Etyka sztucznej inteligencji – próby regulacji

Rozwój sztucznej inteligencji jest nieunikniony – im więcej obszarów ona obejmuje, tym większe zagrożenie, że człowiek stanie się zakładnikiem technologii, którą tworzy. Dlatego niezwykle istotne jest zadbanie o wdrożenie standardów etycznych, którym rozwój SI będzie podlegał. Etyki nie da się zaprogramować, mówimy więc w istocie o zasadach etycznych obowiązujących twórców rozwiązań SI i jej użytkowników.

dr Tomasz Kulisiewicz
sekretarz Sektorowej Rady ds. Kompetencji – Informatyka

dr hab. Andrzej Sobczak
prof. SGH – kierownik Zakładu Zarządzania Informatyką w Szkole Głównej Handlowej w Warszawie. Założyciel inicjatyw polskiej społeczności hiperautomatyzacji,twórca serwisu Robonomika.pl. W swojej działalności naukowej, dydaktycznej i doradczej zajmuje się m. in. zarządzaniem strategicznym IT, ładem danych i architekturą korporacyjną oraz zaawansowaną automatyzacją i robotyzacją procesów biznesowych.

W 1942 r. sławny autor fantastyki Isaac Asimov w opowiadaniu *Zabawa w berka*[1] sformułował trzy prawa robotyki. Od tego czasu pojawiają się próby regulacji stosunków między (przyszłymi) maszynami dysponującymi sztuczną inteligencją a ludźmi, m.in. rozszerzone prawa robotyki Rogera Clarke'a, australijskiego naukowca i konsultanta[2]. Do zakazu wyrządzenia krzywdy rodzajowi ludzkiemu dodał on hierarchizację praw oraz ciekawe z punktu widzenia rozwoju AI Prawo Prokreacji, zgodnie z którym robot nie może brać udziału w projektowaniu i wytwarzaniu robotów, jeśli działania takich robotów byłyby niezgodne z prawami robotyki. Pionierskie prace wywoływały różne reakcje. Na konwencji S-F Novacon w 1995 r. brytyjski pisarz fantastyki Dave Langford sparodiował ówczesne oficjalne próby regulacji zastosowań robotyki i sztucznej inteligencji:

1. Robot nie może wyrządzić żadnej krzywdy upoważnionemu personelowi rządowemu, ale ma likwidować wszystkich jego przeciwników.

1 Polskie wydanie m.in. w zbiorze *Ja, robot* (przekł. Z.A. Królicki), Rebis, Poznań 2013.

2 http://www.rogerclarke.com/SOS/Asimov.html (dostęp: 5.05.2023)

2. Robot ma wypełniać rozkazy wydane przez uprawniony personel, z wyjątkiem przypadków, w których będzie to sprzeczne z trzecim prawem.

3. Robot ma skutecznie chronić siebie samego używając broni przeciwpiechotnej, ponieważ jest cholernie drogi.

Wraz z rozwojem sztucznej inteligencji próby regulacji jej etyki stają się coraz większym wyzwaniem.

Sztuczna inteligencja godna zaufania

W obszernym opracowaniu „Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji", przygotowanym przez Grupę ekspertów wysokiego szczebla ds. sztucznej inteligencji powołaną przez Komisję Europejską w 2018 r. i opublikowanym w kwietniu 2019 r.[3] określono cechy sztucznej inteligencji godnej zaufania.

SI godna zaufania powinna być:

- **zgodna z prawem,**
- **etyczna (zgodna z zasadami i wartościami etycznymi),**
- **solidna zarówno z technicznego, jak i ze społecznego punktu widzenia, ponieważ systemy SI mogą wywoływać niezamierzone szkody nawet wówczas, gdy korzysta się z nich w dobrej wierze.**

W wytycznych określono ramy sprzyjające osiągnięciu godnej zaufania sztucznej inteligencji i przedstawiono pilotażową listę kontrolną oceny godnej zaufania sztucznej inteligencji, zalecając stosowanie (i dostosowywanie) jej w konkretnych przypadkach wdrażania systemów stosujących rozwiązania SI.

Wersja pilotażowa zawierała kilkadziesiąt pytań kontrolnych zgrupowanych w blokach obejmujących kolejno przewodnią i nadzorczą rolę człowieka, techniczną solidność i bezpieczeństwo systemów, ochronę prywatności i zarządzanie danymi, przejrzystość (a w niej wytłumaczalność/wyjaśnialność), zapewnienie różnorodności, niedyskryminacji i sprawiedliwości, dobrostan społeczny i środowiskowy, odpowiedzialność (a w niej możliwość dochodzenia roszczeń). Podkreślono przy tym, że „zapewnienie wdrożenia godnej zaufania sztucznej inteligencji nie polega na mechanicznym „odhaczaniu" pozycji z listy, ale na ciągłym identyfikowaniu wymogów, ocenianiu rozwiązań, zapewnianiu lepszych rezultatów przez cały cykl życia systemu SI oraz włączaniu zainteresowanych stron w podejmowane działania".[4]

Systemy wysokiego ryzyka wg AIA

W dyskusji redakcyjnej w nrze 4/2022 „Domeny" (s. 4) mowa była m.in. o AIA (Artificial Intelligence Act) – procedowanym od 2021 r. projekcie rozporządzenia Parlamentu Europejskiego i Rady w sprawie sztucznej inteligencji[5]. Celami proponowanej regulacji mają być:

- zapewnienie, aby systemy sztucznej inteligencji były bezpieczne i zgodne z prawami podstawowymi oraz unijnymi wartościami;
- pewność prawa dla ułatwienia inwestycji i innowacji w dziedzinie sztucznej inteligencji;
- poprawa zarządzania i skuteczne egzekwowanie obowiązujących przepisów dotyczących praw podstawowych i wymogów bezpieczeństwa;
- ułatwienie rozwoju jednolitego rynku dla zgodnych z prawem, bezpiecznych i wiarygodnych zastosowań sztucznej inteligencji oraz zapobieganie fragmentacji rynku.

W projekcie określono systemy sztucznej inteligencji wysokiego ryzyka. Na podstawie ustawy o Krajowym Systemie Cyberbezpieczeństwa oraz dyrektywy NIS2 pod tym pojęciem skłonni jesteśmy rozumieć tylko systemy dostarczające usługi kluczowe: np. dyrektywa NIS2 określa podmioty niezbędne (*essential entities*) i podmioty istotne (*important entities*). Tymczasem projekt AIA znacznie poszerza definicję systemów wysokiego ryzyka.

Zalicza do nich nie tylko systemy zarządzające infrastrukturą krytyczną (m.in. woda, gaz, ciepło, energia elektryczna), lecz także rozwiązania, które mogą być stosowane w obszarze kształcenia lub szkolenia zawodowego (przy podejmowaniu decyzji o dostępie do instytucji kształcenia i szkolenia oraz do oceniania osób na testach), systemy HR do rekrutacji i wyboru kandydatów, do podejmowania decyzji o awansie czy o rozwiązaniu stosunku pracy oraz do przydzielania zadań, monitorowania lub oceny pracowników, a także systemy do oceny zdolności

3 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (dostęp: 5.05.2023)

4 tamże, s. 42.

5 https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX:52021PC0206 (dostęp: 5.05.2023)

kredytowej osób fizycznych i oceny wniosków obywateli o świadczenia pomocy publicznej. Jako systemy wysokiego ryzyka powinny być także klasyfikowane systemy stosowane do wysyłania służb pierwszej pomocy w sytuacjach nadzwyczajnych, w tym do ustalania priorytetów w ich wysyłaniu.

Definicja systemów SI

Art. 3 projektu AIA stanowi:

1) system sztucznej inteligencji oznacza oprogramowanie opracowane przy użyciu co najmniej jednej spośród technik i podejść wymienionych w załączniku I, które może – dla danego zestawu celów określonych przez człowieka – generować wyniki, takie jak treści, przewidywania, zalecenia lub decyzje wpływające na środowiska, z którymi wchodzi w interakcję.

W przywołanym załączniku I wymieniono:

a) mechanizmy uczenia maszynowego, w tym uczenie nadzorowane, uczenie się maszyn bez nadzoru i uczenie przez wzmacnianie, z wykorzystaniem szerokiej gamy metod, w tym uczenia głębokiego;

b) metody oparte na logice i wiedzy, w tym reprezentacja wiedzy, indukcyjne programowanie (logiczne), bazy wiedzy, silniki inferencyjne i dedukcyjne, rozumowanie (symboliczne) i systemy ekspertowe;

c) podejścia statystyczne, estymacja bayesowska, metody wyszukiwania i optymalizacji.

Jako bardzo ważny obszar wymieniono systemy stosowane przez organy ścigania i wymiaru sprawiedliwości. Jeśli bowiem system będzie trenowany na danych niskiej jakości, nie będzie spełniać odpowiednich wymogów dokładności i solidności lub nie zostanie odpowiednio przetestowany, może dyskryminować jakieś osoby, traktować je niesprawiedliwie, pozbawiać prawa do bezstronnego sądu, do obrony oraz prawa domniemania niewinności – zwłaszcza jeśli systemy sztucznej inteligencji nie będą dostatecznie przejrzyste, wyjaśnialne i udokumentowane. Dlatego systemami wysokiego ryzyka w tym ujęciu są systemy stosowane przez organy ścigania do indywidualnej oceny ryzyka (m.in. poligrafy i inne narzędzia do wykrywania stanu emocjonalnego osoby fizycznej), narzędzia do wykrywania treści typu *deepfake*, do oceny wiarygodności dowodów w postępowaniu karnym, do przewidywania wystąpienia faktycznego lub potencjalnego przestępstwa na podstawie profilowania osób fizycznych lub oceny cech osobowości[6]. Celem tej regulacji jest wyeliminowanie ryzyka tendencyjności oraz tzw. efektu czarnej skrzynki w systemach, które mają wspierać sądy w badaniu i interpretacji faktów oraz w stosowaniu przepisów do konkretnego stanu faktycznego.

W świetle poważnych problemów Europy z imigracją jako systemy wysokiego ryzyka w projekcie AIA wymieniono także systemy wykorzystywane w zarządzaniu migracją, azylem i kontrolą graniczną. Dokładność, niedyskryminujący charakter i przejrzystość systemów SI wykorzystywanych w tych obszarach są szczególnie istotne dla poszanowania praw podstawowych osób, w szczególności prawa do swobodnego przemieszczania się, niedyskryminacji, ochrony życia prywatnego i danych osobowych oraz ochrony międzynarodowej.

Niełatwy do wypełnienia będzie określony w art. 13 projektu obowiązek przejrzystości i udostępniania informacji użytkownikom oraz skutecznego nadzoru systemów SI wysokiego ryzyka przez ludzi (art. 14), przede wszystkim z uwagi na trudne do realizacji postulaty wyjaśnialności systemów. Elementy tego podejścia są już nawet obecne w polskim prawie bankowym. W art. 105a Prawa bankowego jest ust. 1a: „Banki (….) mogą podejmować decyzje, opierając się wyłącznie na zautomatyzowanym przetwarzaniu, w tym profilowaniu, danych osobowych – również stanowiących tajemnicę bankową – pod warunkiem zapewnienia osobie, której dotyczy decyzja podejmowana w sposób zautomatyzowany, prawa do otrzymania stosownych wyjaśnień co do podstaw podjętej decyzji, do uzyskania interwencji ludzkiej w celu podjęcia ponownej decyzji oraz do wyrażenia własnego stanowiska .”[7]

Zakazane praktyki

Zgodnie z AIA zakazane ma być stosowanie w systemach SI technik podprogowych, mających na celu zniekształcenie czy wymuszenie jakiegoś zachowania osób, a także systemów wykorzystujących dowolne słabości określonej grupy osób ze względu na ich wiek, niepełnosprawność ruchową lub zaburzenie psychiczne, wykorzystywania przez organy publiczne lub w ich imieniu systemów dla oceny lub klasyfikacji wiarygodności osób fizycznych. Nie będzie wolno wykorzystywać systemów SI do zdalnej identyfikacji biometrycznej w czasie rzeczywistym „w przestrzeni publicznej do

6 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (dostęp: 5.05.2023)

7 tamże, s. 42.

celów egzekwowania prawa, chyba że jest to absolutnie niezbędne do poszukiwania konkretnych potencjalnych ofiar przestępstw, w tym zaginionych dzieci (…) lub zapobiegnięcia konkretnemu, poważnemu i bezpośredniemu zagrożeniu życia lub bezpieczeństwa fizycznego osób fizycznych lub atakowi terrorystycznemu".

Życie to nie algorytm

W formułowaniu kodeksów etycznych i propozycji regulacyjnych mierzymy się z różnymi wyzwaniami. O kilku z nich jest mowa w rozmowie, jaką w ramach projektu „Strefa Psyche" Uniwersytetu SWPS Ewa Pluta przeprowadziła z doktorantką na Uniwersytecie Stanforda, Agatą Foryciarz, zajmującą się metodami AI/ML wspomagającymi lekarzy. Już sam tytuł rozmowy pokazuje rozległość problemu: „Etyczna sztuczna inteligencja. Czy etykę można zaprogramować?". A. Foryciarz zwraca uwagę, że „…system inteligentny może dobrze działać według definicji sprawiedliwości algorytmicznej w zakresie uśrednień i prognoz. Pojedyncze historie ludzkie mogą jednak w tym wszystkim zginąć. Wracając do przykładu medycznego, na poziomie jednostkowym algorytm może podejmować znacznie gorsze decyzje niż lekarz, który pacjenta postrzega w kontekście, a nie tylko jako daną w modelu statystycznym." Na pytanie postawione w tytule badaczka odpowiada: „Systemy oparte na sztucznej inteligencji realizują ściśle zdefiniowany cel. Niech to będzie zwiększenie użyteczności. Na wstępie należy opisać w matematyczny sposób, czym jest użyteczność. Trudno to zrobić. Podobnie jest z wartościami takimi jak szczęście, zadowolenie, jakość życia. Jak je zdefiniować matematycznie?"[8].

W środowisku firm zajmujących się rozwiązaniami AI/ML pojawiają się obawy dotyczące przeregulowywania różnych obszarów działalności, zwłaszcza w Unii Europejskiej. Krytycy zbytniej regulacji podnoszą kwestie zarówno ekonomicznych kosztów dostosowywania się do legislacji, jak i potencjalnego hamowania innowacyjności i utraty pozycji w globalnym wyścigu w sytuacji, gdy nie wszystkie kraje (zwłaszcza azjatyckie i bliskowschodnie) będą się regulacjami etyki SI przejmować.

Inwencja dostawców rozwiązań SI

Firma NICE, tworząca narzędzia RPA (Robotic Process Automation), zaproponowała etyczne ramy robotyzacji procesów biznesowych (Robo-Ethical Framework) z wykorzystaniem różnych rozwiązań SI, w tym robotów programowych[9].

Zgodnie z nimi roboty muszą być projektowane z uwzględnieniem pozytywnego oddziaływania na otoczenie społeczne i ekonomiczne oraz na środowisko. Nie powinny uwzględniać cech osobniczych: pochodzenia etnicznego, koloru skóry, wyznania, płci, wieku i statusu osób, których dotyczą ich działania. Roboty muszą być projektowane w taki sposób, aby zminimalizować ryzyko wyrządzenia indywidualnej krzywdy, zaś ludzie muszą być w stanie przeprowadzić audyt procesów i decyzji podejmowanych przez automaty. Jeśli okaże się, że robot może zaszkodzić człowiekowi, człowiek musi móc interweniować, by naprawić taką sytuację i zapobiec jej występowaniu w przyszłości. Automaty muszą być trenowane na znanych, zweryfikowanych i zaufanych danych. Dane wykorzystywane do szkolenia algorytmów powinny pozwalać na odwołanie się do ich oryginalnego źródła. Automaty muszą być projektowane z zapewnieniem mechanizmów zarządzania i nadzoru. Platforma wykorzystywana do automatyzacji procesów powinna być zaprojektowana tak, aby chronić przed nadużyciem władzy i nielegalnym dostępem. Musi proaktywnie monitorować i uwierzytelniać każdy dostęp do platformy oraz monitorować każdą akcję podjętą na platformie.

SI po chińsku

Błyskawiczny rozwój zastosowań SI w chatbotach już wywołał nie tylko problemy etyczne, lecz także polityczne. Niedawno opublikowane zostały wytyczne chińskiego Urzędu ds. Cyberprzestrzeni, zgodnie z którymi „treści tworzone przez generatywną SI winny odzwierciedlać fundamentalne wartości socjalizmu, nie mogą zawierać treści nawołujących do obalenia władzy państwowej i ustroju socjalistycznego, podżegać do podziału kraju, podważać jedność narodową" oraz „przekazywać informacje mogące zakłócić działanie gospodarki i porządek społeczny"[10]. Jedna z pierwszych mobilnych aplikacji, ChatYuan, została zamknięta przez wspomniany urząd już po trzech dniach działania[11], a jej autorzy pracują nad oddzielnym rozwiązaniem SI, które ma wykrywać i odsiewać przykłady przyjęcia przez chatbota „niewłaściwego punktu widzenia".

8 https://web.swps.pl/strefa-psyche/blog/relacje/22399-etyczna-sztuczna-inteligencja-czy-etyke-mozna-zaprogramowac (dostęp: 8.05.2023)

9 https://robonomika.pl/etyczne-aspekty-robotyzacji-procesow-biznesowych-propozycja-nice (dostęp: 4.05.2023)

10 https://gizmodo.com/ai-china-regulations-free-speech-baidu-ernie-chatgpt-1850329689 (dostęp 4.05.2023)

11 https://www.taiwannews.com.tw/en/news/4807319 (dostęp 4.05.2023)