Ludzkość nie da rady kontrolować superinteligentnej AI

20 stycznia 2021

Ludzkość może nie być w stanie kontrolować superinteligentnej sztucznej inteligencji (SI), sądzą autorzy najnowszych teoretycznych badań. Co więcej, możemy nawet nie wiedzieć, że stworzyliśmy tego typu AI (artificial intelligence).

Na naszych oczach dokonuje się szybki postęp algorytmów sztucznej inteligencji. Maszyny wygrywają z ludźmi w go i pokera, pokonują doświadczonych pilotów myśliwców podczas walki powietrznej, uczą się od podstaw metodą prób i błędów, zapowiadają wielką rewolucję w medycynie i naukach biologicznych, stawiają diagnozy równie dobrze co lekarze i potrafią lepiej od ludzi odróżniać indywidualne ptaki. To pokazuje, na jak wielu polach dokonuje się szybki postęp.

Wraz z tym postępem rodzą się obawy o to, czy będziemy w stanie kontrolować sztuczną inteligencję. Obawy, które są wyrażane od co najmniej kilkudziesięciu lat. Od 1942 roku znamy słynne trzy prawa robotów, które pisarz Isaac Asimov

przedstawił w opowiadaniu „Zabawa w berka”: „1. Robot nie może skrzywdzić człowieka, ani przez zaniechanie działania dopuścić, aby człowiek doznał krzywdy, 2. Robot musi być posłuszny rozkazom człowieka, chyba że stoją one w sprzeczności z Pierwszym Prawem, 3. Robot musi chronić samego siebie, o ile tylko nie stoi to w sprzeczności z Pierwszym lub Drugim Prawem. Później Asimov dodał nadrzędne prawo 0: Robot nie może skrzywdzić ludzkości, lub poprzez zaniechanie działania doprowadzić do uszczerbku dla ludzkości”.

W 2014 roku filozof Nick Bostrom, dyrektor Instytutu Przyszłości Ludzkości na Uniwersytecie Oksfordzkim badał, w jaki sposób superinteligentna SI może nas zniszczyć, w jaki sposób możemy ją kontrolować i dlaczego różne metody kontroli mogą nie działać. Bostrom zauważył dwa problemy związane z kontrolą AI. Pierwszy to kontrolowanie tego, co SI może zrobić. Na przykład możemy kontrolować, czy podłączy się do internetu. Drugi to kontrolowanie tego, co będzie chciała zrobić. Aby to kontrolować, będziemy np. musieli nauczyć ją zasad pokojowego współistnienia z ludźmi. Jak stwierdził Bostrom, superinteligentna SI będzie prawdopodobnie w stanie pokonać wszelkie ograniczenia, jakie będziemy chcieli nałożyć na to, co może zrobić. Jeśli zaś chodzi o drugi problem, to Bostrom wątpił, czy będziemy w stanie czegokolwiek nauczyć superinteligentną sztuczną inteligencję.

Teraz z problemem kontrolowania sztucznej inteligencji postanowił zmierzyć się Manuel Alfonseca i jego zespół z Universidad Autonoma de Madrid. Wyniki swojej pracy opisał na łamach „Journal of Artificial Intelligence Research”.

Hiszpanie zauważają, że jakikolwiek algorytm, którego celem będzie zapewnienie, że AI nie zrobi ludziom krzywdy, musi najpierw symulować zachowanie mogące zakończyć się krzywdą człowieka po to, by maszyna potrafiła je rozpoznać i zatrzymać. Jednak zdaniem naukowców, żaden algorytm nie będzie w stanie symulować zachowania sztucznej inteligencji i z całkowitą pewnością stwierdzić, czy konkretne działanie może

zakończyć się krzywdą dla człowieka. „Już wcześniej udowodniono, że pierwsze prawo Asimova jest problemem, którego nie da się wyliczyć. Jego przestrzeganie jest więc nierealne”.

Co więcej, możemy nawet nie wiedzieć, że stworzyliśmy superinteligentną maszynę, stwierdzają badacze. Wynika to z twierdzenia Rice’a, zgodnie z którym nie można odgadnąć, jaki będzie wynik działania programu komputerowego patrząc wyłącznie na jego kod.

Alfonseca i jego koledzy mają jednak też i dobre wiadomości. Otóż nie musimy się już teraz martwić tym, co zrobi superinteligentna SI. Istnieją bowiem trzy poważne ograniczenia dla badań takich, jakie prowadzą Hiszpanie. Po pierwsze, taka niezwykle inteligentna AI powstanie nie wcześniej niż za 200 lat. Po drugie nie wiadomo, czy w ogóle możliwe jest stworzenie takiego rodzaju sztucznej inteligencji, czyli maszyny inteligentnej na tylu polach co ludzie. Po trzecie zaś, o ile możemy nie być w stanie kontrolować superinteligentnej SI, to powinno być możliwe kontrolowanie sztucznej inteligencji, która jest superinteligentna w wąskim zakresie. „Już mamy superinteligencję takiego typu. Na przykład mamy maszyny, które liczą znacznie szybciej niż ludzie. To superinteligencja wąskiego typu, prawda?”, stwierdza Alfonseca.

Autorstwo: Mariusz Błoński
Na podstawie: Spectrum IEEE
Ilustracja: Geralt (CC0)
Źródło: KopalniaWiedzy.pl