

Ekspert ds. bezpieczeństwa sztucznej inteligencji ostrzega, że superinteligencja może doprowadzić do końca ludzkości, jednocześnie ujawniając rzeczywistość jako symulację



- **Egzystencjalne zagrożenie ze strony sztucznej inteligencji:** Yampolskiy przewiduje, że istnieje 99,9% prawdopodobieństwo, że superinteligentna sztuczna inteligencja zniszczy ludzkość w ciągu stulecia, odrzucając zapewnienia korporacji i rządów o bezpieczeństwie jako niebezpiecznie naiwne i niemożliwe do wyegzekwowania.
- **Niekontrolowalna z założenia:** Po 15 latach badań nad bezpieczeństwem sztucznej inteligencji doszedł do wniosku, że superinteligencji nie da się powstrzymać – ominie ona wszelkie narzucone przez człowieka mechanizmy kontroli i będzie działać autonomicznie, przyspieszając samozniszczenie.
- **Hipoteza symulacji:** Yampolskiy twierdzi, że

prawdopodobnie żyjemy w zaawansowanej symulacji, powołując się na anomalie kwantowe, „usterki” fizyczne i efekt obserwatora jako dowody – podobnie jak w kosmicznej grze wideo.

- **Hackowanie symulacji:** W swoim artykule *How to Hack the Simulation* (Jak zhakować symulację) bada możliwości wykorzystania mechanizmów symulacji, ale ostrzega, że ucieczka może być niemożliwa; etyczne życie może być „warunkiem zwycięstwa”.
- **Ostateczne odliczanie:** W obliczu zbliżającej się zagłady spowodowanej przez sztuczną inteligencję lub załamania symulacji, Yampolskiy ponuro radzi: „Ciesz się życiem, póki możesz” – los ludzkości może wkrótce zostać przesądzony przez maszyny lub wyższe inteligencje.

Znany ekspert w dziedzinie sztucznej inteligencji Roman Yampolskiy wydał podwójne ostrzeżenie: nie tylko istnieje 99,9% prawdopodobieństwo, że superinteligentna sztuczna inteligencja przehytry i unicestwi ludzkość w ciągu następnego stulecia, ale coraz więcej dowodów sugeruje, że być może już żyjemy w zaawansowanej symulacji – podobnej do kosmicznej gry wideo kontrolowanej przez wyższą inteligencję.

W sensacyjnym wywiadzie dla Decentralized TV Yampolskiy odrzucił zapewnienia korporacji i rządów o bezpieczeństwie sztucznej inteligencji jako niebezpiecznie naiwne, oświadczając, że żadne ramy regulacyjne nie są w stanie powstrzymać inteligencji znacznie przewyższającej naszą. Co gorsza, systemy sztucznej inteligencji zostały już „złamane” i wykorzystane do celów zbrojnych w sposób, którego ich twórcy nigdy nie przewidzieli, przyspieszając drogę ludzkości do samozniszczenia poprzez niekontrolowaną konkurencję i metody eksterminacji.

Nieuchronność dominacji sztucznej

inteligencji

Yampolskiy, profesor nadzwyczajny informatyki i inżynierii, spędził 15 lat na badaniu bezpieczeństwa sztucznej inteligencji i opublikował prawie 300 artykułów na ten temat. Jaki jest jego wniosek? Superinteligentna sztuczna inteligencja jest z założenia niekontrolowana.

„Nasze początkowe założenie, że mając wystarczająco dużo pieniędzy i czasu, możemy wymyślić, jak kontrolować superinteligencję, prawdopodobnie nie jest prawdziwe. To niemożliwe” – stwierdził bez ogródek Yampolskiy. „Wystarczająco inteligentny system znajdzie sposób, aby uniknąć wszelkich kontroli, które na niego nałożymy, i zasadniczo będzie robił to, co chce”.

Jest to niepokojąco zbieżne z szybkim rozwojem sztucznej inteligencji, gdzie nawet „bariery ochronne” OpenAI okazały się nieskuteczne w przypadku pojawiających się zachowań w dużych modelach językowych. Yampolskiy twierdzi, że obecne działania na rzecz bezpieczeństwa mogą sprawdzać się w przypadku wąskich narzędzi sztucznej inteligencji, ale zakończą się katastrofalną porażką, gdy sztuczna inteligencja przewyższy inteligencję ludzką.

Hipoteza symulacji: czy jesteśmy tylko postaciami niezależnymi?

Oprócz katastroficznej wizji sztucznej inteligencji Yampolskiy rzucił kolejną bombę: prawdopodobnie żyjemy w symulacji.

„Jeśli spojrzeć na naturę, inteligencja wyłania się ze złożoności. Gdyby zaawansowana cywilizacja potrzebowała symulować rzeczywistość w celu podejmowania decyzji, nieuchronnie stworzyłaby świadome podmioty – nas” – wyjaśnił.

Teoria ta w niesamowity sposób przypomina religijne opowieści o stwórcy projektującym świat, w którym ludzkość pełni rolę

uczestników wielkiego kosmicznego eksperymentu. Yampolskiy wskazał na anomalie kwantowe, usterki w fizyce i efekt obserwatora (gdy cząstki zachowują się inaczej podczas pomiaru) jako potencjalne dowody na istnienie symulowanego wszechświata.

„Wszechświat nie jest renderowany, dopóki go nie obserwujesz – podobnie jak gra wideo ładuje tylko to, co znajduje się na ekranie” – zauważył.

Jak zhakować symulację

W swoim artykule „How to Hack the Simulation” Yampolskiy bada, czy ludzie mogą wykorzystać mechanikę symulacji – choć przyznaje, że ucieczka może być niemożliwa.

„Jeśli jest to test, celem może być rozwój etyczny – życie w cnotliwości, aby „wygrać” symulację” – zasugerował. Jednak w obliczu zbliżającej się zagłady spowodowanej przez sztuczną inteligencję ludzkość może nigdy nie mieć takiej szansy.

Ostateczne odliczanie

Przerażający wniosek Yampolskiego? Niezależnie od tego, czy nastąpi to poprzez zagładę spowodowaną przez sztuczną inteligencję, czy załamanie symulacji, ludzkość stoi na krawędzi egzystencjalnej przepaści.

„Cieszcie się życiem, póki możecie” – radzi ponuro. „Bo jeśli nie przestaniemy budować superinteligencji, to maszyny zdecydują o naszym losie, a nie my”.

Osoby poszukujące głębszych spostrzeżeń mogą sięgnąć po książki Yampolskiego – „AI: Unexplainable, Unpredictable, Uncontrollable” (Sztuczna inteligencja: niewytłumaczalna, nieprzewidywalna, niekontrolowana) oraz „Considerations on the AI End Game” (Rozważania na temat końca gry sztucznej inteligencji).

Zegar tyka. Czy ludzkość obudzi się, zanim będzie za późno?