Sztuczna inteligencja – trzy mity, które trzeba rozbroić

23 lutego 2020

Sztuczna inteligencja (AI) ma renomę podobną do atomu. Jedni twierdzą, że jest rozwiązaniem na dręczące ludzkość problemy; drudzy, że zdetronizuje, a nawet zgładzi homo sapiens. Obie skrajne opinie karmią się mitami, które zaciemniają prawdziwy obraz i utrudniają debatę nad tym, jaką rolę sztuczna inteligencja ma pełnić w społeczeństwie i czy są jakieś obszary, z których chcemy wykluczyć jej użycie. Czym jest, a czym nie jest sztuczna inteligencja? Jaką rolę w projektowaniu tych systemów pełnią ludzie? Czy AI może pomóc w rozwiązaniu każdego problemu?

Sztuczna inteligencja nie jest autonomiczna wobec człowieka – jest narzędziem przez niego stworzonym i wykorzystywanym

Sztuczna inteligencja nie jest naturalnym żywiołem, nie rozwija się niezależnie od nas, sama z siebie nic nie robi, sama w sobie nie jest ani dobra, ani zła. O tym, jak taki system działa w praktyce – jakie cele realizuje, na jakie błędy pozwala, na jakie efekty jest „zoptymalizowany" – decydują ludzie. A konkretnie projektanci i właściciele tych systemów, którzy przy pomocy sztucznej inteligencji realizują swoje biznesowe i polityczne cele.

Zdefiniowanie celu jest kluczowym zadaniem dla wszystkich chcących korzystać ze sztucznej inteligencji. Jej właścicieli i projektantów stawia przed pytaniem: po czym poznamy, że nasz system dobrze działa? Sztucznej inteligencji nie wystarczy polecić „zrób tak, żeby było lepiej!" i poczekać na twórczą

interpretację tego zadania. Potrzebne są wskaźniki sukcesu, które da się przetłumaczyć na język matematyki. Czy chodzi przede wszystkim o to, żeby osoby zasługujące na kredyt czy świadczenie społeczne nie odchodziły z kwitkiem? Czy raczej o to, żeby nie zmarnować pieniędzy i nie zainwestować w osoby, które okażą się niewiarygodne albo niewypłacalne?

Na pierwszy rzut oka te dwa cele mogą wydawać się podobne. Ale z punktu widzenia sztucznej inteligencji są zupełnie inne, bo opierają się na różnych funkcjach matematycznych i wymagają sformułowania innych poleceń. Jeśli system ma przede wszystkim unikać błędu w postaci przepuszczenia przez sito niewłaściwej osoby, to oznacza, że sito musi być bardzo gęste i nieraz nie przepuści również człowieka, który spełnia kryteria. Jeśli zależy nam na tym, żeby ludzie, którzy zasługują na kredyt czy świadczenie, nie byli pomijani, musimy wybrać sito o większych oczkach. I jednocześnie pogodzić się z tym, że czasem przejdzie przez nie również ktoś, kto naszych kryteriów nie spełnia, ale wydaje się podobny.

Statystyka pomaga nam ograniczyć prawdopodobieństwo wystąpienia sytuacji, którą uważamy za niepożądaną. Ale błędów nie zredukuje do zera. W przypadku systemów, które wpływają na życie i prawa ludzi, decyzja o tym, jakie błędy są preferowane i jakie polecenia dostaje uczący się system, jest polityczna – nie techniczna.

Sztuczna inteligencja nie jest czarną skrzynką – możemy zrozumieć, jak działa

W rozmowie o odpowiedzialności za skutki działania systemów wspieranych przez AI bardzo przeszkadza pokutujący mit czarnej skrzynki. W medialnym dyskursie takie systemy są najczęściej przedstawiane jak magiczne pudełka, których zwykły śmiertelnik nie może ani otworzyć, ani pojąć (gdyby jednak otworzył).

Jednocześnie słyszymy zapowiedzi, że „AI będzie decydować o naszym życiu”: zatrudnieniu, leczeniu, kredycie. To realna, a zarazem wymykająca się społecznej kontroli władza, której ludzie słusznie się obawiają. Szczególnie że z głośnych przykładów – takich jak amerykański system COMPAS wspomagający decyzje sędziów – zdążyli się dowiedzieć, że systemy wykorzystujące AI popełniają tragiczne w skutkach błędy i wzmacniają istniejące w społeczeństwie uprzedzenia.

Ten kryzys zaufania może się rozlać bardzo szeroko, na każde zastosowanie analizy danych do rozwiązywania społecznych problemów. Może, ale nie musi. Systemy wykorzystujące sztuczną inteligencję wcale nie muszą być projektowane na zasadzie czarnej skrzynki – w sposób, który nie pozwala zrozumieć ostatecznego rozstrzygnięcia ani odtworzyć czynników, które miały na nie wpływ. Projektanci tych systemów mają do dyspozycji różne klasy i typy modeli, w tym takie, które jak najbardziej „da się otworzyć”, a nawet zilustrować ich działanie na wykresach zrozumiałych dla przeciętnego odbiorcy. Mogą sięgnąć po liniową lub logiczną regresję albo nieco bardziej rozbudowane drzewo decyzyjne. Co więcej, bardzo często właśnie to robią: decydują się na korzystanie z mniej wyrafinowanych, poddających się interpretacji modeli, ponieważ są one dość dobre, szybsze i tańsze niż złożone i mniej przejrzyste sieci neuronowe.

Owe tradycyjne modele statystyczne były używane przez dekady: w naukach społecznych i przyrodniczych, do przewidywania trendów na giełdzie, w badaniach medycznych i analizie tekstu, tylko dziś mówi się o nich innym językiem (tj. „sztuczna inteligencja”) i podkreśla ich możliwości predykcyjne. Same modele niekoniecznie stały się przez ten czas lepsze, ale z pewnością mamy dziś do dyspozycji więcej danych i więcej doświadczeń z ich zastosowaniem w codziennym życiu. To dodatkowy argument, żeby śmielej korzystać z dobrze rozpoznanych modeli zamiast eksperymentować z sieciami neuronowymi.

Taką właśnie radę projektanci systemów opartych na sztucznej inteligencji znajdą też w wytycznych, które przygotował brytyjski urząd do spraw ochrony danych (ICO) we współpracy z renomowanym Instytutem Alana Turinga. Punktem wyjścia dla tego projektu był art. 22 RODO, który wymaga wyjaśniania logiki automatycznie podejmowanych decyzji. Brytyjski organ uznał, że zanim zacznie kontrolować i karać tych, którzy stosują czarne skrzynki do podejmowania decyzji, za brak ww. wyjaśnień, najpierw powie im, jak powinni postępować. W swoim poradniku podpowiada więc, jak w praktyce wyjaśnić, w jaki sposób działa „AI": jakie czynniki bierze pod uwagę, jakie pomija i na jakie efekty została „zoptymalizowana".

Właściciele i projektanci systemów znajdą też w nim propozycje pytań, na jakie powinni sobie odpowiedzieć, zanim sięgną po wyrafinowany, trudny do zinterpretowania model, tak by podjęta decyzja była odpowiednio przemyślana i uzasadniona. Jeśli projektujący system nie potrafią ocenić ryzyka, jakie się wiąże z rzuceniem sieci neuronowej na nierozpoznany jeszcze obszar, jeśli nie potrafią przewidzieć i zminimalizować błędów, najprawdopodobniej w ogóle nie powinni wchodzić w taki eksperyment – szczególnie jeśli w grę wchodzą poważne decyzje dotyczące ludzi.

Jednocześnie Brytyjczycy zwracają uwagę w swoim poradniku, że nawet trudna w rozszyfrowaniu sieć neuronowa nie jest do końca czarną skrzynką. W tym przypadku nie możemy wprawdzie dokładnie rozszyfrować, jak działa model, ale możemy zrozumieć, na jaki efekt został zoptymalizowany. Chodzi o to, jakie zadanie zostało postawione sztucznej inteligencji i jaki efekt w rzeczywistości ma przynieść jej działanie. Tę decyzję zawsze da się z systemu wydobyć i zawsze można o niej dyskutować. A jest ona kluczowa pod kątem zrozumienia, jakie ryzyka jej działanie niesie dla ludzi.

Sztuczna inteligencja nie sprawdzi się równie dobrze w każdej dziedzinie życia – w niektórych zadaniach radzi sobie lepiej, w innych gorzej od człowieka

Wbrew nazwie nie mamy wcale do czynienia z inteligentną maszyną, tylko z narzędziami do zaawansowanej analizy statystycznej. Jak podpowiada sama nazwa, takie narzędzia najlepiej sobie radzą z wykrywaniem wzorów i prawidłowości (tzw. korelacji statystycznych). Takie wzorce obiektywnie istnieją w języku, naukach ścisłych, przyrodzie. Umiemy odróżnić kota od psa na zdjęciu. I sztuczna inteligencja też może się tego nauczyć. Ale wiele zjawisk społecznych i zachowań człowieka wciąż się jej wymyka. Trudniej bowiem stwierdzić, czym się różni człowiek dobry od złego albo podejrzany od normalnego.

Arvind Narayanan, prof. nauk komputerowych z Princeton, znany z demaskowania systemów AI, które obiecują coś, czego nie da się zrobić, wyróżnia trzy podstawowe cele, jakim może służyć sztuczna inteligencja. AI może wspierać naszą percepcję, automatyzować oceny i przewidywać skutki społeczne. Jego zdaniem sztuczna inteligencja coraz lepiej sobie radzi z tym pierwszym zdaniem (wspieraniem percepcji), ponieważ jest zatrudniona do wykrywania obiektywnie istniejących wzorów (np. klasyfikowania obiektów na zdjęciach, wykrywania zmian chorobowych na zdjęciach rentgenowskich, tłumaczenia mowy na tekst).

Dużo trudniejsze dla AI – i bardziej kontrowersyjne – jest zadanie polegające na zautomatyzowaniu oceny (np. wykryciu przypadków mowy nienawiści w sieci, odróżnieniu osoby chorej od zdrowej, dostosowaniu rekomendacji treści do profilu czytelnika). Przede wszystkim dlatego, że w przypadku oceny

nie ma jednoznacznie poprawnej odpowiedzi, której system może się nauczyć. Ludzie w takich przypadkach też się mylą lub różnią w opiniach, ale łatwiej im zinterpretować kontekst, który może mieć gigantyczne znaczenie. Zdaniem Arvinda Narayanana te systemy nigdy nie będą doskonałe i potrzebujemy dla nich gwarancji prawnych, chroniących ludzi przed błędnymi decyzjami.

Wreszcie, trzecia kategoria – systemy, które mają za zadanie przewidywać przyszłość – które zdaniem Narayanana wiążą się z największym ryzykiem. Jest zasadnicza różnica między wykorzystywaniem sztucznej inteligencji do wykrywania wzorów i prawidłowości, które obiektywnie istnieją i dają się matematycznie opisać, a zatrudnianiem jej do szukania wzorów i prawidłowości tam, gdzie ich nie ma albo bywają nieregularnie. Łatwiej przewidzieć trend niż to, co zrobi konkretny człowiek. Da się przewidzieć wzrost liczby samochodów na drogach albo wzrost zachorowań na grypę jesienią. Mamy reprezentatywne dane i potrafimy zadać pytanie o obiektywnie istniejący problem. Znacznie trudniej przewidzieć, kto popełni przestępstwo, komu opłaca się udzielić pomocy społecznej, kogo warto zatrudnić. A mimo to nadal próbujemy to zrobić przy pomocy sztucznej inteligencji. Dlatego systemy, które wspierają decyzje na temat konkretnych osób (np. sędziów w sprawach karnych, urzędników przyznających pomoc społeczną, analityków w bankach i HR-owców w dużych korporacjach), budzą słuszne kontrowersje.

Konkluzja? Przyjmijmy zasady działania sztucznej inteligencji i poddajmy to działanie kontroli społecznej!

Rozumiejąc, że mityczna sztuczna inteligencja jest w rzeczywistości sumą decyzji człowieka, możemy – zamiast snuć futurystyczne rozważania – zacząć rozmawiać o konkretach.

Zacznijmy od przyjęcia zasad dla tych obszarów zastosowania zaawansowanych metod analizy danych, które już okazały się problematyczne. Od reguł, jakich mają przestrzegać autonomiczne pojazdy, przez profilowanie treści w Internecie i ocenę ryzyka w usługach finansowych, po „optymalizowanie" polityk i usług publicznych. Ich cechą wspólną jest to, że sztuczna inteligencja wspiera decyzje, które będą miały poważne skutki dla ludzi. Innych ludzi niż ci, którzy zaprojektowali system.

Bardzo dobrym narzędziem służącym do takiej analizy jest ocena oddziaływania systemu, którą przeprowadza się we wczesnej fazie projektowania – tzw. impact assessment. Pomysł w żadnym razie nie jest nowy. Stosuje się go w procesie tworzenia prawa oraz jako ocenę ryzyka związanego z przetwarzaniem danych. Algorytmy przypominają reguły prawne, tylko zapisane w języku matematyki. Skutki ich oddziaływania na ludzi da się i trzeba oceniać. Trudniej jest w przypadku systemów uczących się, które mają wyznaczony cel, ale nie posługują się z góry zdefiniowanymi regułami. To, że jest trudniej, nie jest jednak żadną wymówką. Jeśli system ma oddziaływać na innych ludzi, ludzie go projektujący muszą włożyć wysiłek w analizę ryzyka i przewidzenie negatywnych skutków, zanim one się wydarzą. To może wymagać modelowania, testowania, ewaluacji – i wielu kosztownych i czasochłonnych zabiegów, do których nikt w biznesie ani nawet w administracji publicznej dziś się nie wyrywa. Jednak nie ma innej drogi, jeśli te systemy mają być tworzone odpowiedzialnie.

Autorstwo: Katarzyna Szymielewicz
Źródło: Panoptykon.org

Bibliografia

1. https://medium.com/@szymielewicz/black-boxed-politics-cebc0d5a54ad

2. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

3. https://ico.org.uk/media/about-the-ico/consultations/2616433/explaining-ai-decisions-part-2.pdf

4. https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

5. https://panoptykon.org/sites/default/files/publikacje/panoptykon_rodo_praktyczny_poradnik_22.01.2018.pdf