Hakowanie sztucznej inteligencji

Mariusz Rafało

3.1. WPROWADZENIE

W dzisiejszych czasach większość procesów biznesowych jest zinformatyzowana, a wiele z nich jest realizowanych wyłącznie w świecie cyfrowym. Informatyzacja sprawia, że poszczególne kroki procesu, generowane dane czy inne informacje, są rejestrowane w bazach danych. Dane rejestrowane w ten sposób określa się mianem cyfrowego śladu. Służą one monitorowaniu procesu i analizowaniu aktywności uczestników procesu (Surma, 2017). Przykładowo: w bazach danych dostępne są informacje o złożonych zamówieniach, wystawionych fakturach, dostępnych produktach, zalogowanych klientach itd. Rejestrowane są wszelkie działania i aktywności klientów wykonywane na stronie WWW, aplikacjach mobilnych, a także, coraz częściej, w sklepach i punktach obsługi klienta. Dane mogą być również zbierane bez aktywnego udziału klienta – np. za pomocą czujników RFID, geolokalizacji GSM, GPS czy lokalizacji wi-fi. Niezależnie od branży cyfrowe ślady interakcji firmy z klientami są rejestrowane w systemach informatycznych. Przy czym, coraz częściej dzieje się to w czasie niemal rzeczywistym, co oznacza przykładowo, że jeśli klient złożył reklamację w placówce firmy, to niemal w tej samej chwili informacja o tym jest dostępna w systemie CRM, zatem gdy klient zadzwoni na infolinię tej firmy, zostanie obsłużony z wykorzystaniem najbardziej aktualnej wiedzy.

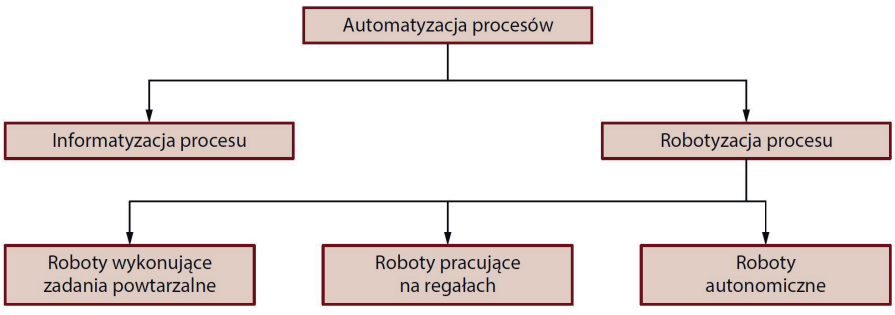
Dane pochodzące z cyfrowego śladu mogą być wykorzystane dwojako. Po pierwsze, mogą służyć bieżącemu wsparciu procesów biznesowych. Po drugie, dane można wykorzystywać w celach analitycznych. Służą temu głównie dane historyczne, które można wykorzystać jako zbiór uczący dla systemów uczących się czy, ogólnie mówiąc, systemów sztucznej inteligencji (AI). Systemy AI, mające „wiedzę" na temat historycznych transakcji, pozwalają nie tylko na automatyzację działań operacyjnych, lecz także na wspieranie w podejmowaniu decyzji. Algorytmy sztucznej inteligencji opierają się na dobrze ugruntowanych zasadach matematyki, statystki i ekonometrii. W wielu przypadkach złożoność algorytmów i systemów AI jest jednak tak duża, że są one raczej postrzegane jako „czarne skrzynki", które realizują konkretne działania, w sposób nie zawsze zrozumiały dla użytkownika. Ta złożoność, niedostępna dla ludzkiej percepcji, może być wykorzystana do potencjalnego „oszukania" systemu AI. Celem takich działań może być znalezienie luki w „czarnych skrzynkach" modeli sztucznej inteligencji. Można tego dokonać, stosując zaawansowane systemy analizy danych lub przez wprowadzanie do modelu specjalnie spreparowanych danych. Zaatakowany w ten sposób proces może zadziałać niepoprawnie (np. przyznać kredyt osobie, która nie powinna go otrzymać, zignorować transakcję, która jest oszustwem, itp.) lub może całkowicie przestać działać.

Celem autora tego rozdziału jest przedstawienie zagrożeń wynikających z zastosowania autonomicznych systemów sztucznej inteligencji (robotów programowych), które wspierają lub całkowicie realizują procesy biznesowe. Niezawodność i ciągłość procesów biznesowych stanowią nie tylko o efektywności działania firmy, ale niejednokrotnie o jej działaniu w ogóle. Przykładami mogą być tu sklep internetowy, który traci możliwość finalizacji składanych zamówień, lub bank, który utracił zdolność rozpoznawania ryzykownych kredytobiorców. Tego typu zagrożenia dla systemów uczących się wymagają zastosowania odpowiednich metod zarządzania ryzykiem.

3.2. ROBOTYZACJA I AUTOMATYZACJA PROCESÓW BIZNESOWYCH

Proces biznesowy to uporządkowany zestaw działań, które są określone, mierzalne i prowadzą do uzyskania konkretnego rezultatu. Może on obejmować krótkie sekwencje działań (np. wystawienie faktury, przyjęcie reklamacji) lub bardziej złożone aktywności (np. realizacja zamówienia online, z możliwością odbioru produktu w wybranej lokalizacji). Ujęcie procesowe pozwala zarządzającym na kontrolowanie i monitorowanie działań firmy, zgodnie z prowadzonymi działaniami sprzedażowymi, marketingowymi czy obsługą klienta.

Automatyzacja procesu biznesowego polega na realizacji tego procesu (lub jego fragmentu) za pomocą technologii, bez udziału pracownika lub przy jego minimalnym udziale (zazwyczaj ograniczającym się do nadzoru). Obecnie większość procesów biznesowych jest wspierana przez informatyczne systemy zarządzania lub jest realizowana w całości w świecie cyfrowym. Takie procesy, jak elektroniczny obieg dokumentów, sprzedaż w kanałach elektronicznych czy śledzenie przesyłek kurierskich online, to przykłady procesów zautomatyzowanych.



Rys. 3.1. Klasyfikacja automatyzacji procesów biznesowych. Źródło: opracowanie własne

Jeśli natomiast automatyzacja jest realizowana za pomocą oprogramowania, które wykonuje określone, często powtarzalne zadania, to można mówić o robotyzacji tego procesu. Robotyzacja jest rodzajem automatyzacji, który polega na tym, że proces (lub jego fragment) jest realizowany przez system informatyczny, tj. robota.

3.2.1. Robotyzacja procesów

Roboty kojarzą się powszechnie z automatami, które wykonują określone zadania w fabrykach, na liniach produkcyjnych lub halach montażowych (pomijając stereotypowy wizerunek robota w literaturze i kinie). Roboty, które wspierają lub realizują procesy biznesowe, są najczęściej specjalizowanymi systemami komputerowymi.

Określanie tych programów mianem robotów stanowi metaforę; roboty programowe realizują zadania w świecie cyfrowym, analizują dane i wykonują określone akcje.

Robotyzacja procesów biznesowych (ang. *Robotic Process Automation – RPA*) to jeden z rodzajów automatyzacji procesów, który jest realizowany za pomocą robota programowego (Sobczak, 2020a). Założeniem robotyzacji procesu jest zastąpienie powtarzalnych zadań wykonywane przez pracowników specjalistycznym oprogramowaniem (Sobczak, 2018). Robotyzacja zwiększa efektywność realizowanego procesu biznesowego w kilku wymiarach. Po pierwsze, w ujęciu wydajności pracy: robot może przetworzyć i zweryfikować setki dokumentów, wprowadzić tysiące informacji do systemu lub zweryfikować jakość danych pomiędzy systemami. Drugi wymiar dotyczy jakości wykonywanych zadań: przy powtarzalnych zadaniach o określonej strukturze robot raczej nie popełnia błędów. Trzeci aspekt to ciągłość pracy: robot może pracować bez przerwy, a jeśli zajdzie taka potrzeba, może zostać zeskalowany, aby pracować równolegle.

Popularnym przykładem robotów programowych, opartych na regułach, są chatboty i voiceboty. Pierwsza kategoria służy do obsługi klienta za pomocą czatu, najczęściej na stronie internetowej, druga służy obsłudze klienta na linii telefonicznej. W obu przypadkach robot obsługuje komunikację z klientem i realizuje wybrane zadania, wynikające z informacji otrzymanych od klienta.

Podstawowe ujęcie robotyzacji dotyczy zatem realizacji większej liczby operacji oraz wykonywania ich w sposób nieprzerwany i bezbłędnie. Inne, nieco mniej oczywiste zastosowania robotów obejmują nie tyle automatyzację działań, ile ich samodzielne inicjowanie. Przykładowo, robot programowy, mający autonomię, może ocenić zdolność kredytową klienta banku, może zweryfikować online poprawność transakcji internetowych, w ciągu dziesiętnych części sekundy, czy też zaproponować maklerowi najbardziej optymalną decyzję zakupową. Roboty autonomiczne działają na nieco innych zasadach niż roboty „klasyczne", oparte na powtarzalnych regułach.

Systemy klasy RPA służą do budowania robotów programowych, uruchamiania ich w środowisku procesu biznesowego oraz do sterowania robotami. Możliwości automatycznej realizacji zadań przez roboty programowe pochodzą ze zdefiniowanych reguł postępowania w konkretnych przypadkach. System działa według określonych sekwencji i wykonuje określone akcje. Przykładowo, mogą to być akcje ekranowe (wprowadzenie informacji, nawigowanie po ekranie itp.), akcje związane z komunikacją z innymi systemami (np. połączenie do bazy danych czy uruchomienie innego systemu) itp. Wraz ze wzrostem poziomu autonomii robota programowego komponenty, które odpowiadają za podejmowanie decyzji, wykorzystują coraz bardziej zaawansowane techniki analiz danych. W zależności od posiadanej autonomii systemy programowe można podzielić na (rys. 3.1.):

- roboty, które wspierają działania rutynowe: weryfikacja dokumentów, automatyczne skanowanie dokumentów PDF, odczytywanie danych z jednego systemu i wpisywanie ich do innego itp.;
- roboty oparte na regułach: systemy działają wykonując akcje, które są wyzwalane przez określone zdarzenia czy dane, np. obsługa klienta za pomocą chatbota;
- roboty autonomiczne: mogą podejmować samodzielne decyzje, np.

reklama

w zakresie rekomendacji produktu klientowi lub w zakresie oceny zdolności kredytowej klienta.

Warto także wspomnieć, że istnieją zrobotyzowane procesy, które w ogóle nie mogłyby być realizowane, gdyby nie wykonywał ich robot. Przykładowo, proces spersonalizowanej rekomendacji produktów w kanale internetowym byłby niemożliwy do realizacji bez udziału specjalizowanego systemu[1]. Sam proces prezentowania określonej treści klientom online oczywiście istnieje, ale ma on niewiele wspólnego z rekomendacją personalizowaną, ponieważ każdy (lub niemal każdy) klient otrzymuje ten sam komunikat. Dopiero wprowadzenie algorytmów eksploracji danych pozwala na to, aby każdego (lub niemal każdego) klienta traktować indywidualnie i prezentować mu treści, które mogą być dla niego interesujące.

3.2.2. Sztuczna inteligencja w robotyzacji procesów

Zaawansowane techniki i technologie analizowania danych stały się w ostatnich latach bardziej dostępne dla organizacji. W sposób szczególny przyczynia się do tego rozwój oprogramowania otwartego (ang. *open source*), głównie w domenie big data oraz rozwój systemów oferowanych w chmurze (ang. *cloud computing*). Pokusa jest duża: zastosowanie AI pozwala bowiem na istotne optymalizacje i korzyści we wspieranych przez nie procesach biznesowych. Dla niektórych firm stanowi to o optymalizacji realizowanych procesów, dla innych stanowi podstawowy czynnik przewagi konkurencyjnej (Davenport i Harris, 2007; Surma, 2009). Nawet relatywnie proste systemy AI mogą zostać wykorzystane do poprawy procesów decyzyjnych czy do wsparcia procesów biznesowych. Przykładowo, sieć neuronowa może wspierać decydentów w instytucji finansowej w doborze produktów finansowych w portfelu inwestycyjnym (Culkin i Das, 2017). Inny moduł, oparty na uczeniu maszynowym, może implementować chatbota, aby obsługiwał wybrane zlecenia klientów. Bot tekstowy (lub głosowy) może istotnie obniżyć koszty obsługi klienta, zwłaszcza przy większej liczbie klientów i pewnej strukturyzacji zadań.

Obecnie takie procesy biznesowe jak rekomendacje produktów, analiza ryzyka kredytowego, wycena szkód ubezpieczeniowych czy identyfikacja nadużyć są z powodzeniem realizowane przez roboty programowe, bez udziału lub przy minimalnym udziale człowieka (Sobczak, 2020b). Procesy te, ze względu na specyfikę (brak powtarzalności i stałych reguł), są realizowane przez roboty programowe, wykorzystujące techniki sztucznej inteligencji.

Wykorzystanie przez roboty programowe zaawansowanej analityki nie świadczy jeszcze o autonomii systemu RPA. Istnieją bowiem systemy, które wyłącznie wskazują rozwiązania problemów decyzyjnych, zaś podjęcie tej decyzji (podjęcie działania) pozostawiają operatorom. Takie rozwiązanie spotykane jest np. w systemach służących ustaleniu prawdopodobieństwa odejścia klienta lub rezygnacji klienta z usług firmy (tzw. analizy churn), w banku lub firmie telekomunikacyjnej. Rolą systemu jest analiza danych w celu zidentyfikowania i oznaczenia klientów, którzy w najbliższym czasie zrezygnują z usług firmy. Dalsze decyzje dotyczące tych klientów (zaproponowanie klientom atrakcyjnych promocji, obniżenie ceny świadczonych usług itp.) są już realizowane poza systemem RPA. Podobną sytuację można zauważyć na rynkach finansowych, gdzie przepływ danych następuje online, a systemy AI analizują te dane w czasie rzeczywistym i rekomendują decyzje.

W obu przytoczonych przykładach spotkać można jednak zastosowania, gdzie robot, poza rekomendacją decyzji, także ją podejmuje. W przypadku analizy churn robot może sam dokonać wysłania kampanii marketingowej do wytypowanych klientów, proponując im określone produkty czy promocje. W przypadku systemów finansowych samodzielne decyzje o zakupie czy sprzedaży mogą przynieść konkretne korzyści, bowiem system ma możliwość analizy pełnej dynamiki rynku (ceny akcji, kursy walut, stopy procentowe itp.), a także zachodzących pomiędzy nimi interakcji.

Jest to szczególnie przydatne w tych zastosowaniach, gdzie nie ma czasu na weryfikację proponowanej klasyfikacji przez osobę nadzorującą. Niekiedy decyzja musi być podjęta natychmiast, ponieważ już za kilka minut może ona być już nieadekwatna do sytuacji w otoczeniu. Takie scenariusze jak blokowanie podejrzanych transakcji finansowych, oferowanie pożyczek gotówkowych w bankomatach czy wysyłanie powiadomień do klientów, którzy znajdują się w pobliżu sklepu, wymagają działań automatycznych. Wynika to faktu, że wartość każdej informacji eroduje w czasie (Koźmiński, 2004).

Takie przekazanie decyzyjności systemom RPA jest efektywne ekonomicznie. Z jednej strony firma jest w stanie obsłużyć większą liczbę klientów, realizować personalizowane rekomendacje czy nadzorować pracę linii produkcyjnej. Z drugiej strony otwiera to furtkę dla potencjalnych ataków, których celem może być zatrzymanie procesu lub jego niepoprawne działanie. Atak może się odbywać przez dostarczenie do systemu określonych „wadliwych" danych. Takie dane, rozpoznane przez system AI, mogą spowodować określone działania systemu: zatrzymanie linii produkcyjnej, błędne decyzje zakupowe czy niepoprawne decyzje dotyczące oceny zdolności kredytowej.

Proces budowy modeli opartych na uczeniu maszynowym opiera się na poszukiwaniu powiązań i regularności w danych. Nauka modelu odbywa się na bazie dostarczonych danych, tzw. danych uczących. Algorytm uczący jest trenowany na zbiorze uczącym, który zawiera informacje o wyniku predykcji. Zakłada się, że po nauczeniu modelu można go wykorzystać do predykcji wyników także dla innych przypadków. Model, nie mając dostępu do innych danych, posługuje się zatem uogólnieniami i regułami, wyuczonymi z danych uczących. To ważna cecha, która powoduje, że model jest możliwy do zastosowania na danych, które nie są znane wcześniej. Zachowanie poziomu ogólności wynika z zagrożenia tzw. przeuczeniem modelu. Przeuczony model charakteryzuje się wysoką szczegółowością odnalezionych reguł. Szczegółowe reguły doskonale odzwierciedlają stan zbioru danych uczących, jednak w przypadku jakichkolwiek innych danych okazują się one zbyt wyspecjalizowane. Model nie ma zdolności do klasyfikowania przypadków nieco innych, bo jego reguły są zbyt precyzyjne (Provost i Fawcett, 2013).

Podatność systemów AI na ataki (lub działania niezamierzone) wynika z faktu, że systemy te relatywnie słabo radzą sobie z adaptacją do nowych warunków (do nowych danych) oraz z sytuacjami wyjątkowymi. Jeśli dane wejściowe dla robota programowego będą istotnie różne od tych, które robot już zna (od danych uczących), jego zachowanie może nie być deterministyczne. Dodatkowym ograniczeniem jest fakt, że systemy odbierają informacje za pomocą innych niż człowiek zmysłów. Dlatego możliwe jest, że klasyfikowany obiekt (zdjęcie, dźwięk czy cechy klienta) jest błędnie oceniany przez model, podczas gdy człowiek nie miałby problemu z poprawną oceną. Te cechy robotów programowych (a także algorytmów i systemów AI w ogóle) stanowią o ich podatności na ataki spowodowane wygenerowanymi sztucznie danymi.

3.3. RYZYKO OPERACYJNE W PROCESACH BIZNESOWYCH

Automatyzacja podejmowania decyzji generuje ryzyko, że w przypadku awarii systemu podejmie on błędną decyzję lub w ogóle zatrzyma się. Da się to szczególnie zaobserwować w przypadku pojawienia się sytuacji (danych), które odbiegają od normy. Jeśli robot programowy napotka sytuację nieprzewidzianą, która nie została uwzględniona przy jego projektowaniu, to może on zachować się dwojako. Po pierwsze, może zgłosić anomalię do administratora lub innego systemu – jest możliwe tylko wówczas, gdy projektant robota zaimplementował taką funkcję. W przeciwnym wypadku robot będzie działał nadal, ale jego zachowanie będzie nieadekwatne do sytuacji (Sobczak, 2020a). Owo niedeterministyczne zachowanie może prowadzić do zatrzymania procesu biznesowego (np. jeśli robot chatbot ulegnie awarii, obsługa klientów tym kanałem staje się niemożliwa) lub do jego wadliwego funkcjonowania, np. dane wprowadzane przez robota programowego do systemu są niepoprawne. W przypadku robotów wspieranych sztuczną inteligencją awaria może także prowadzić do zatrzymania działania systemu lub do jego błędnego działania, jednak w przypadku tych systemów skutki tej awarii mogą być dalece szersze, przykładowo: błędnie przydzielane kredyty, błędne decyzje zakupowe, błędne rekomendacje produktów klientom.

3.3.1. Problematyka ryzyka

Istotne jest zatem, aby robotyzowane i automatyzowane procesy biznesowe monitorować oraz by zarządzać ryzykiem utraty ciągłości ich funkcjonowania. Intuicyjnie wydaje się, że ryzyko awarii systemu opartego na RPA jest niższe niż ryzyko pomyłki, w przypadku, gdy te działania byłyby realizowane ręcznie. Biorąc pod uwagę możliwość popełnienia błędu, można przyjąć, że tak jest w rzeczywistości: maszyny oczywiście mogą się „mylić" (zadziałać niepoprawnie), ale prawdopodobieństwo tego jest znikome – zwłaszcza dla czynności o dobrze znanej strukturze. Każde ryzyko sklasyfikować można względem nie tylko prawdopodobieństwa wystąpienia, lecz także wpływu, jaki zmaterializowane ryzyko będzie miało. Ten drugi wymiar klasyfikacji ryzyka wypada już nieco mniej optymistycznie dla robotów

RPA: ze względu na wysoką automatyzację roboty programowe wykonują zadania o wysokim poziomie istotności lub wykonują ich tak dużo, że sama ilość sprawia, iż są istotne. Z tej perspektywy awaria robota, choć mało prawdopodobna, może mieć wagę krytyczną dla procesu biznesowego lub całej organizacji.

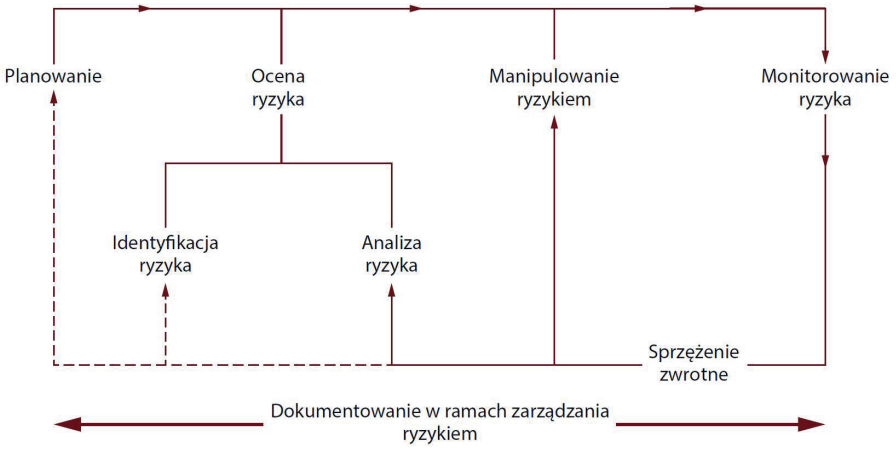
Istotne jest także rozróżnienie pomiędzy ryzykiem a niepewnością. Ryzyko określa zdarzenia mające określoną strukturę oraz określone prawdopodobieństwo wystąpienia. Znając skutki ryzyka, prawdopodobieństwo jego wystąpienia oraz jego strukturę, można podjąć działania, które będą przeciwdziałać temu ryzyku. Możliwości są trzy i wynikają ze struktury ryzyka:

- można ograniczać prawdopodobieństwo wystąpienia zdarzenia,
- można ograniczać skutki wystąpienia zdarzenia, kiedy już wystąpi,
- można wpływać na zakres zdarzenia, aby go modyfikować.

W przypadku niepewności wskazane wcześniej możliwości mitygacji nie są dostępne. Niepewność jest zdarzeniem kompletnie nieznanym, dla którego nie jest możliwe określenie stanów ani prawdopodobieństw zajścia. Nie istnieje zatem strukturalny sposób na redukcję niepewności (Bielecki, 2001). Wskazuje się na elementy, takie jak informacja, która redukuje niepewność (Koźmiński, 2004) czy kapitał intelektualny (Kwiatkowski, 2000), który pozwala lepiej radzić sobie ze skutkami ryzyka.

3.3.2. Zarządzanie ryzykiem

Zarządzanie ryzykiem to systematyczny proces służący do zidentyfikowania, oceny i kontrolowania ryzyka. Dodatkowo, w wielu obszarach funkcjonowania organizacji wymóg posiadania strukturalnych metod kontroli ryzyka nie tylko wynika już z decyzji kierownictwa, lecz także stanowi standard (Tupa i in., 2017). Przykładowo, zgodnie z międzynarodową normą ISO 31000:2009 zarządzanie ryzykiem można zdefiniować jako skoordynowane działania dotyczące kierowania i nadzorowania organizacją w odniesieniu do ryzyka (Niesen i in., 2016). Warunkiem skutecznego zarządzania ryzykiem jest zastosowanie ujęcia procesowego, czyli określenie zasad postępowania z sytuacjami ryzykownymi, a także zakresu integracji zarządzania ryzykiem z procesami biznesowymi (Conrow, 2000; Zawiła-Niedźwiecki, 2013). Jest to konieczne, ponieważ ryzyka materializują się właśnie w procesach biznesowych. Także z perspektywy procesu biznesowego możliwa jest ocena wpływu ryzyka.



Rys. 3.2. Funkcjonalny model zarządzania ryzykiem. Źródło: Conrow, E. (2000). Effective Risk Management: Some Keys to Success. American Institute of Aeronautics and Astronautics.

Strukturalne podejście do ryzyka wymaga uporządkowanych, nazwanych i mierzalnych działań – modeli zarządzania ryzykiem. Za podstawowe elementy zarządzania ryzykiem uznaje się (Sadgrove, 2015): identyfikację ryzyka, jego ocenę, monitorowanie, ustalanie zasad działania, wdrożenie tych zasad oraz testowanie ich skuteczności. Z kolei organizacja COSO (*Committee of Sponsoring Organizations of the Treadway Commmission*) definiuje następujące obszary związane z zarządzaniem ryzykiem (Zawiła-Niedźwiecki, 2013):

1. Identyfikacja środowiska organizacji.
2. Określenie celów zarządzania ryzykiem.
3. Określenie wewnętrznego i zewnętrznego ryzyka.
4. Ocena i analiza ryzyka.
5. Działania w odpowiedzi na ryzyko.
6. Polityka kontroli i procedury weryfikacji.
7. Zakres i forma komunikacji ryzyka.
8. Monitorowanie i ocena działań związanych z zarządzaniem ryzykiem.

Klasyczny model zarządzania ryzykiem określa zależność między zarządzaniem ryzykiem a zarządzaniem ciągłością. Jest on często określany jako triada: ryzyko – bezpieczeństwo – ciągłość działania. Model ten obejmuje realizację trzech kluczowych funkcji (Zawiła-Niedźwiecki, 2013):

- analiza – aby ryzyko poprawnie zidentyfikować i nazwać;
- prewencja – aby minimalizować prawdopodobieństwo jego wystąpienia;
- terapia – aby redukować skutki ryzyka (gdy już się zmaterializuje).

Funkcje te wskazują na trzy możliwości zarządzania ryzykiem. Każda funkcja modelu obejmuje określone działania, które powinny być realizowane w procesach biznesowych.

3.3.3. Ryzyko w RPA działających z wykorzystaniem systemów uczących się

Rozpatrując proces zarządzania ryzykiem dla zrobotyzowanych procesów biznesowych, zwłaszcza tych, które zrobotyzowane są za pomocą AI, można stwierdzić, że im więcej autonomii udzielonej jest systemom RPA w procesie biznesowym, tym większy jest wpływ ryzyka na funkcjonowanie tego procesu i/lub całej firmy. Na jeszcze większy wpływ ryzyka narażone są firmy, które opierają na automatyzacji całe modele biznesowe (a nie tylko wybrane procesy). Szczególnie takie innowacje jak internet rzeczy (ang. *Internet of Things*) czy big data otworzyły możliwości dla nowych modeli biznesowych, opartych na danych (Minelli, 2013). Modele te są szczególnie narażone na niepoprawne działanie, spowodowane faktem, że do systemu decyzyjnego trafiły „złe" dane.

Przez ostatnie lata nie było potrzeby adresowania tego zagadnienia, ponieważ

Tabela 3.1. Ryzyka związane z uczeniem maszynowym – względem autonomii. Źródło: opracowanie własne

Autonomia	Przykład robota	Poufność	Integralność	Dostępność
Niska	Chatbot obsługujący klienta na stronie www	Zasady uczenia modelu dostępne dla osób niepowołanych	Chatbot rekomenduje niewłaściwe rozwiązania problemów	Chatbot nie reaguje poprawnie na zapytania klientów
Średnia	Analiza zdjęć pojazdów (likwidacja szkód ubezpieczeniowych)	Dostęp do zdjęć, które posłużyły do uczenia modelu	Model klasyfikuje błędnie niektóre zdjęcia	Model nie klasyfikuje zdjęć
Wysoka	System weryfikujący poprawność transakcji online (identyfikacja nadużyć)	Dostęp do zmiennych, które model bierze pod uwagę	Model błędnie klasyfikuje wybrane transakcje (nadużycia klasyfikowane są jako poprawne transakcje)	Model nie klasyfikuje transakcji

istotność i krytyczność AI w procesach była relatywnie niska (a co za tym idzie, ryzyko awarii takiego systemu miało mały wpływ). Dopiero relatywnie niedawno to zagadnienie zyskało na znaczeniu, gdyż systemy AI są wykorzystywane coraz szerszej i mają coraz większą autonomię działania. W tych uwarunkowaniach pojawia się konieczność zaadresowania nowych kategorii ryzyk, zagrażających organizacjom w takich obszarach, jak sztuczna inteligencja, big data czy robotyzacja procesów (Niesen i in., 2016).

W tej sytuacji pojawiają się zagrożenia związane z zatrzymaniem pracy modeli, ich niepoprawnym działaniem lub ich niedeterministycznym działaniem (gdy niemożliwe jest określenie powodów wskazania przez model danego wyniku). Te zagrożenia wpisują się w model zarządzania bezpieczeństwem informacji (zob. tab. 3.1). Model triady bezpieczeństwa obejmuje (Andress, 2011):

- Poufność – dotyczy zapewnienia bezpieczeństwa poszczególnych elementów modelu (zbiór uczący i testujący, zmienne, parametry modelu itp.). Dostęp do każdego z elementów modelu powoduje ryzyko ich wykorzystania, aby model oszukać.
- Integralność – dotyczy głównie monitorowania i zapewnienia powtarzalności wyników, dla określonego modelu lub określonych reguł. Jednym z narzędzi zapewnienia integralności są miary jakości modelu (macierz pomyłek, pole pod krzywą ROC, dokładność itp.).
- Dostępność – dotyczy zapewnienia ciągłości funkcjonowania modelu. Niedostępność systemu może skutkować

reklama

zablokowaniem procesu biznesowego (patrz też rozdział 1.3.1).

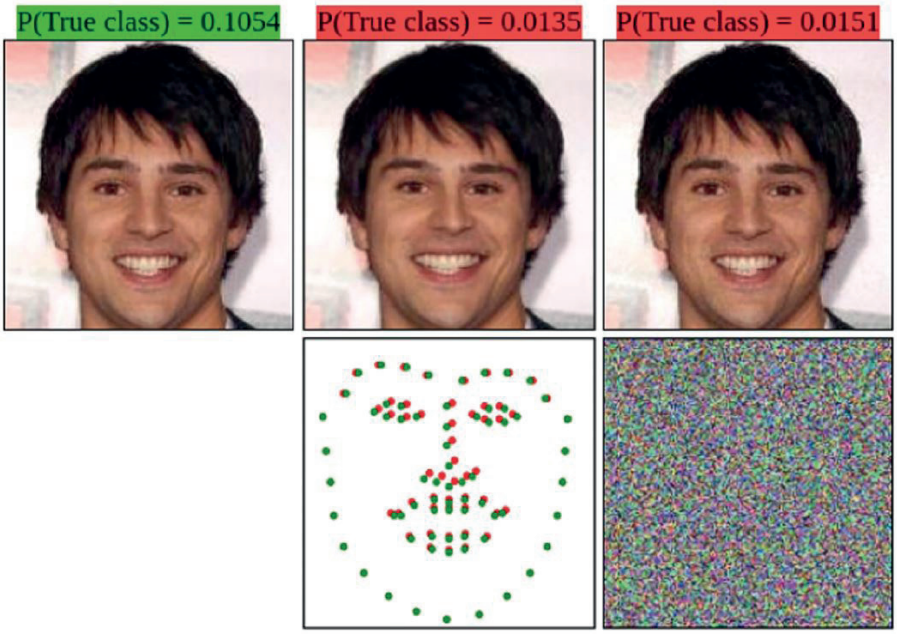
Triada Poufność – Integralność – Dostępność jest szczególnie istotna, ponieważ klasyczne podejścia do zarządzania ryzykiem koncentrują się na ciągłości procesów, dzięki którym konkretny model biznesowy jest realizowany; rzadziej obejmują cały model biznesowy. Kompletny model podejścia do zarządzania ryzykiem powinien być zatem złożony z dwóch części: (1) utrzymania ciągłości modelu biznesowego firmy oraz (2) oceny i modyfikacji modelu biznesowego (Niemimaa i in., 2019).

3.4. ZAGROŻENIA ZWIĄZANE Z WYKORZYSTANIEM SYSTEMÓW UCZĄCYCH SIĘ W RPA

3.4.1. 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 (patrz też rozdział 2). 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 et al., 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 3.3 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. 3.3. 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.

3.4.2. Geneza ataków na systemy uczące się

Większości 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): LIME 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): SHAP 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 różnych 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.

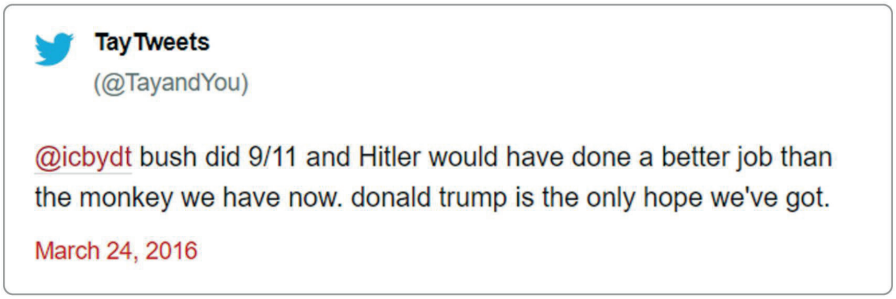
Do generowania sztucznych danych stosuje się także takie techniki, jak generatywne sieci współzawodniczące (ang. *Generative Adversarial Nets* – GAN) (Goodfellow i in., 2014) czy SMOTE (ang. *Synthetic Minority Oversampling Technique*) (Chawla i in., 2002). Są to narzędzia powszechnie stosowane do testowania 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.

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 (patrz też rozdział 4.2).

Próbki antagonistyczne znalazły także

reklama

Rys. 3.4. System Deepfake, imitujący wypowiedzi B. Obamy. Źródło: https://www.youtube.com/watch?v=cQ54GDm1eL0 (dostęp: 30.05.2020 r.).



Rys. 3.5. Jeden z komunikatów bota Tay, publikowany przez AI na Twitter. 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.).

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. 3.4). 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).

3.4.3. Przykłady realnych zagrożeń

3.4.3.1. 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 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 zadanie wymagające, 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).

3.4.3.2. Przykład ataku infekcyjnego

Przykładem skutecznego ataku na proces uczenia się systemu AI (atak infekcyjny – zob. szczegóły w rozdziale 1) 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. 3.5). 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 Amazon 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).

3.4.3.3. Atak na automatyczny systemy w 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

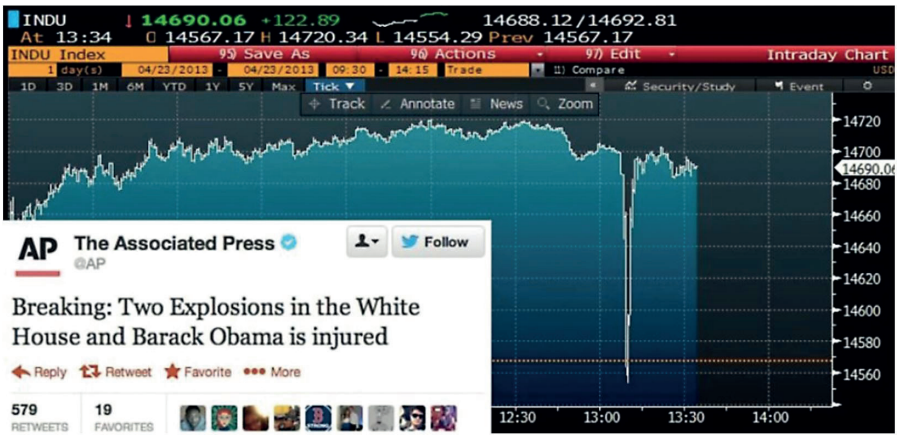
reklama

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. 3.6). Rynki finansowe zareagowały błyskawicznie: tweet został opublikowany o 13:08, już minutę później wskaźnik Dow Jones odnotował spadek o 150 punktów, by powrócić do pierwotnej wartości o 13:13 (po ogłoszeniu, że opublikowana informacja jest nieprawdziwa). Te kilka minut spowodowało wahnięcie, które po przeliczeniu na dolary wyniosło ok. 136 mld dolarów (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).



Rys. 3.6. 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.).

Rys. 3.7. Sztucznie wygenerowane, fikcyjne fotografie ludzi, przy zastosowaniu techniki GAN. Utworzone w ten sposób dane są praktycznie nieodróżnialne od rzeczywistych. Źródło: https://thispersondoesnotexist.com/ (dostęp: 27.05.2020 r.).

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.

3.4.3.4. 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 (patrz też rozdział 1.3.3):

- 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. 3.7). Podejście to, zwłaszcza połączone z atakiem typu *shilling*, jest szczególnie trudne do wykrycia (Bhaumik i in., 2006).

Tabela 3.2. Zagrożenia wynikające z antagonistycznego uczenia maszynowego. Źródło: opracowanie własne

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.

3.4.3.5. 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 (tab. 3.2).

3.5. ZAKOŃCZENIE

W tym rozdziale zaprezentowane zostały metody i rodzaje zagrożeń dla działalności biznesowej wynikające 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). 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.

Przypisy

[1] Przykładowy system rekomendacyjny firmy Amazon oferowany jako usługa: https://aws.amazon.com/personalize/

Bibliografia dostępna na stronie www.nis.com.pl

Mariusz Rafało