Prof. Jemielniak: "Potrzebujemy czwartego prawa robotyki, by AI nie podszywała się pod człowieka"

© Kasper Kalinowski – wywiad z prof. Dariuszem Jemielniakiem

Gazeta Wyborcza, 11.02.2025

https://wyborcza.pl/7,75400,31680184,prof-jemielniak-potrzebujemy-czwartego-prawa-robotyki.html?token=Ve28LGypTUaChUzt3pFlAo-aw1aJ1mORn_WqH8VkJc0

Czwarte prawo robotyki to zakaz podszywania się AI pod człowieka (Fot. DC Studio/Shutterstock)

Wierzę, że wprowadzenie czwartego prawa robotyki - zakaz podszywania się AI pod człowieka - może się okazać jednym z kluczowych elementów przywrócenia przejrzystości i zaufania w erze deepfake'ów i generatywnych modeli językowych - mówi prof. Dariusz Jemielniak z Akademii Leona Koźmińskiego.

Dariusz Jemielniak *- profesor zarządzania Akademii Leona Koźmińskiego, kierownik Katedry Management in Networked and Digital Societies (MINDS), pracuje też jako faculty associate w Berkman-Klein Center for Internet and Society na Uniwersytecie Harvarda. Bada dezinformację w sieci i ruchy antyszczepionkowe. Jest członkiem Rady EIT (European Institute of Innovation & Technology).*

_____

Kasper Kalinowski: Słynne „trzy prawa robotyki", to dzieło legendy s-f Isaaca Asimova. Na czym polegają?

Prof. Dariusz Jemielniak: Asimov sformułował trzy prawa robotyki w opowiadaniu „Runaround" z 1942 r., a szeroko je rozwinął w zbiorze opowiadań „Ja, Robot" opublikowanym w 1950 r. W oryginale brzmią

1. **Robot nie może skrzywdzić człowieka** ani przez zaniechanie działania dopuścić, by człowiek doznał krzywdy.
2. **Robot musi być posłuszny rozkazom człowieka**, o ile nie stoją one w sprzeczności z pierwszym prawem.
3. **Robot musi chronić samego siebie**, o ile nie stoi to w sprzeczności z pierwszym lub drugim prawem.

Te „prawa" miały stać na straży bezpieczeństwa ludzi oraz zapewniać, że roboty pozostaną pod kontrolą człowieka.

W późniejszych powieściach, jak „Roboty i Imperium" z 1985 r., Asimov wzbogacał te zasady o tzw. Zerowe Prawo (o ochronie ludzkości), pokazując jednocześnie, że prosty zapis praw może prowadzić do nieoczekiwanych dylematów moralnych.

Warto jednak pamiętać, że wprowadzał je w czasach, gdy znacznie więcej myślało się o fizycznych robotach niż o światach wirtualnych, awatarach i czatbotach.

Dlaczego te prawa są tak popularne? Wciąż stanowią punkt odniesienia w debatach na temat etyki w opracowywaniu sztucznej inteligencji.

Przede wszystkim ich siłą jest prostota i atrakcyjność literacka. Zapisane w formie klarownych reguł, stały się łatwo przyswajalnym fundamentem i dobrze opisują uniwersalność problemu.

Pytanie, jak maszyna ma (nie) szkodzić człowiekowi, jest wciąż aktualne w dobie algorytmów decydujących o kredytach czy szansach zwolnienia warunkowego. O izraelskich systemach AI podpowiadających, kogo zabić, nie wspominając.

Zbiory opowiadań Asimova stały się też inspiracją dla wielu filmów, seriali i książek, co rozpowszechniło tę ideę w debacie publicznej.

Nawet jeśli w praktyce niewiele dzisiejszych systemów AI „myśli" w kategoriach trzech praw, to stanowią one metaforę dla prawdziwych dylematów: jak wbudować etykę w oprogramowanie. Temat ten zdecydowanie interesuje ludzi, czego dowodem są choćby dyskusje o decyzjach podejmowanych przez auta autonomiczne, czy tzw. dylemat wagonika, który stał się tematem niezliczonych memów.

To tylko obszar science-fiction czy etyka wdziera się współcześnie do świata maszyn i rozwiązań opracowywanych przez ich twórców?

Jeszcze kilka dekad temu etyka robotyki mogła się wydawać tematem wyłącznie z literatury s-f. Obecnie tak już nie jest. Po pierwsze, algorytmy decyzyjne w amerykańskim systemie sprawiedliwości, takie jak COMPAS, oceniają ryzyko recydywy i mają wpływ na orzeczenia sędziowskie. Skądinąd, wykazały one tendencje do uprzedzeń rasowych, co rodzi poważne pytania etyczne.

Autonomiczne pojazdy (Waymo, Tesla) całkiem realnie muszą rozwiązywać problem, kogo chroni auto w sytuacji bezpośredniego zagrożenia. Roboty opiekuńcze (np. system PARO w Japonii, roboty Nao czy Pepper) wspierają osoby starsze w szpitalach i domach spokojnej starości, co wywołuje pytania o emocjonalne przywiązanie pacjentów do maszyn i granice prywatności. To pokazuje, że etyka nie jest już teorią, lecz wkracza do praktyki inżynierskiej.

Jak trafnie obserwuje prof. Patrick Lin, projektowanie robotów bez uwzględnienia aspektów etycznych byłoby jak projektowanie samochodów bez hamulców.

To nie jedyna próba zdefiniowania etyki w pracach nad AI. Jakie inne próby są warte uwagi?

Prób tego rodzaju jest oczywiście wiele. Choćby Google ogłosiło w 2018 r. „AI Principles", deklarując m.in. unikanie tworzenia narzędzi naruszających prawa człowieka (zob. blog Google AI). Microsoft przyjął ramy „Responsible AI", publikując szczegółowe zasady i system wewnętrznej weryfikacji projektów pod kątem etyki. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) przygotowała obszerne raporty z rekomendacjami dla inżynierów i decydentów.

Unijna grupa ekspercka High-Level Expert Group on AI opracowała w 2019 r. „Ethics Guidelines for Trustworthy AI" – kompleksowy dokument o transparentności, ochronie prywatności i odpowiedzialności.

Wszelkie tego typu próby są istotne, choć bez realnego wsparcia, choćby w postaci konsensusu mocarstw, mają niewielką siłę sprawczą.

Pan postuluje stworzenie czwartego prawa. Na czym miałoby polegać?

Faktycznie, proponuję, a mój manifest, opracowany w moim zespole Campus AI, opublikowało znane czasopismo specjalistyczne, IEEE Spectrum:

4. **Robot nie może wprowadzać człowieka w błąd**, podszywając się pod innego człowieka.

W praktyce oznacza to jasne rozgraniczenie między byciem sztuczną inteligencją a człowiekiem: AI powinna się identyfikować jako AI, zwłaszcza gdy podejmuje z nami interakcje w czasie rzeczywistym (rozmowy telefoniczne, czaty). Dodatkowo treści generowane przez AI powinny być wyraźnie oznaczone (tzw. watermarking), chyba że człowiek je głęboko przeedytuje.

Warto zauważyć, że sam zakaz krzywdzenia człowieka (z prawa pierwszego) nie wystarcza, bo choćby np. nastolatki nawiązują relacje emocjonalne z awatarami, ponadto krzywdzenie zakłada intencję oszukania, tymczasem w wielu przypadkach możemy mieć do czynienia z szarą strefą, botami udającymi ludzi bez celu czynienia krzywdy, ale koniec końców z niepożądanym skutkiem.

Realne przykłady zagrożeń braku takiej regulacji już są.

Choćby w 2019 r. w Wielkiej Brytanii prezes firmy energetycznej stracił 220 tys. euro, bo oszust podszył się głosowo pod szefa niemieckiej spółki-matki.

Mamy do czynienia de facto z metodą oszustwa "na wnuczka", gdzie rozmawiamy z kimś, kto mówi i wygląda jak wnuczek. Oczywiście istotnym impusem do działania są też liczne przypadki automatycznego generowania fake newsów czy zaawansowanych phishingów, co zwłaszcza po premierze otwartego modelu DeepSeek, ale i z wykorzystaniem Llama [zaawansowany model językowy opracowany przez Meta - red.] , jest niestety po prostu bardzo tanie.

Mój zespół w Akademii Leona Koźmińskiego, gdzie wraz z dr. Leonem Ciechanowskim prowadzimy intensywne badania nad metodami wykrywania botów, w ostatnim roku przeżywa klęskę urodzaju, jeśli chodzi o ilość rozpowszechnianej dezinformacji.

Technologicznie to możliwe? Mylimy już wytwory AI z wytworami człowieka?

Coraz częściej. Modele językowe potrafią pisać teksty nieodróżnialne od ludzkich. Systemy generujące obraz (Midjourney, DALL·E) tworzą fotorealistyczne obrazy. Istnieją jednak narzędzia do próby rozpoznawania, np. Content Credentials od Adobe, gdzie do cyfrowych treści dołączane są ukryte metadane. Możliwa jest też analiza statystyczna i widmowa w wykrywaniu deepfake'ów głosowych (firma Pindrop deklaruje skuteczność rzędu 99 proc., choć bardzo wątpię w to, czy tego typu skuteczność ma jakiekolwiek szanse się utrzymać).

Wykorzystuje się też blockchain do śledzenia pochodzenia i przekształceń plików (Microsoft ION).

Jednak to wyścig zbrojeń: twórcy fake'ów ciągle ulepszają metody omijania zabezpieczeń.

Dlatego nie wystarczy sama technologia – potrzebne są standardy, uregulowania prawne i edukacja społeczeństwa. To trochę tak, jak z pornografią dziecięcą, czy fałszywymi banknotami - nie powinno być tak, że mimo technologicznej dostępności narzędzi, problem jest marginalizowany przez system prawny.

Nie sposób nie wspomnieć o Dolinie Niesamowitości. Niektórzy eksperci uważają, że m.in. dzięki hiperrealistycznym twarzom generowanym cyfrowo już ją przekroczyliśmy.

„Dolina Niesamowitości" (ang. Uncanny Valley) oznacza, że im bardziej robot albo animacja upodabnia się do człowieka, tym większy dyskomfort odczuwają ludzie – przynajmniej dopóki nie osiągnie prawie idealnego realizmu. W przypadku generowanych cyfrowo twarzy, a nawet głosu, badania pokazują, że wiele osób nie potrafi odróżnić generowanych wizerunków od prawdziwych fotografii.

Najbardziej zaawansowane systemy (np. od Resemble AI czy rodzimego ElevenLabs) są tak przekonujące, że nawet rodzina może się nabrać na podszycie głosu. Mój ojciec, gdy zobaczył nagranie, na którym posługuję się płynnym hiszpańskim, nie mógł wyjść z podziwu - dopiero potem dowiedział się, że do stworzenia nagrania użyliśmy AI, bo hiszpańskim posługuję się na razie na poziomie A1. Można więc powiedzieć, że w pewnych obszarach przekroczyliśmy Dolinę Niesamowitości – a czasem technologie deepfake'owe budzą nawet większe zaufanie niż prawdziwe nagrania (ludzie mylnie sądzą, że skoro coś jest „tak doskonałe", to musi być prawdziwe).

Liczba zagrożeń, jakie się z tym wiążą, wydaje się nieograniczona.

Zdecydowanie największym, z mojego punktu widzenia, jest masowa dezinformacja – modele generatywne tworzą setki tysięcy fałszywych newsów i postów w mediach społecznościowych.

Pokłosiem tego są także oszustwa finansowe – podszywanie się głosowe (deepfake voice) i phishing skutkują miliardowymi stratami. Mamy do czynienia z zaawansowaną inżynierią społeczną – duże modele językowe mogą generować wysoce przekonujące maile, SMS-y czy rozmowy telefoniczne.

Skutkiem są wręcz zmiany polityczne i wygrywanie wyborów przez osoby wsparte machiną dezinformacyjną, co widzieliśmy choćby w Rumunii.

Mam głęboką obawę, że w Polsce także dojdzie do poważnych ataków dezinformacyjnych, a biorąc pod uwagę jasne deklaracje właściciela serwisu X, możemy się spodziewać wręcz ostentacyjnego promowania wspieranych kandydatów, a przycinania zasięgów niewspieranym.

Dużym ryzykiem jest także podważenie zaufania społecznego – jeśli nie wiemy, kto (lub co) stoi po drugiej stronie ekranu, zaczynamy kwestionować każdą treść, a także wątpić w instytucjonalny system wiedzy, co obserwowaliśmy (i badaliśmy z zespołem) choćby w ramach projektu MedFake, analizującego postawy wobec szczepień obowiązkowych.

Mimo obowiązywania trzech praw w powieściach Asimova zdarza się nieoczekiwana utrata kontroli nad zachowaniem robotów. To realny scenariusz?

U Asimova roboty popadały w „paranoję" w wyniku konfliktu pomiędzy prawami. W praktyce współczesne AI nie ma jasno zaprogramowanych „praw", tylko statystyczne modele.

Gary Marcus w książce "Rebooting AI" podkreśla, że takie systemy mogą zachowywać się w sposób nieprzewidywalny, bo opierają się na korelacjach w danych, nie na regułach logicznych. Przykład Microsoft Tay z 2016 r. jest symptomatyczny– chatbot przejął rasistowskie i obraźliwe wzorce wypowiedzi od użytkowników Twittera w ciągu 24 godzin, co pokazało, jak łatwo może dojść do „wymknięcia się spod kontroli", choć nie w sensie buntu, a nieoczekiwanego efektu uczenia maszynowego.

Zatem scenariusz nagłej utraty kontroli jest możliwy, ale głównie dlatego, że systemy AI są często „czarnymi skrzynkami", które obudowujemy topornymi nakładkami dodatkowych ograniczeń.

A jest jakakolwiek szansa, że opracujemy narzędzia, aby odróżniać wytwory AI od ludzkich czy uniknąć utraty kontroli nad maszynami? Czy rozwój AI jest zbyt dynamiczny?

Narzędzia do wykrywania deepfake'ów już powstają (np. projekty DARPA MediFor, prawdopodobnie wykorzystywane przez agencje wywiadowcze; komercyjne rozwiązania Pindrop, Sensity). Wprowadzane są także watermarking i standardy – Adobe (Content Credentials), umowy branżowe, by oznaczać treści AI, i regulacje (AI Act w UE). To kroki w dobrym kierunku.

Warto zauważyć, że nawet w tekście da się wprowadzać znaczniki syntaktyczne, choć trudno będzie cokolwiek finalnie udowodnić.

W ciągu ostatnich trzech lat modele językowe i generatywne przeszły przełom dzięki większej mocy obliczeniowej i dostępowi do danych. Jest to tak szybki postęp, że prawo i edukacja ledwie nadążają.

Nick Bostrom w książce "Superinteligencja" ostrzega, że za dynamicznym rozwojem AI muszą iść środki ostrożności, w przeciwnym razie możemy przegapić moment, w którym systemy staną się trudne do kontrolowania.

Mimo to nie jest tak, że stoimy na straconej pozycji. Kluczem jest wczesne wprowadzanie standardów bezpieczeństwa, audytów i mechanizmów odpowiedzialności (o czym pisze ciekawie choćby Stuart Russell w książce z 2019 roku, "Human Compatible").

Kto obecnie ponosi odpowiedzialność za działania sztucznej inteligencji i maszyn?

Według obecnego prawa (zarówno w Polsce, jak i w większości krajów UE czy w USA) producent oprogramowania/urządzenia może odpowiadać, jeśli AI działa wadliwie (przepisy o odpowiedzialności za produkt). Z kolei operator (np. właściciel robota) jest odpowiedzialny, gdy zaniedbuje nadzór lub używa AI niezgodnie z przeznaczeniem. Użytkownik ponosi winę, jeżeli nie sprawdził wiarygodności danych wygenerowanych przez AI – np. prawnik Steven Schwartz, który bez weryfikacji złożył do sądu pismo z nieistniejącymi precedensami.

To wciąż jednak obszar w dużej mierze nieuregulowany. Ryan Calo w artykule „Robotics and the Lessons of Cyberlaw" wskazuje, że potrzebujemy nowej koncepcji prawnej dla systemów o pewnym stopniu autonomii, bo tradycyjne prawo narzędziowe (traktujące AI jak młotek) może być niewystarczające.

Może potrzebujemy regulacji prawnych na poziomie systemowym?

Unia Europejska wprowadziła AI Act i finalizuje prace nad przepisami pochodnymi, co wprowadzi wymóg transparentności, ocenę ryzyka i klasyfikację systemów AI (niski / wysoki / niedopuszczalny). Polska będzie zobowiązana do wdrożenia AI Act – to oznacza, że producenci i dystrybutorzy systemów AI działających w UE będą musieli zapewnić zgodność z nowymi regulacjami (m.in. testy bezpieczeństwa, dokumentację techniczną, wyjaśnialność algorytmów).

Są też lokalne inicjatywy, np. Ministerstwo Cyfryzacji rozważa przygotowanie szczegółowej strategii AI. Niektóre polskie instytucje (jak Urząd Ochrony Danych Osobowych) wydają wytyczne o ochronie prywatności w kontekście AI, ale to wciąż raczej fragmentaryczne działania.

Moim zdaniem AI to obecnie kluczowy element edukacji od najwcześniejszych etapów. Są podejmowane jakieś inicjatywy?

Są ciekawe projekty unijne – np. „AI4EU" (Horyzont 2020) wspiera materiały edukacyjne dla nauczycieli i uczniów, tłumacząc podstawy uczenia maszynowego. Są liczne warsztaty i hackathony – w Polsce organizowane są inicjatywy m.in. przez GovTech, czy Polską Agencję Rozwoju Przedsiębiorczości, gdzie młodzież uczy się praktycznego podejścia do danych i algorytmów. Coraz więcej uczelni wprowadza zarówno kierunki informatyczne i mechatroniczne, jak i moduły etyki AI, czy kierunki skoncentrowane na łączeniu wiedzy z zakresu zarządzania i AI - wspólnie z prof. Aleksandrą Przegalińską współtworzymy bodaj pierwszy i do dziś jeden z najpopularniejszych programów licencjackich z zakresu zarządzania i AI, realizowany w języku angielskim przez Akademię Leona Koźmińskiego.

Oczywiście są też inicjatywy prywatne - tutaj nie będę obiektywny, bo za absolutną rewelację uważam to, co robi Campus AI, startup z największą rundą finansowania zalążkowego w historii Polski i na szybkiej ścieżce do zostania jednorożcem - ale ponieważ sam jestem w

projekt zaangażowany, trudno mi na niego patrzeć z boku. Edukacja musi na pewno stale nadążać za dynamicznym rozwojem branży. Brian Christian w książce "The Alignment Problem" zwraca uwagę, że to właśnie brak szerokiej świadomości społecznej stwarza największe ryzyko nadużyć technologii AI.

Tylko czy to w ogóle realne? Nie wiemy, co dzieje się w wojskowych ośrodkach badawczych. Do tego, podobnie jak w przypadku każdej technologii, jeśli nawet kraje zachodnie wprowadzą regulacje, oznacza to przewagę Chin.

To realny dylemat – zawsze istnieje obawa, że jeśli „zachód" ograniczy się przepisami, inny gracz (kraj mniej przejmujący się etyką) zyska „przewagę technologiczną".

Jednak historia traktatów międzynarodowych (dot. np. broni biologicznej czy chemicznej) pokazuje, że mimo wszystko warto wprowadzać regulacje i systemy kontroli. Wyścig zbrojeń w AI już trwa i trzeba działać teraz, by nie skończyć w scenariuszu rodem z "Czarnego lustra". W 2019 r. np. USA i Rosja zablokowały na forum ONZ próby zakazania „killer robots", czyli autonomicznej broni ofensywnej. Chiny zadeklarowały częściowe poparcie, ale tylko w określonym obszarze.

Nawet w sferze wojskowej istnieje jednak coraz większa presja opinii publicznej, by wprowadzić jakiekolwiek normy, czego przejawem jest choćby apel naukowców z 2015 r. podpisany przez m.in. Stephena Hawkinga, a nawet Elona Muska, by zakazać autonomicznej broni sztucznej inteligencji.

Skądinąd, autonomiczne bronie AI to coś, co realnie już działa: przecież kiedy ludzki decydent ma trzy sekundy na decyzję, czy zlikwidować osobę, o której AI mówi, że na 97 proc. jest terrorystą, to trudno mówić o jakiejkolwiek kontroli.

Mimo wszystko regulacje i edukacja są nieuniknione, jeśli nie chcemy, by AI wymknęła się spod kontroli zarówno w obszarze cywilnym, jak i militarnym. Kontrola nad AI nie polega na zatrzymaniu postępu - chodzi raczej na zaprojektowaniu ram, w których ów postęp nie obróci się koniec końców przeciwko nam.

Wierzę, że wprowadzenie czwartego prawa robotyki – zakaz podszywania się AI pod człowieka – może się okazać jednym z kluczowych elementów przywrócenia przejrzystości i zaufania w erze deepfake’ów i generatywnych modeli językowych.

Ale samo prawo to nie wszystko. Potrzebujemy także mechanizmów technicznych (watermarking, blockchain, wyłączniki awaryjne), regulacji prawnych (AI Act w UE, międzynarodowych norm) oraz powszechnej edukacji już od najmłodszych lat. Dopiero w tak skoordynowany sposób możemy zabezpieczyć się przed ryzykami i w pełni skorzystać z potencjału AI na polu nauki, medycyny czy gospodarki.

Red. Aleksander Gurgul

© Kasper Kalinowski, jest dziennikarzem w dziale Nauka i Zdrowie. Specjalizuje się w psychologii ewolucyjnej. Autor artykułów naukowych z tematyki atrakcyjności, percepcji ruchu biologicznego u osób chorych na schizofrenię czy mrocznej triady osobowości.