

Nie da się kontrolować sztucznej inteligencji

16 października 2022

Pomysł zniszczenia ludzkości za pomocą sztucznej inteligencji był dyskutowany od kilkudziesięciu lat, a naukowcy niedawno wydali swój werdykt, czy można kontrolować superinteligencję komputerową wysokiego poziomu. Naukowcy twierdzą, że haczyk polega na tym, że aby kontrolować superinteligencję, która wykracza daleko poza ludzkie zrozumienie, konieczne jest modelowanie tej superinteligencji, którą można analizować i kontrolować. Ale jeśli ludzie nie są w stanie tego zrozumieć, to nie da się stworzyć takiej symulacji.

Naukowcy twierdzą, że zasady takie jak „nie krzywdź ludzi” nie mogą zostać ustalone, jeśli ludzie nie rozumieją, jakie scenariusze ma do zaoferowania sztuczna inteligencja. Gdy system komputerowy działa na poziomie przekraczającym możliwości naszych programistów, nie można nałożyć żadnych ograniczeń.

Superinteligencja jest zasadniczo innym problemem niż te, które zwykle bada się w ramach „etyki robotów”. Dzieje się tak, ponieważ superinteligencja jest wieloaspektowa i dlatego potencjalnie jest w stanie zmobilizować różnorodne zasoby, aby osiągnąć cele, które są potencjalnie niezrozumiałe dla ludzi, nie wspominając już o tym, że można ją kontrolować.

Część rozumowania zespołu opierała się na problemie zatrzymania postawionym przez Alana Turinga w 1936 roku. Problem z zatrzymaniem – biorąc pod uwagę, że program/algorytm kiedykolwiek się zatrzyma, czy nie? Zatrzymanie oznacza, że $\square$ program na określonym wejściu zaakceptuje je i zatrzyma się lub odrzuci i zatrzyma się, i nigdy nie wejdzie w nieskończoną pętlę. W istocie zatrzymanie oznacza zatrzymanie pracy. Jak udowodnił Turing za pomocą skomplikowanych obliczeń

matematycznych, chociaż możemy to wiedzieć dla niektórych konkretnych programów, logicznie niemożliwe jest znalezienie sposobu, który pozwoli nam to wiedzieć dla każdego potencjalnego programu, który kiedykolwiek mógłby zostać napisany. To prowadzi nas z powrotem do sztucznej inteligencji, która w stanie hiperinteligentnym mogłaby jednocześnie przechowywać w swojej pamięci wszystkie możliwe programy komputerowe.

Każdy program napisany na przykład po to, by powstrzymać sztuczną inteligencję przed krzywdzeniem ludzi i niszczeniem świata, może dojść do określonego rezultatu i zatrzymać się lub nie – i tak matematycznie niemożliwe jest uzyskanie absolutnej pewności, co oznacza, że $\square$ nie da się tego powstrzymać. Według naukowców alternatywą dla nauczania etyki AI i zakazu niszczenia świata – czego żaden algorytm nie może być absolutnie pewien – jest ograniczenie możliwości superinteligencji.

Badanie również odrzuciło ten pomysł, sugerując, że ograniczyłoby to zakres sztucznej inteligencji; argument brzmi: jeśli nie zamierzamy go wykorzystać do rozwiązywania zadań, które wykraczają poza człowieka, to po co w ogóle go tworzyć? Jeśli zamierzamy promować sztuczną inteligencję, możemy nawet nie wiedzieć, kiedy pojawi się superinteligencja poza naszą kontrolą, to jest tak niezrozumiałe. Oznacza to, że musimy zacząć zadawać poważne pytania dotyczące tego, dokąd zmierzamy”.

Źródło: InneMedium.pl