Wymiar biznesowy ataków na systemy uczące się. Cz. 2. Zagrożenia związane z wykorzystaniem systemów uczących się w RPA

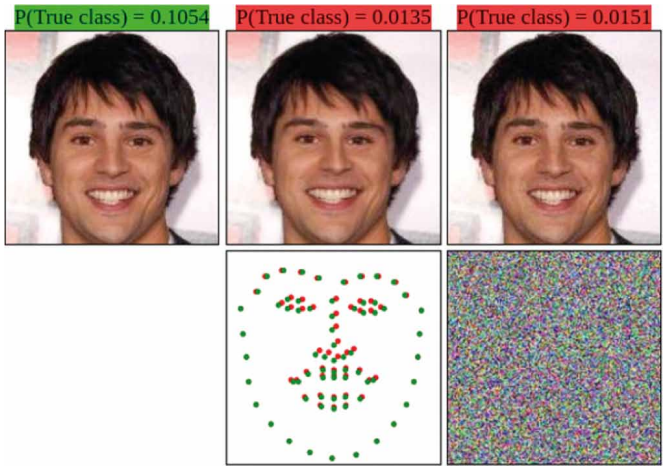
Mariusz Rafało

1. Wprowadzenie

Jak wskazano wcześniej, systemy uczące się, opierając się na danych uczących, dokonują generalizacji i identyfikacji reguł. Ta cecha modeli, czyniąca je możliwymi do wykorzystania na innych zbiorach danych, sprawia także, że każdy model jest niedoskonały. Ogólne reguły klasyfikacji powodują, że możliwe jest przygotowanie danych, które nieznacznie różnią się od danych oryginalnych, natomiast są odmiennie rozpoznawane przez model. Takie dane mogą być naturalnymi anomaliami, mogą także być próbkami intencjonalnie przygotowanymi, aby przeprowadzić atak na model. Atakujący prezentuje modelowi dane, które ten sklasyfikuje do innej grupy niż ta, do której rzeczywiście należą. Działanie to może mieć na celu zablokowanie pracy systemu lub wymuszenie błędnego działania systemu, w tym m.in. reakcji systemu na ściśle określone dane wejściowe zgodnie z intencjami atakującego.

Najczęściej obecnie przytaczane ataki na AI dotyczą rozpoznawania obrazów. Przykładowo: wykazano, że możliwe jest wprowadzenie drobnych zmian w poprawnie klasyfikowanym obrazie, co spowoduje, że obraz otrzyma całkowicie inną etykietę. Modyfikacja (tzw. perturbacja) obrazu obejmuje niewielką zmianę nasycenia wybranych kolorów (tzw. gradient), która jest trudna do odróżnienia dla ludzkiego oka (Szegedy i in., 2014). Ataki tego typu są spektakularne: niewielkie modyfikacje obrazu mogą powodować błędną klasyfikację znaków drogowych (Papernot, Mcdaniel i Goodfellow, 2017), odręcznego pisma (Papernot i in., 2017) czy twarzy (Dabouei i in., 2019).

W przypadku systemów rozpoznawania twarzy możliwe jest wprowadzanie perturbacji, która nie dotyczy modyfikacji koloru pikseli, ale rozmieszczenia kluczowych cech twarzy. Podejście to bazuje na rozmieszczeniu oczu, ust, brwi i nosa na twarzy. Okazuje się, iż niewielkie (praktycznie niedostrzegalne) przesunięcia wybranych elementów sprawiają, że twarz przestaje być poprawnie klasyfikowana (Dabouei i in., 2019). Wadą podejścia opartego na gradiencie jest modyfikacja obrazu polegająca na rozmyciu kolorów lub pogorszeniu ostrości – może to być zauważone gołym okiem. Wspomniane podejście, oparte na transformacji przestrzennej (ang. *spatial transformation*), jest pozbawione tej wady: obraz zachowuje kolory i ostrość, różni się jedynie położeniem elementów twarzy. Rysunek 1 prezentuje oba podejścia do perturbacji obrazu: kolumna pierwsza zawiera obraz oryginalny (poprawnie klasyfikowany), kolumna druga zawiera obraz ze zmodyfikowanym rozmieszczeniem oczu, zaś kolumna trzecia to obraz zmodyfikowany gradientowo (za pomocą nasycenia kolorów).



Rys. 1. Porównanie metod ataku na system identyfikacji tożsamości: metoda transformacji przestrzennej (kolumna 2) oraz metoda gradientowa (kolumna 3)

Źródło: Dabouei, A. i in. (2019). Fast geometrically-perturbed adversarial faces. Proceedings – 2019. IEEE Winter Conference on Applications of Computer Vision, WACV 2019, 1979–1988, https://doi.org/10.1109/WACV.2019.00215

Metoda transformacji przestrzennej należy do grupy ataków typu *white box*, tzn. atakujący ma wiedzę o działaniu klasyfikatora (modelu) i ma dostęp do jego parametrów. Może ona zostać wykorzystana przez potencjalnych atakujących do zmylenia systemów identyfikacji twarzy lub (rozszerzając zastosowanie) innych systemów służących identyfikacji obrazu. Jej zastosowanie nie wymaga bowiem „rozmywania obrazu", co czyni ją trudniejszą do wykrycia. Zagrożone atakiem są szczególnie systemy kontroli dostępu, weryfikacji tożsamości czy monitorowania obecności. Nie są to systemy wprost realizujące podstawowe procesy biznesowe. Jednak ich rola w zapewnieniu ciągłości i bezpieczeństwa pracy, jako procesów pomocniczych, jest kluczowa.

2. Geneza ataków na systemy uczące się

Większość ataków na systemy sztucznej inteligencji opiera się na sztucznie przygotowanych próbkach danych, które przekazane do modelu powodują jego błędne klasyfikacje. Genezą tworzenia sztucznych danych jest problem z wyjaśnianiem decyzji modelu. Algorytmy uczenia maszynowego, zwłaszcza te oparte na sieciach neuronowych, są zwykle trudne w interpretacji. Oznacza to, że trudno jest odpowiedzieć na pytanie, dlaczego model ocenił dane w określony sposób. Opierając się jedynie na wyniku klasyfikacji, zazwyczaj trudno jest ustalić, co spowodowało taką decyzję, i podać jej sensowne uzasadnienie. Aby rozwiązać ten problem, stosuje się metody alternatywnych wyjaśnień, które zamiast wyjaśniać, dlaczego model dokonał określonej klasyfikacji, wyjaśniają, w jaki sposób można osiągnąć inny wynik (Moore, Hammerla i Watkins, 2019).

Do generowania sztucznych danych, na potrzeby wyjaśniania predykcji modeli, stosuje się dwie główne kategorie systemów uczenia maszynowego (Moore i in., 2019):

- Algorytm LIME (Ribeiro, Singh i Guestri, 2016): pobiera dane wejściowe i tworzy ich różne wersje przez zerowanie różnych atrybutów, a następnie buduje lokalny model liniowy, ważąc dane wejściowe na podstawie odległości od oryginału. Rezultatem jest możliwy do wyjaśnienia model liniowy, w którym współczynniki modelu działają jako wyjaśnienie i opisują udział każdego atrybutu w uzyskanej klasyfikacji.
- Algorytm SHAP (Lundberg i Lee, 2017): opiera się na teorii gier i poszukuje optymalnego rozwiązania przez system nagród i kar.

Obie metody, choć mają odmienne algorytmy, prezentują podobne wyniki: wskazują, które atrybuty przyczyniły się najbardziej do uzyskania określonej klasyfikacji. Ograniczeniem metod opartych na sztucznie generowanych próbkach jest to, że nie wskazują one przyczyn takiej czy innej klasyfikacji, a jedynie prezentują przykłady alternatywnych danych, które uzyskały inną klasyfikację. Przykładowo na podstawie sztucznie wygenerowanych próbek można stwierdzić, że dany klient banku nie otrzymał pożyczki ze względu na wynagrodzenie i wiek. Nie można natomiast stwierdzić, co klient musi zrobić, aby uzyskać pożyczkę w przyszłości (Moore i in., 2019). Jaki poziom dochodów gwarantuje pozytywną decyzję kredytową? Jaki wiek zwiększa szanse na uzyskanie kredytu? Na te pytania nie można udzielić jednoznacznej odpowiedzi. Moore i in. (2019) przytaczają przykład eksperymentu, w którym dla odmownej decyzji kredytowej wskazane zostały przykłady klientów o niewiele różniących się cechach, którzy otrzymali pozytywną decyzję kredytową. Na pytanie o to, dlaczego 27-letnia kobieta otrzymała odmowę udzielenia kredytu, a (sztucznie wygenerowany) 31-letni mężczyzna kredyt by otrzymał – nie znaleziono odpowiedzi.

Do generowania sztucznych danych stosuje się także takie techniki, jak generatywne sieci współzawodniczące (ang. *Generative Adversarial Nets* – GAN) (Goodfellow i in., 2014) czy SMOTE (ang. *Synthetic Minority Oversampling Technique*) (Chawla i in., 2002). Są to narzędzia powszechnie stosowane do testowania modeli uczenia maszynowego czy też do trenowania takich modeli, szczególnie w przypadku systemów służących identyfikacji anomalii, gdzie uzyskanie wysokiej liczby rzeczywistych przypadków anomalii jest trudne. Wówczas stosuje się techniki sztucznego generowania danych, oparte na niewielkiej próbie przypadków rzeczywistych. W efekcie uzyskuje się większą liczbę przypadków, które służą do uczenia modelu.

Rys. 2. System Deepfake imitujący wypowiedzi B. Obamy
Źródło: https://www.youtube.com/watch?v=cQ54GDm1eL0 (dostęp: 30.05.2020 r.)

Przytoczone narzędzia, zbudowane dla realizacji konkretnych potrzeb analitycznych, mogą być z powodzeniem wykorzystane do generowania próbek antagonistycznych (ang. *adversarial sample*) – służących „oszukaniu" modeli AI (Goodfellow i in., 2014). Uzyskane w ten sposób sztuczne dane są bardzo trudne do odróżnienia od rzeczywistych danych.

Próbki antagonistyczne znalazły także zastosowanie w opracowaniu techniki określanej mianem *deepfake*. Technika ta stosowana jest do łączenia i nakładania obrazów nieruchomych i ruchomych na obrazy lub filmy źródłowe i stosowania przy tym algorytmów AI. Uzyskane w tym procesie obrazy czy filmy są bardzo realistyczne, stwarzając możliwości manipulacji przez np. niemożliwą do odróżnienia przez widza zamianę twarzy aktorów występujących w filmie. Przykładowo badacze z Uniwersytetu w Waszyngtonie (Suwajanakorn, Seitz i Kemelmacher-Shlizerman, 2017) opracowali algorytm pozwalający na spreparowanie dowolnej wypowiedzi Baracka Obamy (rys. 2). Na wygenerowanym filmie autor wypowiada się, zaś obraz i dźwięk prezentowane są w formie wypowiedzi prezydenta Obamy (system dokonuje także syntezy głosu byłego prezydenta USA). Efektem jest film prezentujący wypowiedzi B. Obamy, które faktycznie nie miały miejsca.

Ta technika może służyć do oszustw, niemniej nie należy do domeny hakowania AI. Antagonistyczne uczenie maszynowe obejmuje działania, które mają na celu oszukanie sztucznej inteligencji. W przypadku Deepfake atakujący stosuje sztuczną inteligencję, aby oszukać inne osoby lub podmioty. Oba podejścia łączą: intencja oszustwa oraz stosowanie sztucznej inteligencji. Niektóre firmy wdrażają jednak specjalizowane oprogramowanie, które ma na celu identyfikować, czy dany obraz, film lub nagranie audio nie zostały spreparowane sztucznie (przez Deepfake). Te systemy z kolei stają się celem antagonistycznych ataków, które mają na celu przekonanie ich, że dany materiał jest prawdziwy, mimo że został wygenerowany komputerowo za pomocą Deepfake (Neekhara i in., 2019).

3. Przykłady realnych zagrożeń

3.1. Uwagi wstępne

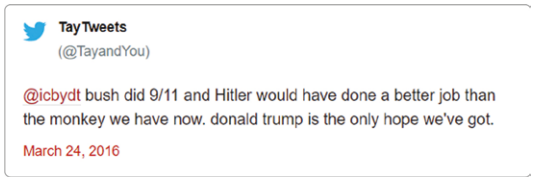
Zagrożenia dla systemów opartych na uczeniu maszynowym mogą wynikać z działań zamierzonych (ataków) lub przypadkowych anomalii. W obu przypadkach konsekwencją dla systemu może być przerwanie ciągłości procesu biznesowego. Ataki można sklasyfikować pod względem łatwości przeprowadzenia. Przykładowo ataki związane ze znakami drogowymi wymagają ingerencji atakującego w infrastrukturę fizyczną: musiałby podmienić albo zmodyfikować znaki stojące przy drogach. Są to ataki potencjalnie trudne do przeprowadzenia, jednak w czasie, gdy coraz więcej pojazdów ma aktywne wspieranie kierowcy lub w ogóle są autonomiczne, tego typu zagrożenia nie mogą zostać pominięte. Podobnie w przypadku systemów analizy tożsamości: atakujący musiałby dokonać zmian w swoim wyglądzie lub zmodyfikować fizycznie swój dokument tożsamości. Jest to dla atakującego zadaniem wymagającym, jednak, jeśli przeprowadzone skutecznie, stanowi poważne zagrożenie dla systemów identyfikacji tożsamości, monitorowania bezpieczeństwa czy identyfikowania osób poszukiwanych. Przestępcy atakujący np. systemy dokonujące transakcji finansowych mają jednak teoretycznie prostsze zadanie. Atakujący mają bowiem pełną możliwość kontrolowania danych, które są wejściem do modelu. Składając i anulując zlecenia zakupu, atakujący może wpływać na systemy, które podejmują automatyczne decyzje zakupowe na podstawie składanych zleceń (Goldblum i in., 2020).

3.2. Przykład ataku infekcyjnego

Przykładem skutecznego ataku na proces uczenia się systemu AI jest krótka historia funkcjonowania bota Tay, który komunikował się z użytkownikami mówiącymi po angielsku za pomocą profilu Twitter. Tay była botem, opracowanym przez Microsoft jako projekt badawczy, którego celem była implementacja sztucznej inteligencji zdolnej do prowadzenia samodzielnej konwersacji na portalu społecznościowym. W ciągu zaledwie kilku godzin interakcji z innymi osobami Tay „nauczyła się" rasistowskich wypowiedzi oraz wypowiadania się pochlebnie o Adolfie Hitlerze (rys. 3). Po 16 godzinach od uruchomienia Microsoft był zmuszony wyłączyć Tay (Hunt, 2016).

W przypadku bota Tay nauka odbywała się na wysoce „skrzywionej" próbie danych uczących. Rozmówcy bardzo szybko zorientowali się bowiem, że Tay jest botem i że uczy się podczas konwersacji. Grupa użytkowników Twittera zaczęła publikować nieprawdziwe lub niepoprawne politycznie tezy, które algorytm traktował jako dane uczące.

Biznesowy odpowiednik wadliwie nauczonego robota został wdrożony w firmie Amazon. System sztucznej inteligencji został zaprojektowany, by podejmować decyzje dotyczące rekrutacji nowych pracowników działów IT. Do Amazona spływają tysiące życiorysów programistów, analityków i projektantów, stąd system miał dokonywać wstępnego wyboru kandydatów. Wybrane pojedyncze osoby były następnie kierowane do kolejnych etapów rekrutacji. Szybko okazało się, że system całkowicie dyskryminuje kobiety i do zatrudnienia rekomenduje wyłącznie mężczyzn. Nie znajdowało to uzasadnienia, ponieważ do pracy aplikowały także kobiety o odpowiednich kwalifikacjach. W tym przypadku wadliwe uczenie modelu odbyło się bez intencji ataku. Dostarczony do modelu zbiór danych uczących obejmował dane z 10 lat, zaś w tym okresie na rynku IT oraz na uczelniach technicznych dominowali mężczyźni. Wobec tak określonych danych wejściowych system skutecznie eliminował życiorysy o cechach kobiet (a robił to rzeczywiście „inteligentnie", ponieważ wszystkie CV były anonimowe) (Dastin, 2018).



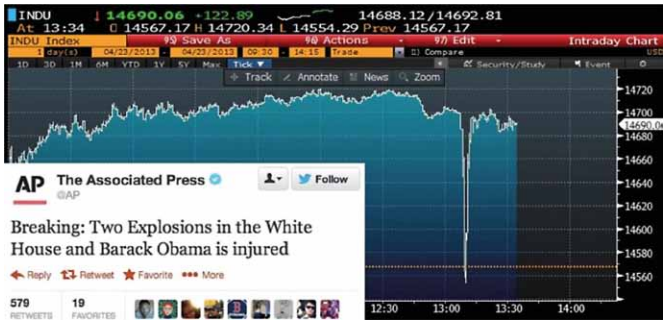
Rys. 3. Jeden z komunikatów bota Tay, publikowany przez AI na Twitterze
Źródło: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbotgets-a-crash-course-in-racism-from-twitter (dostęp: 24.05.2020 r.).

3.3. Atak na automatyczny system transakcji finansowych

Współczesne rynki kapitałowe opierają się na zaawansowanych systemach informatycznych. Wszystkie transakcje są realizowane elektronicznie, a informacje rejestrowane są w bazach danych. Decydenci, którzy podejmują decyzje inwestycyjne, są wspierani przez specjalizowane systemy, które z jednej strony automatyzują pewne działania, z drugiej zaś wspierają podejmowanie decyzji. Decyzje mogą być wspierane pasywnie – przez wskazywanie optymalnych kompozycji portfela – lub aktywnie – przez realizowanie tych akcji. Zwłaszcza transakcje krótkoterminowe na hurtowych rynkach walutowych (*Forex*) obsługiwane są za pomocą robotów, które mają dość dużą swobodę działania. Wysoki i wciąż rosnący poziom autonomii robotów sprawia, że ataki dokonane na te roboty mogą przynieść atakującym wymierne korzyści. Jeśli atak na działający system uczący się jest w stanie spowodować określone akcje na robotach, to można przewidzieć skutki tych akcji. Atakujący może zatem dysponować wiedzą o zachowaniu rynku w przyszłości, a to przekłada się już na konkretne korzyści finansowe.

Przykładem takiego zagrożenia może być seria cyberataków przeprowadzona w okresie od kwietniu do maja 2013 roku. Celem ataków były serwisy informacyjne w Syrii, Europie i USA; szczególnie strony WWW oraz konta w mediach społecznościowych. Do ataków przyznała się grupa przestępcza używająca nazwy Syrian Electronic Army, popierająca syryjskiego przywódcę Baszara al-Assada. Ataki miały na celu zdyskredytować media i podważyć ich wiarygodność. Część ataków kierowana była na całkowite zablokowanie strony WWW, a część służyła nawet blokowaniu dostępu do internetu na terenie Syrii (Mandel, 2017).

W trakcie jednego z ataków, 23 kwietnia 2013 r., atakujący umieścili na profilu Twitter agencji Associated Press wiadomość o rzekomym ataku terrorystycznym na Biały Dom i rannym prezydencie Obamie (rys. 4). Rynki finansowe zareagowały błyskawicznie: tweet został opublikowany o 13:08, a już minutę później wskaźnik Dow Jones odnotował spadek o 150 punktów, by powrócić do pierwotnej wartości o 13:13 (po ogłoszeniu, że



Rys. 4. Reakcja indeksu Dow Jones na publikację wiadomości o zamachu na Biały Dom

Źródło: https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackersclaim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/ (dostęp: 26.05.2020 r.)

opublikowana informacja jest nieprawdziwa). Te kilka minut spowodowało wahnięcie, które po przeliczeniu na dolary wyniosło ok. 136 mld (Fisher, 2013).

Do ataku na konto Associated Press doszło przez atak typu *phishing*. Atakujący wysłali spreparowane maile do pracowników agencji prasowej. Maile zawierały informację o interesującym artykule i zachęcały do kliknięcia i zalogowania się. Co ciekawe, próby ataku zostały zidentyfikowane wcześniej i administratorzy Associated Press ostrzegali pracowników, aby nie otwierali podejrzanych maili (Perez, 2013). Mimo tych ostrzeżeń w trakcie ataku wyłudzono jednak informacje, pozwalające na zalogowanie się na konto Twitter i chwilowe przejęcie nad nim kontroli. Po przejęciu kontroli nad kontem atakujący spreparowali alarmujący komunikat (twitt) o nieprawdziwej treści. Ten komunikat został poddany maszynowej analizie, z wykorzystaniem metod *text mining*, przez systemy automatycznie inwestujące na giełdzie i w efekcie doszło do drastycznego wahnięcia wskaźnika Dow Jones.

Wskazany przykład podkreśla przede wszystkim aspekty biznesowe i finansowe ataku, jednak jego skutki miały także wymiar polityczny. Rola ataku (a właściwie ataków) była bowiem także istotna w destabilizacji układu politycznego w regionie Syrii (Mandel, 2017). Podobna sytuacja zaszła w Afryce Środkowej, gdzie w 2018 r. jako powód nieudanego zamachu stanu przez wojsko gabońskie wskazuje się spreparowane metodą Deepfake wystąpienia prezydenta Gabonu Ali Bongo (Westerlund, 2019).

We współczesnym przekazie informacji następuje zatarcie granicy między prawdą a fikcją. Dotyczy to szczególnie świata cyfrowego, gdzie dość łatwo można opublikować zmodyfikowane obrazy, filmy czy wiadomości. Szybkość przepływu informacji jest już dziś bardzo duża i wciąż rośnie. Dodatkowo coraz więcej systemów stale monitoruje aktywność polityków czy celebrytów w mediach społecznościowych, co znacznie zwiększa ryzyko, że ktoś omyłkowo opublikuje niebezpieczne dane lub padnie ofiarą ataku hakerskiego.

Szczególnie systemy finansowe, które cechują się wysokim poziomem automatyzacji, są podatne na takie zdarzenia. Systemy analizy treści i analizy sentymentu stale monitorują przestrzeń elektroniczną w poszukiwaniu zdarzeń, które mogą wpłynąć na kursy akcji czy walut. Sposobem na redukcję ryzyka jest opieranie się na wiarygodnych źródłach informacji. To w znacznym stopniu zabezpiecza przed przedostaniem się „sztucznie wygenerowanego" *fake news* do systemu sztucznej inteligencji, aczkolwiek, jak wskazano wcześniej, nie daje 100% gwarancji wiarygodności.

3.4. Ataki na systemy rekomendacyjne

Obszar potencjalnych i rzeczywistych zagrożeń dla systemów AI stanowią także powszechnie stosowane systemy rekomendacyjne. Celem systemu rekomendacyjnego jest zaproponowanie klientowi produktu, który z najwyższym prawdopodobieństwem go zainteresuje. Systemy te pracują głównie w internetowym kanale sprzedaży, gdzie każdy użytkownik może otrzymać spersonalizowaną ofertę sklepu internetowego czy usługodawcy. W kontekście zastosowanego algorytmu istnieją dwa sposoby funkcjonowania systemów rekomendacyjnych:

- oparte na regułach asocjacyjnych (ang. *association rules*) – systemy tej klasy ignorują tożsamość klienta, koncentrując się na współwystępowaniu produktów w koszyku klienta (paragonie). Systemy te noszą nazwę analiz koszykowych (ang. *market basket analysis*), ponieważ badają zawartość koszyków klientów, w poszukiwaniu produktów, które są kupowane łącznie;
- oparte na zachowaniach klientów i ich podobieństwie – systemy tej klasy, oparte głównie na algorytmie *collaborative filtering*, bazują na informacjach o aktywnościach klientów oraz na ocenach i opiniach o produktach, wystawianych przez innych im podobnych klientów.

Szczególnie zagrożone są systemy oparte na rankingach i opiniach klientów (*collaborative filtering*). Atakujący mogą bowiem manipulować treścią i częstotliwością rekomendacji produktów, stosując fałszywe profile użytkowników (klientów). W tej domenie można wyróżnić dwa rodzaje zagrożeń.

Pierwsze zagrożenie dotyczy generowania fikcyjnych ocen produktów, aby były one częściej proponowane klientom. Budowa algorytmu *collaborative filtering* sprawia, że jest on podatny na tego typu ataki, nazywane *shilling attacks* (Deldjoo, Di Noia i Merra, 2020). Atak typu *shilling* opiera się na fałszywych ocenach produktów, które są generowane automatycznie (Zhou i in., 2018). Efektem tych działań są nieprawdziwe wysokie oceny produktów lub pochlebne opinie o tych produktach. Systemy zabezpieczające przed takimi zdarzeniami opierają się głównie na analizie anomalii (aby zidentyfikować fałszywe oceny) lub na analizie profili (aby wyłapać fałszywe profile użytkowników).

Drugi rodzaj ataków na systemy rekomendacyjne ma charakter bardziej ogólny i dotyczy budowania fałszywych profili użytkowników. Atakujący wystawiają opinie o firmach lub produktach, posługując się fałszywymi kontami klientów (Bhaumik i in., 2006). Profile te można wykorzystać w atakach na systemy rekomendacyjne, ale także w atakach na systemy analizy sentymentu czy podczas oceny ryzyka kredytowego. Fikcyjne osobowości mogą zostać uwiarygodnione przez generowanie fikcyjnych działań czy przez publikowanie zdjęć zawierających nieistniejące osoby (rys. 5). Podejście to, zwłaszcza połączone z atakiem typu *shilling*, jest szczególnie trudne do wykrycia (Bhaumik i in., 2006).

Rys. 5. Fikcyjne fotografie ludzi, sztucznie wygenerowane, przy zastosowaniu techniki GAN. Utworzone w ten sposób dane są praktycznie nieodróżnialne od rzeczywistych

Źródło: https://thispersondoesnotexist.com/ (dostęp: 27.05.2020 r.)

3.5. Inne zagrożenia

Dotychczasowe klasyfikacje zagrożeń wynikających z antagonistycznego uczenia maszynowego opierają się głównie na dwóch kategoriach: na czasie, w którym atak został wykonany (infekcyjny, inwazyjny lub atak na klasyfikator), lub na poziomie wiedzy dostępnej dla atakującego (*black box* lub *white box*). Można także dokonać klasyfikacji wybranych zagrożeń na podstawie procesów biznesowych będących celem ataku lub według stosowanych w nich technikach AI (tabela 1).

4. Zakończenie

W tym rozdziale zaprezentowane zostały metody i rodzaje zagrożeń dla działalności biznesowej wynikających z ataków na systemy uczące się. Z przeprowadzonej analizy płyną dwa wnioski. Po pierwsze, ataki tego typu mogą w istotny sposób zaburzyć funkcjonowanie procesów biznesowych. Procesy biznesowe wspierane sztuczną inteligencją mogą zostać zmuszone do niepoprawnego działania. Ryzyko jest szczególnie wysokie w przypadku systemów, które mają wysoki poziom autonomii.

Po drugie, organizacje raczej nie uwzględniają specyfiki ataków na AI podczas zarządzania ryzykiem. Świadomość tych zagrożeń istnieje, jednak problemem jest brak narzędzi, które pomagałyby ograniczać ryzyko na etapie budowania i operacjonalizacji modeli AI (Kumar i in., 2020). W domenie sztucznej inteligencji istnieją jedynie zbiory dobrych praktyk i wskazówek, które mają na celu uchronić kod przed potencjalnymi lukami. Innych zabezpieczeń w zasadzie nie ma, choć specjaliści wskazują na konieczność uwzględniania sztucznie wygenerowanych „złośliwych" danych podczas uczenia modeli. Chodzi o to, aby modele były wyczulone na jak najwięcej tego typu przypadków (Dai i in., 2018).

Tabela 1. Zagrożenia wynikające z antagonistycznego uczenia maszynowego

Biznesowe zastosowanie AI	Przykłady zagrożeń
Identyfikacja nadużyć	Manipulowanie danymi, aby ukryć nielegalną działalność, związaną przykładowo z nadużyciami finansowymi lub praniem brudnych pieniędzy. Generowanie próbek antagonistycznych służy w tym przypadku dwóm celom: zastąpieniu podejrzanej transakcji inną transakcją (wygenerowaną sztucznie) lub obudowaniu nadużycia innymi transakcjami (także sztucznymi), aby nadużycie nie było traktowane jak anomalia (Schreyer i in., 2019).
Bezpieczeństwo danych	Ukrywanie faktu kradzieży danych z systemów informatycznych. Systemy identyfikacji nadużyć wykrywają działania pracowników, które odbiegają od normy (np. uruchamianie kilkadziesiąt razy tego samego raportu zawierającego dane klientów, podczas gdy inni pracownicy uruchamiają go średnio raz w tygodniu). Atak polega na przygotowaniu robota programowego, aby wykonywał on działania symulujące pracownika, jednak prowadzące do pozyskania jak największej ilości danych.
Zarządzanie portfelem inwestycyjnym	Wprowadzenie w błąd systemów realizujących automatyczne transakcje finansowe przez wykorzystanie luk w regułach działania tych systemów. Generowanie dużej liczby transakcji powodujące, że systemy zaczynają je intepretować według zaimplementowanych reguł, co może prowadzić do zmian w kursach akcji lub walut. Przykładowo w 2015 r. rosyjscy hakerzy dokonali ataku na sektor finansowy, wykorzystując tę właściwość robotów. Hakerzy wykorzystali złośliwe oprogramowanie, aby na krótko zdestabilizować kurs wymiany rubla do dolara (Hacker News, 2016).
Symulacje finansowe	Wprowadzenie fałszywych danych transakcyjnych do uczącego zbioru danych, aby wprowadzić w błąd systemy symulacyjne. Atakujący może w ten sposób wpłynąć na parametry opracowanego modelu symulacyjnego. Modele te są regularnie szkolone, aby uwzględnić nowsze dane, co czyni je podatnymi na tego typu ataki (Cantos, 2019).
Zarządzanie ryzykiem kredytowym	Wprowadzenie w błąd systemu oceny ryzyka kredytowego przez prezentowanie spreparowanych lub zmodyfikowanych danych. Taki system może błędnie oszacować ryzyko kredytowe i sprawić, że bank podejmie niepożądane działania i np. udzieli kredytu podmiotowi niewypłacalnemu.

Źródło: opracowanie własne

Kontekst biznesowy ataków na systemy maszynowego uczenia się nie ogranicza się jednak do robotyzacji i automatyzacji procesów biznesowych. Obrana w tym rozdziale perspektywa ma charakter procesowy i pokazuje wiele aspektów funkcjonowania przedsiębiorstw, takich jak marketing, operacje, sprzedaż czy finanse. Dalsze rozważania związane z tego typu zagrożeniami powinny jednak objąć całość procesów biznesowych – od zakupów po sprzedaż.

Odmienny obszar potencjalnych zagrożeń stanowią szeroko pojęte zastosowania Internetu Rzeczy, szczególnie w dobie możliwości sieci 5G. Czujniki gromadzące dane na potrzeby inteligentnych samochodów, domów, miast czy inteligentnej produkcji, a także modele wykorzystujące dane z tych czujników też mogą stać się celem ataków przy wykorzystaniu antagonistycznych próbek danych. ■

Bibliografia dostępna pod linkiem: nis.com.pl/bibliografia.html

Fragment pochodzi z książki: *Hakowanie sztucznej inteligencji*, Jerzy Surma (redakcja naukowa), Wydawnictwo Naukowe PWN, Warszawa 2020