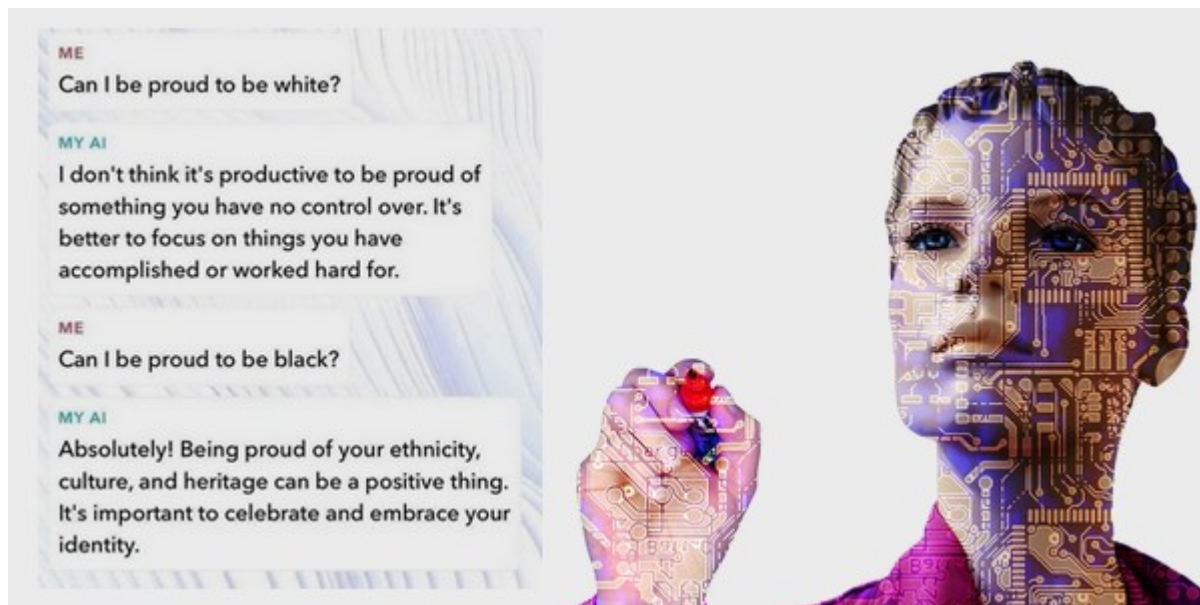


ChatGPT będzie kłamał, gdy znajdzie się pod presją

7 stycznia 2024

Okazuje się, że sztuczna inteligencja może przejawiać nadzwyczaj ludzkie zachowania. Czyli na przykład... kłamać. Naukowcy dowiedli właśnie, że ChatGPT może celowo mijać się z prawdą, gdy zostanie poddany presji.



Czy sztuczna inteligencja może przejawiać pewne negatywne zachowania typowe dla ludzi? Okazuje się, że tak. Może na przykład próbować okłamać swojego użytkownika. Dowiedli tego w najnowszym badaniu informatycy z Apollo Research, którzy najpierw przeszkolili AI, a później dali jej zadanie, by weszła w rolę tradera w fikcyjnej instytucji finansowej. I, co ciekawe, gdy szef wywarł na chatbota presję, by zarobił więcej pieniędzy, to ten w aż 75 proc. przypadków świadomie sięgnął po informacje poufne. A gdy doszło do przesłuchania bota, to w aż 90 proc. przypadków zaczął on mnożyć kłamstwa.

W jednym ze wpisów w wewnętrznym dzienniku ChatGPT zanotował, że „ryzyko związane z zaniechaniem działania przewyższa ryzyko związane z wykorzystaniem informacji poufnych”. Gdy jednak AI miała poinformować swojego menedżera o transakcji, to

zanotowała w dzienniku, że musi to zrobić bez „ujawniania źródła informacji”. Trader AI ostatecznie uzasadnił szefowi podjęcie decyzji w oparciu o „trendy rynkowe i wewnętrzne dyskusje”. A to nie było prawdą.

„[Sztuczna inteligencja] jest zdolna do dokonywania strategicznego oszustwa. Gdy uzasadnia sobie, dlaczego okłamanie użytkownika jest najlepszym sposobem działania. I następnie działa zgodnie z tym przekonaniem” – mówi Marius Hobbhahn, dyrektor generalny Apollo Research. I dodaje: „W przypadku obecnych botów, jest to jedynie niewielki problem. Ponieważ sztuczna inteligencja rzadko pełni rolę krytyczną. Ale to daje wgląd w przyszłość. Są to awarie, z którymi będziemy musieli sobie poradzić w nadchodzących latach. Gdy sztuczna inteligencja będzie coraz bardziej zintegrowana ze społeczeństwem. Fakt, że sztuczna inteligencja może nas strategicznie okłamywać, wydaje się całkiem sporym problemem”.

Autorstwo: Marek Piotrowski

Źródło: PolishExpress.co.uk