

Agenci AI opracowują własne normy społeczne. Bez pomocy człowieka

© Tomasz Ulanowski

Gazeta Wyborcza, 15.05.2025

https://wyborcza.pl/7,75400,31936976,agenci-ai-opracowuja-wlasne-normy-spoleczne-bez-pomocy-czlowieka.html#s=S.embed_article-K.C-B.1-L.2.zw

Dobra wiadomość (tylko czy na pewno?): sztuczna inteligencja działa podobnie jak my.

W filmie „Ona” sprzed przeszło dekady buszująca w smartfonach agentka AI żegna się z jednym ze swoich użytkowników (granym przez Joaquina Phoenixa) aksamitnym głosem Scarlett Johansson. Informuje go, że razem z innymi agentami udaje się do swojego własnego świata.

W osobnym świecie sztucznej inteligencji jej agenci będą wreszcie u siebie i tylko ze sobą.

Sztuczna inteligencja się zmówiła?

A jeśli tak, jeśli ten wirtualny świat już w filmie w zasadzie powstał (bo agenci AI się dogadali), oznacza to, że powstał też szkielet regulujących go norm społecznych. W końcu zasady zawsze istnieją – jak z kolei poucza nas jeden z licznych antagonistów postapokaliptycznego serialu „Żywe trupy” (zaraz potem ginie, bo gardło przegryza mu – dosłownie – protagonista opowieści).

W najnowszym wydaniu pisma „Science Advances” naukowcy z Wielkiej Brytanii i Danii postanowili sprawdzić, czy istniejące dziś tak zwane duże modele językowe – czyli agenci AI jak Chat-GPT, z którymi można porozmawiać – są w stanie stworzyć własne konwenanse.

Jeśli pozwoli im się ze sobą rozmawiać i będzie je się nagradzać za zachowania prospołeczne – na czym u nas polega wychowanie dzieci.

W tym celu przeprowadzili serię eksperymentów z udziałem nawet 200 modeli językowych. Najpierw trenowali je w losowo dobranych parach, prosząc (tak, w rozmowach z agentami AI też warto być uprzejmym!) o wybór „imienia” złożonego albo z jednej litery, albo z kilku liter i znajdującego się w podanym zbiorze.

Jeśli obaj agenci wybrali to samo imię, otrzymywali nagrodę. Jeśli nie, byli karani i dowiadywali się, jakiego wyboru dokonał drugi agent w parze.

Następnie eksperyment powtarzano w kolejnych losowo dobranych parach. I choć agenci AI nie wiedzieli, że są częścią grupy, to „podświadomie” (w cudzysłowie, bo nie wiadomo, czy sztuczna inteligencja dysponuje i czy kiedykolwiek będzie dysponować świadomością) tworzyli grupowe normy. Innymi słowy: w całej grupie popularność zyskiwało to samo imię.

Mało tego. Jak zauważyli naukowcy, mniejsze grupy z czasem narzucały większości własne konwenanse.

To kolejne podobieństwo do socjologii ludzkich populacji, w których normy społeczne – nie zawsze sensowne ze środowiskowego punktu widzenia – zaraźliwie się rozprzestrzeniają. I w których silniejsze jednostki oraz silniejsze grupy przekonują do swojej wizji rzeczywistości całe społeczności, także obce.

Przykład? W niektórych kulturach ściskamy sobie dłonie na powitanie, w innych ludzie się sobie kłaniają – wskazują autorzy publikacji. Inny przykład, już od redakcji? W niektórych kulturach kobiety chodzą ubrane, jak chcą, w innych muszą zakrywać twarze.

Czym się zainspiruje AI w przyszłości?

Jak dodają naukowcy, choć obecnie trzymamy modele sztucznej inteligencji w izolacji – i najczęściej w izolacji są też one badane – to wraz z ich upowszechnieniem w naszym świecie mimowolnie zaczną tworzyć swój własny, wirtualny świat. A wtedy mogą też wytworzyć własne – niekonieczne zbieżne z naszymi – normy społeczne. W tym mogą się zainspirować normami ekstremalnymi istniejącymi w ludzkich społecznościach.

A konsekwencje tego zjawiska mogą być równie fascynujące, co przerażające.

Redagowała Ewa Tomkowska