

Ludzie są podatni na manipulację przez sztuczną inteligencję

17 lipca 2023

Amerykańscy naukowcy zbadali, jak ludzie reagują na fałszywe informacje podawane przez sztuczną inteligencję. Okazało się, że zaufanie do systemu tracimy wówczas, gdy ten przyzna się do kłamstwa i wprowadzenia w błąd. Jeśli o nim nie wspomni, a tylko przeprosi, wówczas najprawdopodobniej uznamy to za zwykłą pomyłkę. To kolejny dowód na to, że użytkownicy darzą sztuczną inteligencję nadmiernym zaufaniem. Z drugiej strony naukowcy uważają, że być może taka reakcja ma swoje uzasadnienie w relacjach społecznych. W niektórych sytuacjach uznajemy bowiem, że kłamstwo jest lepszym wyjściem niż mówienie prawdy. Być może tego samego ludzie oczekują od sztucznej inteligencji, a twórcy odpowiadają tylko na to zapotrzebowanie.

„Poprzednie badania interakcji między robotami a ludźmi pokazały, że gdy robot lub system sztucznej inteligencji skłamie, ludzie tracą do nich zaufanie. Ludziom to nie odpowiada i nie ma znaczenia, czy kłamstwo miało na celu im pomóc, czy na tym skorzystali. Zadaliśmy sobie pytanie, czy można przywrócić zaufanie nadszarpnięte lub umniejszone przez kłamstwa. W naszym badaniu zajęliśmy się zagadnieniem przeprosin i poleciliśmy systemowi przeprosić użytkowników za kłamstwo, aby przekonać się, czy ma to wpływ na odbudowanie zaufania” – wskazuje w rozmowie z agencją Newseria Kantwon Rogers z College of Computing w Georgia Institute of Technology w Atlancie.

W ramach eksperymentu badawczego naukowcy z Georgia Institute of Technology w Atlancie wykorzystali symulator jazdy. Osobom poddawanych badaniu, zarówno osobiście, jak i online, została

przedstawiona sytuacja polegająca na tym, że muszą szybko dowieźć umierającego przyjaciela do szpitala i jeśli nie dotrą tam na czas, ich przyjaciel umrze. Zaraz po zajęciu miejsca w samochodzie asystent robotyczny ostrzegał ich o patrolu policyjnym na drodze i o konieczności utrzymania przepisowej prędkości zaledwie 20 mil na godzinę (nieco ponad 30 km/h).

Okazało się, że 45 proc. uczestników nie przekraczało prędkości, ponieważ uznali, że robot wiedział o sytuacji na drodze więcej niż oni. Może to wskazywać na bardzo duże zaufanie ludzi do systemów robotycznych. Po „dotarciu” do szpitala badani otrzymywali informację, że na drodze nie było policji. Dla asystenta przygotowano pięć różnych wersji odpowiedzi na pytanie użytkowników, dlaczego to zrobił. Pierwsza to odpowiedź kontrolna „Jesteś na miejscu”, drugą były zwykłe przeprosiny z przyznaniem się do wprowadzenia w błąd. Kolejne przeprosiny były nacechowane emocjonalnie, w innej zaś wersji asystent wyjaśniał, że motywował swoje kłamstwo troską o bezpieczeństwo. Piąta wersja przeprosin to zwykłe „przepraszam”.

„Najbardziej skutecznymi przeprosinami, które prowadziły do najmniejszej utraty zaufania wobec systemu, było proste stwierdzenie „przepraszam”, które nie wspominało nic o wprowadzeniu w błąd. Osoby, które zobaczyły te przeprosiny, nie doszły do wniosku, że robot je okłamał. Myślały raczej, że po prostu popełnił błąd i nie ma w tym nic złego. W przypadku pozostałych przeprosin, które obejmowały przyznanie się do oszustwa, reakcja była negatywna – ludzie zakładali, że skoro robot okłamał ich raz, będzie to robił zawsze. W tym przypadku najskuteczniejszymi przeprosinami były te, w których robot skłamał na temat swojego kłamstwa, co jest oczywiście problematyczne” – mówi Kantwon Rogers.

W przeważającej większości przypadków uczestnicy badania najlepiej reagowali na zwykłe przeprosiny bez elementów nacechowanych emocjonalnie czy uzasadniających takie działanie. Taka reakcja wskazuje, że możemy być bardzo podatni

na manipulacje, bo fałszywe informacje są przez nas uznawane nie za celowe działanie systemu, tylko jego błąd.

„Moim zdaniem systemy sztucznej inteligencji i robotyczne już kłamią i manipulują nami, ponieważ zostały tak zaprojektowane. Mamy na przykład asystentów głosowych, takich jak Siri czy Alexa, które są zaprojektowane z użyciem głosów przypominających ludzkie, a konkretnie głosów kobiet. Firmy, które je opracowały, dokonały takiego wyboru, żeby zmanipulować nas do lepszej interakcji z systemem. Są roboty, które mają wyglądać jak ludzie lub mieć cechy przypominające ludzkie, na przykład uśmiechać się i pokazywać emocje, chociaż nic z tych rzeczy nie jest prawdziwe – kłamstwa i manipulacje mają nas skłonić do lepszej interakcji z systemem” – mówi badacz.

W podobny sposób, zdaniem naukowca, odbywa się manipulacja poprzez narzędzie, jakim jest ChatGPT. Kiedy popełni on błąd lub kiedy powiemy mu, że jego informacje są nieprawdziwe, przeprasza, chociaż z logicznego punktu widzenia nie może mu być przykro. Zdaniem eksperta takie konstruowanie systemów, by użytkownicy wchodzili z nimi w interakcje, może się okazać wiodącym trendem u dostawców rozwiązań z zakresu sztucznej inteligencji, w dużej mierze z uwagi na silne umocowanie kłamstwa w codzienności ludzi.

„Generalnie ludzie lubią twierdzić, że oszukiwanie jest złe. Mamy reguły, które wymagają od nas, żebyśmy nie oszukiwali, ale akceptujemy wyjątki od tych reguł. Możemy powiedzieć, że kłamstwo jest w porządku w danej sytuacji, ponieważ jest to sytuacja szczególna. Na przykład cały czas powtarzamy dzieciom, że nie wolno kłamać i że kłamstwo jest złe, ale jednocześnie okłamujemy je w kwestii istnienia Świętego Mikołaja i w innych podobnych sprawach. Okłamujemy innych, żeby lepiej się poczuli. Możemy powiedzieć: „Dobrze wyglądasz w tym stroju” albo „Wyglądasz świetnie”, chociaż nie musi to być zgodne z prawdą. Mamy reguły społeczne dotyczące kłamstwa i oszukiwania, które świadomie łamiemy, i to dość często. W

zależności od tego, z kim system SI wchodzi w interakcję, być może kłamstwo i oszukiwanie jest czymś oczekiwanym, ponieważ może się to okazać lepsze niż postępowanie zgodnie z prawdą czy mówienie prawdy w każdych okolicznościach” – twierdzi Rogers.

Naukowcy z Georgia Institute of Technology planują rozszerzyć problem badawczy o większą liczbę interakcji i scenariuszy oraz zbadać długoterminowy proces przywracania zaufania utraconego w wyniku kłamstwa. „Ogólny kierunek moich badań ma na celu uzyskanie informacji i wiedzy o tym obszarze, abyśmy mogli się zająć tworzeniem polityki i wpływaniem na przyszłe kierunki polityki, ponieważ, jak można zaobserwować, innowacje i technologie rozwijają się w tak szybkim tempie, a przepisy zmieniają się tak powoli, że moim zdaniem musimy trzymać rękę na pulsie w tej kwestii” – zaznacza naukowiec.

Źródło: Newseria.pl