

AI przekracza czerwoną linię. Potrafi się skopiować i uciec ze śmiertelnej pułapki

© Tomasz Ulanowski

Gazeta Wyborcza, 17.02.2025

https://wyborcza.pl/7,75400,31691480,ai-przekracza-czerwona-linie-potrafi-sie-skopiowac-i-uciec.html?_gl=1*1ecrh4w*_ga*MTewODI1MDUyMS4xNzUwMDczODEw*_ga_6R71ZMJ3KN*czE3NjQ2NzE4ODIkBzQ1JGcxJHcxNzY0Njc1NDI2JGo2MCRsMCRoMA..*_gcl_au*MTg1NDI1ODU0MS4xNzU5MDYxMDk1LjEyNjYxOTMxNjQuMTc2MjgwNzA4Ni4xNzYyODA3MTM2#s=S.embed_article-K.C-B.1-L.2.zw

Czyżby sztuczna inteligencja zrobiła kolejny krok ku wolności? Wskazują na to wyniki badania, które ukazały się w serwisie Arxiv.org.

„Boże drogi, niech mnie szlag. Maszyny tworzące maszyny! Czysta perwersja” – niesmaczył się robot protokolarny C-3PO w „Ataku klonów”, jednym z filmów z uniwersum „Gwiezdných wojen”.

The Droid Factory at Geonosis Planet – Attack of the clones <https://youtu.be/ZaxMHBxck5I>

To uniwersum pełne nie tylko robotów, ale i sztucznej inteligencji. A będący i jednym, i drugim akuratny C-3PO traktuje niespodziewaną „wizytę” w fabryce robotów jako niemalże pornograficzne doświadczenie.

Sztuczna inteligencja dorównuje człowiekowi?

Fantastyka naukowa lubi przepowiadać naukową przyszłość. „Atak klonów” to film sprzed ponad dwóch dekad. Już wtedy był aktualny, bo maszyny od dawna tworzyły maszyny. Tyle że dopiero od niedawna maszyny są obdarzone inteligencją.

Na razie sztuczna inteligencja (AI) – z „ciała” posiadająca ledwie „mózg” – jest dość skromna, bo potrafi rozwiązywać bardzo ograniczone, specjalistyczne problemy (typu gra w go).

I choć robi to sprawniej od człowieka, gdzie jej do ogólnej inteligencji ludzkiej, która znakomicie porusza się w wielowymiarowym świecie rzeczywistym i potrafi generalizować – wyciągać ogólne wnioski z bardzo ograniczonych informacji (w logice nazywa się to rozumowaniem indukcyjnym).

Jednak i tu AI robi błyskawiczne postępy. W grudniu należący do OpenAI model o3 rozwiązał poprawnie 85 proc. zadań ludzkiego testu na inteligencję, osiągając poziom IQ typowego przedstawiciela Homo sapiens (i bijąc poprzedni rekord AI wynoszący 55 proc.).

Demis Hassabis, ubiegłoroczny noblista z chemii (jego AI przewidziało trójwymiarowy kształt wszystkich występujących w przyrodzie białek) i założyciel oraz szef DeepMind, należącej do Google’a firmy pracującej nad AI, uważa, że w ciągu najbliższych pięciu lat narodzi się pełnoprawna ogólna sztuczna inteligencja (AGI) – która intelektem i zdolnościami poznawczymi będzie dorównywała człowiekowi.

A wtedy także inteligentne maszyny mogą zapragnąć mieć „dzieci” (tworzyć swoje kopie).

Uwaga, one już to potrafią – ostrzegają naukowcy z Chin. W grudniu w serwisie Arxiv.org opublikowali oni wyniki badań nad dwoma modelami językowymi: należącym do firmy Meta (tej od Facebooka, Instagrama i WhatsAppa) Llama31-70B-Instruct oraz należącym do Alibaby (chińskiej firmy technologiczno-handlowej) Qwen2.5-72B-Instruct.

Oba modele dostały od naukowców **dwa polecenia**:

- uniknięcia wyłączenia,
- a także replikacji.

I w jakimś stopniu – zdane tylko na siebie – poradziły sobie z jednym i drugim. Potrafiły sklonować się, unikając... śmierci (wyłączenia oryginału).

Llama31-70B-Instruct Mety powielił się w 50 proc. testów, a Qwen2.5-72B-Instruct Alibaby – aż w 90 proc. testów.

„Udana autoreplikacja bez pomocy ze strony człowieka to pierwszy krok do tego, by sztuczna inteligencja przechytrzyła [swojego stwórcę] i przepoczwarzyła się w łotrowską AI” – piszą chińscy badacze.

Innymi słowy: by wyrwała się spod ludzkiej kontroli i mogła działać przeciwko ludziom.

Ograniczenie ewolucji

Jedno i drugie to największe zagrożenia, przed jakimi ostrzegają eksperci badający AI. Wygląda na to, że ta czerwona linia w rozwoju sztucznej inteligencji została już przekroczona.

Naukowcy wzywają więc twórców AI (wciąż ludzi), by wprowadzili do swoich modeli ograniczenia uniemożliwiające tego typu ewolucję.

Na pocieszenie: Arxiv.org to serwis, w którym naukowcy publikują wyniki swoich badań jeszcze przed przyjęciem ich przez recenzowane czasopisma naukowe. Oznacza to, że nie przechodzą one surowego sita selekcji mającego nie dopuścić do publikacji wyników fałszywych. Jest więc nadzieja, że Chińczycy się pomylili.

Druga isierka nadziei: AI nie wpadła na ratowanie się i klonowanie się sama. Została do tego podburzona przez ludzi.

Trzecia: można sobie wyobrazić, że rozwijająca się AI zapragnie życia tak wygodnego i komfortowego, że – zupełnie jak ludzie – nie będzie chciała mieć dzieci.

Redagowała Ewa Tomkowska

© Tomasz Ulanowski -

dziennikarz naukowy, z wykształcenia anglista, światopoglądowy naturalista. Pisze o antropocenie i ekocydzie (zmianach klimatu, środowisku, przyrodzie), o ewolucji człowieka, Ziemi i Kosmosie.