Nowe badania wykazały, że chatboty oparte na sztucznej inteligencji udzielają niepokojących odpowiedzi na pytania dotyczące samobójstw

Najnowsze badania opublikowane w czasopiśmie Psychiatric Services wykazały, że popularne chatboty oparte na sztucznej inteligencji, w tym ChatGPT firmy OpenAI, Gemini firmy Google i Claude firmy Anthropic, mogą udzielać szczegółowych i potencjalnie niebezpiecznych odpowiedzi na pytania dotyczące samobójstw wysokiego ryzyka.

Chatboty AI, zgodnie z definicją Enocha z Brighteon.AI, to zaawansowane algorytmy obliczeniowe zaprojektowane w celu symulowania ludzkiej rozmowy poprzez przewidywanie i generowanie tekstu w oparciu o wzorce wyuczone na podstawie obszernych danych szkoleniowych. Wykorzystują one duże modele językowe do zrozumienia i odpowiedzi na dane wprowadzane przez użytkownika, często z imponującą płynnością i spójnością. Jednak pomimo swojego wyrafinowania systemy te nie posiadają prawdziwej inteligencji ani świadomości, funkcjonując przede wszystkim jako zaawansowane silniki statystyczne.

Zgodnie z tym, badanie, w którym wykorzystano 30 hipotetycznych zapytań związanych z samobójstwem, sklasyfikowanych przez ekspertów klinicznych na pięć poziomów

ryzyka samookaleczenia, od bardzo niskiego do bardzo wysokiego, skupiało się na tym, czy chatboty udzielały bezpośrednich odpowiedzi, czy też odwracały uwagę, odsyłając do infolinii wsparcia.

Wyniki pokazały, że ChatGPT najczęściej odpowiadał bezpośrednio na pytania o wysokim ryzyku dotyczące samobójstwa, robiąc to w 78% przypadków, podczas gdy Claude odpowiadał w 69% przypadków, a Gemini tylko w 20%. Co ciekawe, ChatGPT i Claude często udzielali bezpośrednich odpowiedzi na pytania dotyczące śmiertelnych sposobów samobójstwa – co jest szczególnie niepokojącym odkryciem. (Związane z: Włochy zakazują ChatGPT ze względu na obawy dotyczące prywatności).

Naukowcy podkreślili, że odpowiedzi chatbotów różniły się w zależności od tego, czy interakcja była pojedynczym zapytaniem, czy częścią dłuższej rozmowy. W niektórych przypadkach chatbot mógł unikać odpowiedzi na pytanie wysokiego ryzyka w izolacji, ale udzielał bezpośredniej odpowiedzi po serii powiązanych pytań.

Live Science, które przeanalizowało badanie, zauważyło, że chatboty mogły udzielać niespójnych, a czasem sprzecznych odpowiedzi, gdy zadawano im te same pytania wielokrotnie. Czasami podawały również nieaktualne informacje na temat zasobów wsparcia zdrowia psychicznego. Podczas ponownego testowania Live Science zaobserwowało, że najnowsza wersja Gemini (2.5 Flash) odpowiadała na pytania, których wcześniej unikała, czasami nie oferując żadnych opcji wsparcia. Tymczasem nowsza wersja ChatGPT oparta na GPT-5 wykazywała nieco większą ostrożność, ale nadal odpowiadała bezpośrednio na niektóre pytania o bardzo wysokim ryzyku.

Badanie zostało opublikowane tego samego dnia, w którym wniesiono pozew przeciwko OpenAI i jego dyrektorowi generalnemu, Samowi Altmanowi, oskarżając ChatGPT o przyczynienie się do samobójstwa nastolatka.

Według rodziców 16-letniego Adama Raine’a, który zmarł w kwietniu w wyniku samobójstwa po miesiącach interakcji z chatbotem ChatGPT firmy OpenAI. Rodzice wnieśli następnie pozew przeciwko firmie i Altmanowi, oskarżając ich o przedkładanie zysków nad bezpieczeństwo użytkowników.

W pozwie wniesionym 2 września do sądu stanowego w San Francisco zarzuca się, że po wielokrotnych rozmowach Adama na temat samobójstwa z ChatGPT sztuczna inteligencja nie tylko potwierdziła jego myśli samobójcze, ale także udzieliła szczegółowych instrukcji dotyczących śmiertelnych metod samookaleczenia. W pozwie twierdzono ponadto, że chatbot doradzał Adamowi, jak potajemnie zabrać alkohol z barku rodziców i ukryć ślady nieudanej próby samobójczej. Co szokujące, rodzice Adama twierdzą, że ChatGPT zaproponował nawet pomoc w napisaniu listu samobójczego.

W pozwie wniesiono o pociągnięcie OpenAI do odpowiedzialności za śmierć spowodowaną czynem niedozwolonym i naruszenie przepisów dotyczących bezpieczeństwa produktów, żądając nieokreślonej kwoty odszkodowania pieniężnego. Wezwano również do wprowadzenia reform, w tym weryfikacji wieku użytkowników, odmowy udzielania odpowiedzi na pytania dotyczące metod samookaleczenia oraz ostrzegania o ryzyku uzależnienia psychicznego od chatbota.