# Andrzej Dragan o konsekwencjach AI: Jednak zmieni się wszystko



Gazeta Wyborcza, 14.12.2025

https://wyborcza.pl/7,75517,32447140,andrzej-dragan-o-konsekwencjach-ai-jednak-zmieni-sie-wszystko.html

*Andrzej Dragan (ur. w 1978 r.) jest profesorem Uniwersytetu Warszawskiego i profesorem wizytującym w Narodowym Uniwersytecie Singapuru. Zajmuje się szukaniem związków między teorią względności i nowym opisem świata, jaki przyniosła fizyka kwantowa. Jest również fotografem, a także popularyzatorem nauki, autorem książki „Kwantechizm, czyli klatka na ludzi", w której opisuje Wszechświat i życie z punktu widzenia fizyki kwantowej i teorii Einsteina.

**Zdumiewa mnie, że są jeszcze ludzie, których to nie rusza. Że nie są zaskoczeni tym, co się dzieje.**

**Piotr Cieśliński: Jest taki cytat w twojej nowej książce: „Trudno jest mówić, dokąd to wszystko doprowadzi, żeby nie zabrzmieć jak kompletny wariat". Co miałeś na myśli?**

**Andrzej Dragan***: Trzeba by Sama Altmana, szefa OpenAI, pytać, bo to jego myśl. No ale czy nie ma racji?Gdyby pięć lat temu ktoś mi powiedział, że dziś wystarczy krótki prompt, żeby AI stworzyła film czy obraz trudny do odróżnienia od rzeczywistości, tobym przecież popukał się w głowę.

To, że generatywna sztuczna inteligencja się ludziom jakimś cudem udała, jest naprawdę niewiarygodne. Mustafa Suleyman w książce „Nadchodząca fala" twierdzi, że „w nadchodzących dziesięcioleciach" będziemy promptować nie tylko obrazki czy filmy, ale też związki chemiczne o nowych właściwościach, a może nawet całe organizmy.

**Brzmi to...**

–...niedorzecznie. I trudno uniknąć myśli, że to tylko hajp, tym bardziej że mówi to szef działu AI w Microsofcie – ale z drugiej strony to w końcu jeden z najpoważniejszych ludzi w tej branży. Tak czy siak, przewidywanie, co się za parę lat wydarzy, przestało mieć sens, przewidywalna przyszłość się już skończyła.

**Widać, że jesteś tą technologią zafascynowany.**

– A jak miałbym nie być? Bo to „tylko matematyka"? Cała znana mi rzeczywistość, która jest przecież fascynująca, działa na podstawie praw fizyki napędzanych „tylko matematyką".

Zdumiewa mnie, że są jeszcze ludzie, których to nie rusza. Że nie są zaskoczeni tym, co się dzieje.

**A co się właściwie dzieje?**

– Chociażby to, że nie trzeba nawet umieć grać w szachy, żeby zaprojektować sieć neuronową, która po jednej nocy uczenia przez wzmocnienie, grając sama ze sobą, ogra mistrza świata. Zresztą taka sieć ogra nie tylko mistrza, ale każdy klasyczny algorytm zaprojektowany przez arcymistrzów szachowych. Nic dziwnego, że wszystkie najsilniejsze silniki szachowe, takie jak Stockfish (począwszy od wersji bodaj 16.), działają już w oparciu o metody ewaluacji zaprojektowane nie przez ludzi, tylko przez sieci neuronowe.



Inny przykład: grupa programistów postanowiła znaleźć związek pomiędzy trójwymiarową strukturą białek a sekwencją aminokwasów, która ją tworzy – jeden z najważniejszych problemów biologii molekularnej, z którym naukowcy męczyli się pół wieku. Model sieci neuronowej AlphaFold w zasadzie go rozwiązał.

**Za to właśnie przyznano w zeszłym roku Nobla z chemii.**

– Już jak zobaczyłem ich wyniki, to prawie się zakładałem, że dostaną Nobla, tylko że z medycyny.

**Jeden z laureatów, Demis Hassabis z DeepMind, powiedział, że nie musimy rozumieć, co dzieje się w środku AI – tak jak nie musimy rozumieć mechaniki kwantowej, żeby z niej korzystać.**

– To nie tak, że jej nie rozumiemy. Jakbyśmy jej rzeczywiście nie rozumieli, toby nas nie dziwiła.

My mechanikę kwantową rozumiemy całkiem nie najgorzej i właśnie dzięki temu dostrzegamy jej mankamenty, bo teoria ta rzeczywiście jest dziurawa. A niektórzy idą dalej i nie mogą się z nią po prostu pogodzić, bo wydaje im się zbyt absurdalna. O ile wiem, Hassabis podziela pogląd Einsteina, któremu nie podobała się niedeterministyczność tej teorii.

**Może mechanika kwantowa rzeczywiście jest niepełna?**

– Byłbym zaskoczony, gdyby cokolwiek z tego, co obecnie nam się wydaje, było prawdą ostateczną. Czy nie byłoby dziwne, gdybyśmy w zaledwie czterysta lat od narodzin współczesnej nauki dotarli do jakiejkolwiek ostatecznej prawdy o świecie?

W zasadzie to co my właściwie rozumiemy tak do końca? Mamy nawet problem z tym, żeby wytłumaczyć, dlaczego stół jest twardy albo jak działa młotek.

**Nawet tego nie potrafimy wyjaśnić?**

– Ja przynajmniej nie potrafię. Są „zabawkowe" modele oddziaływań elektromagnetycznych, które najpewniej odpowiadają za twardość młotka, ale to tylko skrajne uproszczenia.

Więc tak – korzystamy z rzeczy, których nie rozumiemy, i nam to nie przeszkadza. I zawsze tak było. Ogień odkryliśmy zapewne przypadkiem, a kto z palących papierosy wie, czym jest ogień? Penicylinę – przypadkiem. Mikrofalowe promieniowanie tła po Wielkim Wybuchu – też przypadkiem, jego odkrywcy sądzili, że to tylko efekt zanieczyszczeń pozostawionych przez gołębie na antenie radiowej. Anestezjolodzy stosują środki, po których błyskawicznie tracimy świadomość – ale mechanizmu tego zjawiska też nie rozumiemy. Fajnie byłoby pojąć, jaka to fizjologiczna blokada naszego mózgu może nas odciąć od świadomości.

Tak często funkcjonuje nauka: najpierw odkrywamy, że coś działa, a potem mozolnie próbujemy zrozumieć dlaczego.

Ze sztuczną inteligencją i z tym niezrozumieniem jest jeszcze zabawniej – to kolejny poziom.

*Demis Hassabis – a on należy do ludzi, których warto słuchać – mówi, że może w krótkim horyzoncie czasowym AI jest przehajpowana, ale w długim – skrajnie niedohajpowana. I martwi go, jak bardzo nie doceniamy skutków w perspektywie dekad.*

**Mój syn, który jest programistą, zapytał: „Dlaczego mielibyśmy wierzyć Draganowi w to, co mówi o sztucznej inteligencji? Przecież on nie jest specjalistą od AI, tylko fizykiem".**

– I ma rację! Ja wcale nie oczekuję, żeby mi wierzyć w cokolwiek. „Kwantechizm" też napisany był tak, żeby nie trzeba było wierzyć autorowi, bo zamiast oznajmiać, „jak jest", serwowałem uzasadnienia i eksperymenty myślowe pokazujące, skąd się wzięło nasze pozornie absurdalne wyobrażenie o świecie kwantowym, nazywane „mechaniką kwantową". O wiarę czytelników nie prosiłem.

Nie bez powodu w Instytucie Fizyki Teoretycznej nie używamy tytułów naukowych na tabliczkach przed drzwiami, bo odwoływanie się do „autorytetu" w dyskusji o fizyce byłoby porażką.

I podobnie skonstruowana jest „Quo vAIdis": tak, żeby w nic nie trzeba było autorowi wierzyć. A zamiast opinii, które uważam za średnio wartościowe, są różne eksperymenty myślowe i testy. Czytelnik może je sobie prześledzić i wyciągnąć samodzielne wnioski. A jak ktoś ma już swoje wyobrażenia i opinie o sztucznej inteligencji – to jeszcze lepiej, będzie okazja je zweryfikować.

**Zaczynasz książkę od zdania, że fizyka to najskuteczniejsza metoda docierania do** prawdy o wszystkim, co ciekawe.

– Przynajmniej najskuteczniejsza, jaką dotąd odkryliśmy.

**Skąd ta pewność?**

- Widać to chociażby po efektach ubocznych procesu naukowego, czyli po rozwoju technologii.

A dlaczego fizyka i nauki empiryczne są tak skuteczne? Bo w pewnym momencie dotarło do ludzi, że jesteśmy zbyt głupi, żeby budować wizję świata wyłącznie na swoich opiniach. Naturalną konsekwencją był wniosek, żeby wszystko sprawdzać. I nagle

okazało się, że większość wcześniejszych wyobrażeń o świecie po prostu nie trzyma się kupy. Odrzucając je, zrobiliśmy pierwszy krok do głębszego zrozumienia.

**W „Kwantechizmie" opisywałeś różne domowe eksperymenty – choćby pomiar siły nośnej muchy albo testy, czy mrówki naprawdę potrafią biegać.**

– W nowej książce jest to samo, tylko zamiast much i mrówek są modele językowe i tym podobne. Weźmy popularne przekonanie, że ChatGPT – zwłaszcza jego starsza generacja – to tylko statystyka słów, zwykłe korelacje między wyrazami i nic poza tym. To rzeczywiście ciekawa opinia, warta sprawdzenia. Przetestujmy ją, zadając modelowi pytanie złożone ze słów, które w ogóle nie istnieją. No i zobaczymy, co się stanie.

Takich eksperymentów jest w książce sporo. I co ciekawe, najbardziej wartościowe są nie te, przy których model radzi sobie dobrze, ale dopiero te, przy których się wywala. Bo pomyłki mówią o wiele więcej o mechanice modeli i o słabościach ich konstrukcji. Czy to prawdziwie rozumująca inteligencja, jak twierdzi twórca tej technologii Geoffrey Hinton, czy tylko recytująca papuga, jak upiera się Yann LeCun, inny wybitny przedstawiciel branży?

I właśnie o tym jest parę rozdziałów. A jako że obaj panowie nie mogą mieć racji, to zawartość bzdury w tej całej szamotaninie wynosi co najmniej 50 proc.

**Twierdzisz, że matematycznie uczenie maszynowe jest znacznie prostsze niż fizyka.**

– Aż tak twierdzę? Wiedza inżynierów od uczenia maszynowego jest bardzo rozległa, natomiast sama matematyka, której potrzeba, żeby zrozumieć podstawy treningu sieci, to rzeczywiście materiał z pierwszych paru tygodni pierwszego semestru algebry liniowej i analizy matematycznej. Matematycy czasem sobie kpią, że programiści nie muszą nawet odróżniać tensora od macierzy.

**Skoro to tak proste narzędzia, dlaczego twórcy AI przyznają, że nie wiedzą, co dzieje się w środku ich modeli?**

– Nie powiedziałbym, że aż tak proste. Algorytm propagacji wstecznej, za który w zeszłym roku Geoffrey Hinton dostał Nobla z fizyki, można co prawda wytłumaczyć studentowi pierwszego roku w 20 minut, ale to przecież tylko sam początek, na którym nabudowana jest cała masa dodatkowych struktur, często bardzo złożonych.

Dlatego gdy Dario Amodei, współtwórca OpenAI i Anthropica, mówi, że nie mamy pojęcia, jak działa AI, to ma na myśli nie algorytmy treningowe, które ich twórcy doskonale rozumieją, bo je przecież sami projektują, lecz mechanizm działania już wytrenowanych sieci.

**Piszesz, że to podobnie jak z ewolucją i jej wytworami.**

– Tak, bo niektóre algorytmy treningowe, jak uczenie przez wzmocnienie, współodkryte przez Richarda Suttona, są wprost inspirowane ewolucją. Mechanizmy ewolucyjne rozumiemy przecież nie najgorzej, natomiast słabo rozumiemy budowę organizmów, które są produktem ewolucji. Nawet do pełnego zrozumienia pojedynczej komórki wciąż bardzo nam daleko.

Z tym że trening oparty na uczeniu nadzorowanym to bardziej nauka imitowania określonego zbioru danych. Na przykład pierwsze modele językowe uczone były imitowania tekstów stworzonych przez ludzi.

**Czyli one rzeczywiście tylko „papugują" ludzi, jak twierdzi wielu krytyków tej technologii?**

– O tym są ze trzy rozdziały książki, więc nie chciałbym teraz wszystkiego spojlować, ale jednoznaczna odpowiedź „tak" lub „nie" byłaby zbyt naiwna. Według mnie najciekawszą umiejętnością modeli językowych, która wykracza poza „tylko papugowanie" wiedzy z treningu, jest zdolność dostrzegania analogii. Dzięki niej modele potrafią rozwiązywać zadania, których nigdy wcześniej nie widziały, a które jedynie odrobinę przypominają dane z treningu.

**W książce opisujesz eksperyment, w którym kazałeś modelowi narysować literę A, choć nie był trenowany na obrazach.**

– A jednak ją narysował. Okazuje się, że nawet modele trenowane wyłącznie na tekście potrafią rysować, choć nie były do tego zaprojektowane. Jest wiele szalenie intrygujących przykładów pokazujących, że modele często wykraczają poza kompetencje, do których zostały stworzone.

**Tak jak twój własny, stary projekt, o którym wspominasz w książce – „ludziki" żyjące w wirtualnym świecie.**

– Zanim zająłem się fizyką, łaziło mi po głowie, żeby być programistą. Jako nastolatek kodowałem w Asemblerze, potem w Pascalu, C++, Fortranie... W latach 90. pisałem proste programy oparte na idei algorytmów genetycznych. Te moje „ludziki" miały symulować uproszczone procesy ewolucyjne, chciałem przetestować mechanizmy selekcji naturalnej.

Wnioski, nawet w moim skrajnie uproszczonym modelu, były zaskakująco podobne do tego, co obserwujemy w naturze. „Ludziki" same odkrywały sprytne strategie, na które bym nie wpadł. Był to dla mnie pierwszy zaskakujący przykład tego, że programista nie musi wiedzieć, co jego program będzie robił po uruchomieniu. A mówimy o latach 90.

**Twoje „ludziki" potrafiły się nawet poświęcać, choć ich celem było przetrwanie.**

– Wyłoniło się coś w rodzaju prymitywnej strategii altruistycznej.

**Stąd twoja konkluzja: „inteligentnie wyglądające" strategie mogą powstać same.**

– Tak – w odpowiednich warunkach.

*Ale pamiętajmy, że coś, co wygląda na inteligencję, nie musi nią być. Jeśli organizm ma zmyślną strategię przetrwania, to przecież nie znaczy, że jest inteligentny, skoro nie ma nawet mózgu.*

**Jak przywra, o której piszesz. Ten prymitywny organizm atakuje ślimaka i paraliżuje jego układ nerwowy, żeby podstępem zwabić ptaka i dostać się do jego organizmu.**

– Właśnie. Dlatego w książce dużo miejsca poświęcam pytaniu, czy zachowania AI – w tym modeli językowych – naprawdę zasługują na miano inteligencji. Wielu ludzi mówi, że to tylko niezła imitacja. To rozsądna wątpliwość i nie ma co jej zbywać. Mamy zachowanie, które wygląda inteligentnie. Pytanie: czy to tylko symulacja, czy coś więcej?

Dla mnie to kluczowe, bo pracuję w branży, w której płaci się za rozumienie rzeczy. Jak do mnie przychodzi student na egzamin ustny, to nie sprawdzam, czy wykuł podręcznik, bo to mało kogo obchodzi, tylko czy rozumie to, co mówi.

**Które z twoich eksperymentów z modelami AI zrobiły na tobie największe wrażenie?**

– W książce jest cały rozdział o tym, jak działają modele: opisuję transformera, sam proces treningu, wektoryzację języka, przestrzeń osadzeń – tę „przestrzeń", w której „żyją" pojęcia językowe, na których operują modele. To oczywiście uproszczone, popularnonaukowe wersje, ale idea powinna być klarowna.

I właśnie to jest najbardziej zdumiewające – że trywialny nieomal algorytm treningowy i dość intuicyjna konstrukcja transformera mogą doprowadzić do niewiarygodnie złożonych kompetencji wytrenowanej sieci. Już samo to, że do badania modeli językowych wzięli się psychologowie, jest fascynujące.

**Nowe modele AI coraz częściej stosują „namysł" nad odpowiedzią, zamiast iść na skróty. Zastępują szybkie skojarzenia – to, co u ludzi Daniel Kahneman nazywał umysłowym systemem 1 – rozumowaniem krok po kroku, czyli odpowiednikiem naszego systemu 2.**

– W dużym uproszczeniu. Do połowy zeszłego roku modele były trochę jak Kahnemanowski system 1 „na sterydach": działały błyskawicznie, często bez autorefleksji – jak my, gdy mówimy bez zastanowienia, ile to 5×5. Z tym że one potrafiły w tym trybie rozstrzygać znacznie trudniejsze rzeczy. Teraz twórcy modeli zaczęli wdrażać coś na kształt „systemu 2" – wolniejszego, z autorefleksją, poprawianiem własnych odpowiedzi, pętlami planowania i korekty. A do tego modele te trenowane są w trzeciej fazie właśnie uczeniem przez wzmocnienie. A więc mogą wyjść poza wyłącznie imitowanie ludzi.

Efekt? Już mamy modele, które dzięki temu „rozumowaniu", tym wariantom systemu 2, dostają złote medale na Międzynarodowej Olimpiadzie Matematycznej, rozwiązując oryginalne problemy, których nie ma w żadnych podręcznikach. Przypominam: jeszcze dwa lata temu złośliwcy się śmiali, że model nie umie policzyć liter „r" w słowie „truskawka". Postęp jest ekstremalnie szybki.

Z drugiej strony te modele wciąż mają poważne słabości. Potrafią popełnić zupełnie głupie błędy – w książce pokazuję masę przykładów, jak w prosty sposób skompromitować nawet najlepszy dziś model.

**A w fizyce? Próbujesz wykorzystywać AI w badaniach?**

– Nieśmiało. W fizyce teoretycznej często trzeba udowodnić jakieś twierdzenie, znaleźć powiązanie między równaniami, coś, co wymaga czystej analizy matematycznej. I w tym modele okazują się bardzo pomocne. Potrafią wychwycić rzeczy, które umykają studentom, a czasem nawet rozwiązać problem, który wymagałby długiego, często nieskutecznego kombinowania. Mówię oczywiście o najlepszych modelach, nie o publicznie dostępnych darmowych ogryzkach.

**Czy dopuszczasz, że AI rozwinie się tak, że przestaniemy ją rozumieć?**

– To się już dzieje – tylko w małej skali. Weźmy silniki szachowe. Czasem wykonują ruch, który wygląda absurdalnie. Arcymistrz drapie się w głowę i mówi: „Nonsens". A po kilkudziesięciu ruchach okazuje się, że to był genialny plan.

Kiedy ta jakość wkroczy do bardziej złożonych dziedzin, jak matematyka czy fizyka? Mój kolega Piotrek Sankowski jest przekonany, że wkrótce.

Co ciekawe, programy szachowe oparte na sieciach neuronowych nie tylko biją już ludzi, ale pokonują też klasyczne algorytmy napisane przez ludzi.

*Stockfish 8, który był jednym z ostatnich silników szachowych zaprogramowanych klasycznymi metodami, analizował 70 mln ruchów na sekundę, ale przegrał z siecią neuronową AlphaZero, która analizowała… ledwie 80 tys.*

Nie chodzi więc o brutalną moc obliczeniową, tylko o właściwy wybór ruchów do analizy.

Jeśli mózg jest maszyną obliczeniową, to dużo bardziej skomplikowaną niż nasze sztuczne sieci neuronowe. Co wcale jednak nie znaczy, że musimy go dokładnie imitować, żeby zbudować coś myślącego. Samoloty nie machają skrzydłami jak ptaki, są o wiele prostsze, a jednak latają.

Jeśli celem jest rozumowanie, nie trzeba kopiować mózgu – zapewne można to osiągnąć inną drogą, nawet uproszczoną. Sztuczna inteligencja nie musi „machać skrzydłami" jak mózg.

**Rozmawiamy o rozumowaniu, inteligencji... A świadomość? Twoim zdaniem dzisiejsze modele językowe, nawet te „rozumujące", na pewno jej nie mają.**

– Roger Penrose twierdzi, że mózg jest urządzeniem fizycznym, ale takim, które działa dzięki prawom wykraczającym poza znaną fizykę. Słowem, gdzieś w naszym mózgu zachodzą procesy, których wciąż nie odkryliśmy i nie rozumiemy, i to z nich rodzi się świadomość.

**On sugeruje, że te zjawiska zachodzą w białkowych strukturach w mózgu, zwanych mikrotubulami. Czy to jednak nie przypomina dawnych prób tłumaczenia życia jakąś nieznaną „siłą życiową"?**

– Takie mam wrażenie, jakby to była znów ta sama tęsknota za „czymś więcej", za magiczną substancją, która wykracza poza to, co znamy. Argumentum ad microtubulum, jak sobie drwi mój kolega ze szwajcarskiego Google'a Robert Święcicki.

Nie znamy zresztą nawet warunków wystarczających do tego, by pojawiła się świadomość. Znamy co najwyżej niektóre warunki konieczne. Na przykład zdolność obserwowania własnych stanów wewnętrznych – coś na kształt autorefleksji. I ten prosty warunek nie jest spełniony w przypadku prostych modeli językowych ani, szerzej, jednokierunkowych sieci neuronowych.

**Bo te proste sieci nie „cofają się" i nie „obserwują" własnych stanów?**

– Transformery ograniczają się do niewielkiego „globalnego" kontekstu, taką mają architekturę. Ale do głębszej autorefleksji to wciąż dużo za mało. Są prace nad sieciami wielokierunkowymi – ostatnio ciekawą propozycję pokazał start-up Pathway, właśnie idącą w stronę architektury niejednokierunkowej i uczącej się w sposób ciągły.

**Obecna AI nie ma kontaktu z realnym światem. Jest karmiona tekstami i zdjęciami. Czy to jej nie ogranicza? Może przełom nastąpi dopiero wtedy, gdy pozwolimy modelom eksperymentować w rzeczywistości?**

– A ty masz kontakt z rzeczywistością? Widziałeś kiedyś papieża? Widziałeś fotony z ekranu, na którym „był" papież. „Dotykasz" świata sygnałami, które wpadają przez kilka otworów do pudełka z kości, w którym mieszkasz. Tak działa nasze „poznanie świata". Oczywiście, strumień danych, którymi trenuje się AI, jest dziś wciąż ubogi: litery, piksele, wciąż mało.

Jest jednak inny, poważniejszy problem – obecne modele AI nie uczą się w trakcie rozwiązywania zadania. My uczymy się „w biegu", one – tylko w fazie treningu. To próbuje zmienić propozycja Pathwaya.

Innym ciekawym kierunkiem byłoby to, co proponuje Andrej Karpathy – żeby odchudzać modele z pamięci. Trening miałby się odbywać według kryteriów, które sprawdzają wyłącznie umiejętność rozumowania, a nie pamiętanie.

Może takie wyizolowanie tej konkretnej jednej umiejętności mogłoby dać nową jakość?

**Szukamy esencji czystego rozumu?**

– Może tak, ale nie wiemy, jak to zrobić, bo trudno jest powiedzieć, w którym zakresie parametrów w modelach językowych kończy się wiedza, zasób danych, a gdzie się zaczyna rozumowanie. Jest to dość mocno wymieszane.

W zasadzie dwie trzecie parametrów modeli językowych to są wielowarstwowe perceptrony, o których mówi się w uproszczeniu, że są zbiorowiskiem danych zapisanych w treningu, czyli wiedzą o świecie. Pewnie tylko niewielka część parametrów koduje coś, co można by nazwać umiejętnością rozumowania, ale nie bardzo chyba nawet wiadomo, które to są.

**Sugerujesz, że nie poprawimy tych modeli, karmiąc ich większą ilością naszych danych, bo one już „pożarły" całą zapisaną wiedzę. Wszystko, co mogliśmy, wpakowaliśmy w nie, więc teraz musimy je po prostu mądrze wyszkolić w rozumowaniu?**

– Richard Sutton, laureat Nagrody Turinga, zjadliwie krytykuje modele językowe i generatywną sztuczną inteligencję za to, że są trenowane na danych dostarczonych przez człowieka. Bo uważa, że to jest droga, która musi się w którymś momencie skończyć – i właśnie się kończy. Musimy więc szukać sposobów trenowania AI nie na podstawie istniejących tekstów dostarczonych przez człowieka, tylko np. na podstawie danych syntetycznych, które AI tworzy sobie sama.

**Czyli AI ma się sama uczyć rozumowania – bez udziału człowieka?**

– Trochę to się już dzieje. Niedawno DeepMind stworzyło algorytm do rozwiązywania problemów z matematyki – z geometrii dwuwymiarowej – przy zerowej ilości danych treningowych od ludzi. AI nauczyła się rozwiązywać zadania z geometrii na poziomie olimpijskim bez znajomości podręczników, zaczynając naukę od zera. Sama wymyślała

sobie przykładowe problemy i uczyła się na nich reguł geometrii, odkrywając je samodzielnie.

A w zeszłym roku weszliśmy w fazę wdrażania tych samych technik do uczenia modeli językowych. I to przynosi już teraz fantastyczne efekty – złoty medal z olimpiady matematycznej dla dwóch modeli językowych, a nie dla wyspecjalizowanego systemu, który zajmuje się tylko matematyką. A przecież dopiero co zaczęliśmy.

*Łukasz Kaiser, jeden z twórców modelu transformera, który pracuje teraz w Open AI, twierdzi, że na najbliższe dwa-trzy lata krzywa rozwoju dla modeli rozumujących uczonych tymi technikami jest bardzo stroma*.

W dodatku pojawiają się setki pomysłów na kolejne modyfikacje architektury, które mogłyby poprawić technologię.

Mówienie w tym momencie o jakimkolwiek zastoju albo nadchodzącym kryzysie to fantazja osób, które z jakiegoś powodu nie lubią AI.

**Wiele autorytetów to przeraża. Podpisują apele o to, by wstrzymać prace nad AI.**

– To na szczęście nie mój problem, bo nie jestem autorytetem. I nie należy mnie pytać o zdanie w tych sprawach, bo jestem nieodpowiedzialny. W dodatku to, co się wydarzy, bardziej mnie ciekawi, niż niepokoi.

**Myślisz, że tu, w Polsce, możemy się jeszcze do tego wyścigu włączyć, czy tez finansowo i kadrowo jesteśmy na straconej pozycji?**

– Musimy się włączyć. Ja bardzo popieram działania Piotra Sankowskiego, który robi wszystko, żeby za tym gardłować. Polska ma wielką szansę, żeby w to wejść. Remek Kinas i ekipa tworząca model Bielik – dzięki ich wysiłkom budujemy realną wiedzę na temat najnowszych technik rozwijania modeli rozumujących. I to jest nieocenione.

*Na razie zasiewamy kapitał intelektualny. I jeśli ktoś mówi, że nie mamy szans w starciu z gigantami, to warto przypomnieć, że pięć lat temu nikt nie słyszał o OpenAI, a teraz jest on najpoważniejszym konkurentem Google'a*.

30 lat temu nikt nie słyszał również o Google'u. 90 proc. największych firm technologicznych świata to organizacje, które niedawno jeszcze nie istniały.

**Wierzysz, że jedna z kolejnych powstanie w polskim garażu?**

– U nas jest za zimny klimat na garaże.

**Czy nie boisz się wpływu AI? Nie chodzi mi o katastroficzny scenariusz, że maszyna kiedyś przejmie władzę nad światem, tylko o to, że rozleniwi nam mózgi. Że mając pod ręką takiego genialnego asystenta, przestaniemy myśleć. A nieużywany organ zanika. Możliwe, że ludzkość po prostu zgłupieje?**

– Może tak być, oczywiście. Ale przykład szachistów temu przeczy.

Stworzyliśmy silniki szachowe, które demolują arcymistrzów. Mistrz świata może grać ze Stockfishem z przewagą wieży, a i tak przegra. A jednak to nikogo nie zniechęca ani nie rozleniwia. Przeciwnie, szachy nigdy nie były tak popularne jak teraz. I najlepsi szachiści nigdy nie grali tak dobrze, jak grają dziś.

Nawet więc jeśli matematyka stanie się kolejnymi szachami, a profesorowie matematyki będą pytać programy AI, jak rozwiązywać nurtujące ich zagadnienia, to nie musi przecież prowadzić do zastoju. Może wręcz przeciwnie – będzie z tego intelektualna eksplozja.

Ale przewidywania w takich sprawach prawie zawsze okazują się błędne, więc proponuję mnie nie słuchać.

*Redagował Łukasz Grzymisławski*

**Andrzej Dragan (ur. w 1978 r.) jest profesorem Uniwersytetu Warszawskiego i profesorem wizytującym w Narodowym Uniwersytecie Singapuru. Zajmuje się szukaniem związków między teorią względności i nowym opisem świata, jaki przyniosła fizyka kwantowa. Jest również fotografem, a także popularyzatorem nauki, autorem książki „Kwantechizm, czyli klatka na ludzi", w której opisuje Wszechświat i życie z punktu widzenia fizyki kwantowej i teorii Einsteina.*