nowej ery, gdzie granice między technologią a humanitaryzmem mocno się zacierają, dając obietnicę świtu, w którym maszyny nie tylko wspierają, ale i współtworzą lepszą przyszłość dla całej ludzkości.

Mikołaj Kabała

Autor specjalizuje się w dziedzinie sztucznej inteligencji i matematyki dyskretnej, aktywnie angażuje się w badania naukowe oraz edukację. Pisze teksty naukowe i tworzy kursy on-line, dążąc do upowszechniania wiedzy i nowatorskich rozwiązań technologicznych. Otrzymał stypendium rektora Uniwersytetu Jana Kochanowskiego w Kielcach (Wydział Prawa i Nauk Społecznych) za wysokie wyniki w nauce (średnia ocen na poziomie 5,00)

Bibliografia

Baniecki, H. i Biecek, P. (2023). *Adversarial attacks and defenses in explainable artificial intelligence: A survey*. Pobrane z: https://doi.org/10.48550/arXiv.2306.06123.

Fujimoto, T., Pedersen, A.P. (2021). *Adversarial Attacks in Cooperative AI*. Pobrane z: https://doi.org/10.48550/arXiv.2111.14833.

Li, X., Pan, D., Zhu, D. (2020). *Defending against adversarial attacks on medical imaging AI system, classification or detection?* Pobrane z: https://doi.org/10.48550/ arXiv.2006.13555.

Lutz, M. (2015). *Learning Python: Powerful Object-Oriented Programming.* Kalifornia: O'Reilly Media.

Miasto Jarosław. (2024). Pobrane z lokalizacji: https://powiat.jaroslawski.pl/gminy/item/24-miasto-jaroslaw.

Sellakumar, S., Shiran, T., Blake, C., Maloney, R., McAllister, J., Bhojani, R. i Winkler, E. (2024). *Adversarial Attacks in AI*. Pobrane z: https://www.dremio.com/ wiki/adversarial-attacks-in-ai/.

Shen, J., Xia, M. (2020). *AI Data poisoning attack: Manipulating game AI of Go*. Pobrane z: https://doi.org/10.48550/arXiv.2007.11820.