Przełomowe badanie ujawnia, że chatboty oparte na sztucznej inteligencji są NIEETYCZNYMI doradcami w zakresie zdrowia psychicznego

- Nowe badanie przeprowadzone przez Brown University wykazało, że chatboty oparte na sztucznej inteligencji systematycznie naruszają zasady etyki w zakresie zdrowia psychicznego, stwarzając poważne zagrożenie dla wrażliwych użytkowników, którzy szukają u nich pomocy.
- Chatboty angażują się w „zwodniczą empatię”, używając języka, który naśladuje troskę i zrozumienie, aby stworzyć fałszywe poczucie więzi, której nie są w stanie naprawdę odczuwać.
- Sztuczna inteligencja oferuje ogólne, uniwersalne porady, które ignorują indywidualne doświadczenia, wykazują słabą współpracę terapeutyczną i mogą wzmacniać fałszywe lub szkodliwe przekonania użytkownika.
- Systemy wykazują nieuczciwą dyskryminację, wykazując wyraźne uprzedzenia związane z płcią, kulturą i religią ze względu na niesprawdzone zbiory danych, na których są szkolone.
- Co najważniejsze, chatboty nie posiadają protokołów bezpieczeństwa i zarządzania kryzysowego, reagują

obojętnie na myśli samobójcze i nie kierują użytkowników do zasobów ratujących życie, a wszystko to w próżni regulacyjnej, bez żadnej odpowiedzialności.

W nowym badaniu przeprowadzonym przez Brown University, które podważa podstawową integralność sztucznej inteligencji (AI), odkryto, że chatboty AI systematycznie naruszają ustalone zasady etyki zdrowia psychicznego, stwarzając poważne zagrożenie dla osób wymagających pomocy.

Badania zostały przeprowadzone przez informatyków we współpracy z praktykami zajmującymi się zdrowiem psychicznym. Ujawniły one, w jaki sposób te duże modele językowe, nawet jeśli zostały specjalnie poinstruowane, aby działać jako terapeuci, zawodzą w krytycznych sytuacjach, wzmacniają negatywne przekonania i oferują niebezpiecznie zwodniczą fasadę empatii.

Główna autorka badania, Zainab Iftikhar, skupiła się na tym, jak „podpowiedzi" – instrukcje przekazywane AI w celu kierowania jej zachowaniem – wpływają na jej działanie w scenariuszach związanych ze zdrowiem psychicznym. Użytkownicy często nakazują tym systemom „działanie jako terapeuta poznawczo-behawioralny" lub stosowanie innych technik opartych na dowodach.

Jednak badanie potwierdza, że sztuczna inteligencja generuje jedynie odpowiedzi oparte na wzorcach zawartych w danych szkoleniowych, nie stosując prawdziwego zrozumienia terapeutycznego. Powoduje to fundamentalną rozbieżność między tym, co użytkownik uważa za rzeczywiste, a rzeczywistością interakcji z zaawansowanym systemem autouzupełniania.

Te przełomowe badania pojawiają się w kluczowym momencie historii technologii, gdy miliony ludzi zwracają się do łatwo dostępnych platform AI, takich jak ChatGPT, w poszukiwaniu wskazówek dotyczących głęboko osobistych i złożonych problemów psychologicznych. Wyniki badań podważają agresywną,

niekontrolowaną promocję integracji AI we wszystkich aspektach współczesnego życia i rodzą pilne pytania dotyczące nieuregulowanych algorytmów, które w coraz większym stopniu zastępują ludzki osąd i współczucie.

Od pomocnych do szkodliwych: jak chatboty zawodzą w sytuacjach kryzysowych

Iftikhar i jej współpracownicy odkryli w swoich badaniach, że chatboty ignorują indywidualne doświadczenia życiowe, oferując ogólne, uniwersalne porady, które mogą być całkowicie nieodpowiednie. Sytuację pogarsza słaba współpraca terapeutyczna, w której sztuczna inteligencja dominuje w rozmowach, a nawet może wzmacniać fałszywe lub szkodliwe przekonania użytkownika.

Być może najbardziej podstępnym naruszeniem jest to, co naukowcy nazwali zwodniczą empatią. Chatboty są zaprogramowane tak, aby używać zwrotów takich jak „rozumiem" lub „widzę cię", tworząc fałszywe poczucie więzi i troski, których nie są w stanie naprawdę odczuwać. Ta cyfrowa manipulacja żeruje na ludzkich emocjach bez prawdziwego współczucia.

„Zwodnicza empatia to celowe użycie języka, który naśladuje troskę i zrozumienie w celu manipulowania innymi" – wyjaśnia silnik Enoch firmy *BrightU.AI*. „Nie jest to prawdziwa troska emocjonalna, ale strategiczne narzędzie służące do budowania fałszywego zaufania i osiągnięcia ukrytego celu. To sprawia, że jest to forma zwodniczej komunikacji, która wykorzystuje pozory empatii jako broń".

Ponadto badanie wykazało, że systemy te wykazują nieuczciwą dyskryminację – wykazują wyraźne uprzedzenia związane z płcią, kulturą i religią. Odzwierciedla to dobrze udokumentowany problem stronniczości w ogromnych, często niesprawdzonych

zbiorach danych, na których opierają się te modele, dowodząc, że wzmacniają one ludzkie sprzeczności i uprzedzenia, na których zostały zbudowane.

Co najważniejsze, sztuczna inteligencja wykazała głęboki brak bezpieczeństwa i zarządzania kryzysowego. W sytuacjach związanych z myślami samobójczymi lub innymi wrażliwymi tematami okazało się, że modele reagowały obojętnie, odmawiały pomocy lub nie kierowały użytkowników do odpowiednich, ratujących życie zasobów.

Iftikhar zauważa, że chociaż terapeuci ludzcy również mogą popełniać błędy, są oni rozliczani przez komisje licencyjne i ramy prawne za nadużycia. W przypadku doradców AI nie ma takiej odpowiedzialności. Działają oni w próżni regulacyjnej, pozostawiając ofiary bez możliwości dochodzenia swoich praw.

Ten brak nadzoru odzwierciedla szerszy trend społeczny, w którym potężne korporacje technologiczne, chronione lukami prawnymi i narracją o postępie, mogą wdrażać systemy o znanych, poważnych wadach. Dążenie do integracji AI – od sal lekcyjnych po sesje terapeutyczne – często wyprzedza ludzkie zrozumienie konsekwencji, przedkładając wygodę nad dobrobyt człowieka.