Naukowiec ostrzega o zagrożeniu ze strony super-AI w ciągu 5 lat

6 maja 2024

W obliczu szybko rozwijających się technologii AI, naukowcy jak Eliezer Yudkowsky alarmują o potencjalnych zagrożeniach, które mogą wynikać z braku kontroli nad superinteligencjami. Według Yudkowsky'ego, który jest badaczem w Machine Intelligence Research Institute, możemy mieć tylko pięć lat na zapobieżenie scenariuszom, w których niekontrolowana AI prowadzi do katastrofy na skalę globalną.

Zawrotne tempo, z jakim ta nowa technologia postępuje, sprawiło, że niektórzy eksperci w tej dziedzinie zaczynają się niepokoić, a przewidywania o ponurej i dystopijnej przyszłości naszego gatunku spełniają się szybciej, niż się spodziewano. Eliezer Yudkowsky, amerykański badacz sztucznej inteligencji i pisarz specjalizujący się w teorii decyzji i etyce, jest powszechnie znany z popularyzowania idei przyjaznej sztucznej inteligencji (AI), w tym koncepcji, że może nie być „alarmu", który poinformowałby nas, kiedy przestanie być ona przyjazna.

Jego prace na temat możliwości niekontrolowanego wybuchu AI wpłynęły na książkę „Superinteligencja: Ścieżki, niebezpieczeństwa, strategie” filozofa Nicka Bostroma, opublikowaną w 2014 roku. Teraz, w rozmowie z brytyjskim medium „The Guardian”, Yudkowsky przewiduje, że brak kontroli może nastąpić szybciej, niż wielu sądzi.

Trudność polega na tym, że ludzie nie zdają sobie sprawy, że mamy niewielkie szanse na przetrwanie ludzkości. Jego zdaniem jeśli by przypisać prawdopodobieństwa, to nasz obecny pozostały czas to raczej pięć lat, a nie pięćdziesiąt. Może to być dwa lata, może być dziesięć. A przez „pozostały czas” Yudkowsky ma na myśli czas, jaki mamy do końca, zanim stanąrzymy w obliczu końca wszystkiego spowodowanego przez maszyny, w scenariuszu, który mógłby się wahać od stylu „Terminatora” (wojna spowodowana lub napędzana przez AI) do „Matrixa” (cywilizacja zależna od AI, która zachowuje się jak Bóg).

Próbując wyrwać ludzkość z jej obojętności w tej sprawie, Yudkowsky opublikował artykuł w „Time” wiosną zeszłego roku, w którym posunął się tak daleko, że zasugerował, jako ostatnią alternatywę przetrwania ludzkości, bombardowanie farm komputerowych, w których rozwijane i szkolone są AI. Ostrzega, że jeśli nie podejmiemy natychmiastowych działań, by zapanować nad rozwojem tej technologii, to w ciągu najbliższych pięciu lat możemy stanąć w obliczu całkowitej zagłady naszego gatunku.

Yudkowsky podkreśla, że rozwijające się AI mogą zacząć działać w sposób, którego nie jesteśmy w stanie przewidzieć ani kontrolować. Porównuje to do nieprzewidzianych exploitów w architekturze komputerowej, jak np. row-hammer, które były nieoczekiwane przez projektantów systemów komputerowych. Podobne niespodziewane zachowania mogą pochodzić od AI, które mogłyby manipulować lub działać w sposób szkodliwy dla ludzkości.

Debata na temat ryzyk związanych z AI nie jest jednomyślna. Niektórzy eksperci, jak rozmówcy w podcaście Econlib, argumentują, że obawy dotyczące autonomicznych działań AI mogą być przesadzone. Wskazują oni, że tak jak narzędzia można używać zarówno na dobre, jak i na złe cele, podobnie jest z AI. Krytycy jak Mort Dubois zauważają, że wiele aspektów fizycznego świata może stanowić barierę dla nadmiernych możliwości AI, które mogą nie być w stanie samodzielnie rozwiązać wszystkich problemów.

Niezależnie od różnic w opiniach, jasne jest, że rosnąca moc AI wymaga odpowiedzialnego podejścia zarówno w kwestii rozwoju technologii, jak i implementacji odpowiednich środków bezpieczeństwa. Ostatecznie, jak zauważa Yudkowsky i inni, nieprzewidywalność AI, szczególnie tych o wyższych zdolnościach, stanowi wyzwanie, które nie może być ignorowane.

Ilustracja: KELLEPICS (CC0)
Źródło: ZmianyNaZiemi.pl