Sztuczna inteligencja właśnie zdała testy na teorię umysłu. Naukowcy nie dowierzają

Gazeta Wyborcza, 21.05.2024

https://wyborcza.pl/7,75400,30988780,sztuczna-inteligencja-czesciowo-bije-ludzi-w-rozumieniu-jak.html#s=S.embed_article-K.C-B.1-L.1.zw

Naukowcy poddali testom dwie rodziny modeli językowych - GPT i LLaMA2. Wzdragają się jednak przed przyznaniem, że wykryli u nich tzw. teorię umysłu.

Pod hasłem "teoria umysłu" uczeni opisują zdolność do wczucia się w skórę innego. Zrozumienie, że ktoś może inaczej postrzegać świat niż ja. A nawet przewidzenie, co inny myśli i czuje. Na poziomie fundamentalnym u zwierząt niezbędna jest do tego empatia.

A więc uczucia. A jeśli one, to także świadomość, która według najprostszej definicji jest niczym więcej niż zdolnością do przeżywania subiektywnych doświadczeń.

Do niedawna zakładano więc, że teoria umysłu dotyczy tylko ludzi.

W ostatnich latach coraz więcej biologów podkreśla jednak, że świadomością dysponują także inne zwierzęta. A teorię umysłu zaobserwowano m.in. u kruków, szympansów, bonobo i orangutanów (więcej o empatycznych małpach pod linkiem powyżej).

Małpy człekokształtne zdały choćby tzw. test fałszywego przekonania, z którym ludzkie dzieci radzą sobie dopiero po czwartym roku życia. Podczas takiego sprawdzianu maluchy oglądają film, na którym jedna lalka chowa piłeczkę. Kiedy wychodzi z pokoju, pojawia się w nim druga lalka, która przekłada piłkę w inne miejsce.

Dzieci są wtedy pytane o to, w które miejsce zajrzy pierwsza postać, kiedy wróci do pokoju po swoją własność.

Te młodsze wskazują na miejsce, w którym rzeczywiście znajduje się piłka przełożona przez drugą lalkę. Małe dzieci nie potrafią sobie wyobrazić, że ktoś może doświadczać czegoś innego niż one.

Te starsze rozumieją już, że pierwsza lalka nie musi podzielać ich doświadczeń i sądzi, iż jej zabawka znajduje się w pierwotnym miejscu. Rozumieją jej błędne przekonanie.

Sztuczna inteligencja i teoria umysłu

W najnowszym wydaniu pisma „Nature Human Behaviour" badacze z Niemiec, USA, Wielkiej Brytanii i Włoch udowadniają, że test fałszywego przekonania zdają także modele językowe z rodziny GPT, "wyhodowane" przez naukowców z firmy OpenAI.

ChatGPT-4 nawet bije w tym ludzi!

W różnych testach porównawczych sprawdzających składniki teorii umysłu wzięło udział ponad 1,9 tys. osób. A spośród mówiących algorytmów sztucznej inteligencji (AI) wzięły w nich udział właśnie ChatGPT oraz model językowy LLaMA2 opracowany przez firmę Meta (tę od Facebooka).

Jak się okazało, choć LLaMA2 radził sobie gorzej od ludzi w testach fałszywego przekonania, to bił nas i ChataGPT w rozpoznawaniu faux pas (wszyscy znamy osoby, które bez przerwy je popełniają, kompletnie nie pojmując, dlaczego zwraca im się uwagę).

Z kolei ChatGPT pokonywał i ludzi, i LLaMA2 nie tylko w testach fałszywego przekonania, ale także w rozumieniu owijania w bawełnę – kiedy ktoś nie mówi wprost, czego chce, ale z grzeczności wyraża się pośrednio. Np. zamiast niezbyt grzecznego „podaj sól", zapyta: „Przepraszam bardzo, czy mógłbyś mi podać sól?". Wiadomo, że pytany może to zrobić, ale z grzeczności wypada właśnie tak prośbę o sól sformułować (Polacy są w tym kiepscy, jak wiemy).

ChatGPT świetnie (czasem lepiej niż ludzie, a zawsze lepiej od LLaMA2) orientował się też, kiedy próbowano wprowadzić go w błąd.

Co ciekawe, inni naukowcy niedawno zaraportowali, że sztuczna inteligencja sama potrafi oszukiwać i manipulować, a mistrzem oszustwa okazał się inny algorytm AI stworzony przez firmę Meta (więcej o tym badaniu pod poniższym linkiem).

Naukowcy z USA i Australii ostrzegają, że sztuczna inteligencja już nauczyła się oszukiwania i manipulowania ludźmi.

Naukowcy nie dowierzają temu, co mierzą

Autorzy publikacji w „Nature Human Behaviour" zastrzegają, że to, iż potężne modele językowe radzą sobie – czasem świetnie – w zadaniach sprawdzających teorię umysłu, nie oznacza, że ją posiadają albo że są bardziej świadome cudzych stanów umysłowych niż ludzie. Nie oznacza też, że modele te mają świadomość.

Innymi słowy: według naukowców, choć badane modele zdały stosowne testy, to najwyraźniej bez zrozumienia tego, z czym mają do czynienia!

Jak, nomen omen, rozumieć to zaskakujące, paradoksalne zastrzeżenie, które przeczy właśnie opublikowanym dowodom?

Być może tak, że część osób rozmawiających z modelami językowymi czuje, jakby rozmawiała z ludźmi. A ponieważ wie, że rozmawia z programami komputerowymi (opartymi na sieciach neuronowych), przeżywa dysonans poznawczy, przez który odmawia sztucznej inteligencji cech uznawanych za wyłącznie ludzkie lub zwierzęce.

Ale w końcu jaki mamy dowód na to, że w innym człowieku tlą się świadomość, empatia i wrażliwość na innego? U nas te cechy są generowane przez sygnały elektryczno-chemiczne wymieniane przez neurony w mózgu. U AI ta wymiana jest wyłącznie elektryczna.

Tak naprawdę jednak nie wiemy, co się dzieje pomiędzy miliardami sztucznych neuronów i jaki jest efekt ich samoorganizacji. Wiemy tylko tyle, ile programy sztucznej inteligencji nam mówią. I orientujemy się w ich umiejętnościach, patrząc na to, jak sobie radzą z różnymi zadaniami.

Podobnie wiele wiemy o zdolnościach poznawczych innych ludzi.