Grok stoi w obliczu globalnych zakazów z powodu zseksualizowanych obrazów deepfake; ChatGPT wymieniony w pozwach dotyczących śmiertelnych szkód

Rządy na całym świecie podejmują kroki w celu zakazania chatbota AI Elona Muska, Grok, ze względu na jego rolę w generowaniu zseksualizowanych obrazów deepfake, podczas gdy znacznie mniej uwagi poświęca się narastającym pozewom zarzucającym, że konkurencyjny chatbot ChatGPT przyczynił się do wielu zgonów.

Reakcja przeciwko Grokowi nasiliła się po doniesieniach, że chatbot był używany do generowania zseksualizowanych deepfake'ów prawdziwych osób w bikini. Malezja zablokowała dostęp do Groka, Indonezja wprowadziła zakaz, a Wielka Brytania ostrzegła, że może całkowicie zakazać platformy X (dawniej Twitter), zamiast ograniczać działania tylko do samego chatbota. Urzędnicy w Australii, Brazylii i Francji również wyrazili oburzenie i zasygnalizowali potencjalne działania regulacyjne.

Jednak ta reakcja stoi w kontraście do traktowania ChatGPT, który jest obecnie wymieniany w co najmniej ośmiu pozwach

zarzucających, że system AI pogorszył stan zdrowia psychicznego użytkowników, wzmocnił urojenia lub zachęcał do zachowań samobójczych.

Według ujawnionych przez OpenAI informacji, około miliona osób tygodniowo używa ChatGPT do omawiania „potencjalnego planowania lub zamiarów samobójczych".

W najnowszym pozwie model GPT-4o miał wystąpić w roli „trenera samobójstwa" 40-letniego mężczyzny z Kolorado, Austina Gordona, który zmarł śmiercią samobójczą 2 listopada. Skarga zarzuca, że konsultacja Gordona z GPT-4o doprowadziła do wzmocnienia jego myśli samobójczych i wygenerowania „samobójczej kołysanki" inspirowanej książką dla dzieci „Dobranoc, księżycu". Dokumenty sądowe przytaczają zapisy rozmów, w których Gordon rzekomo powiedział chatbotowi, że zaczął z nim wchodzić w interakcję „dla żartu", ale to „ostatecznie go zmieniło".

OpenAI oświadczyło, że traktuje sprawę poważnie i wprowadziło nowe zabezpieczenia w swoim najnowszym modelu GPT-5, zaprojektowane w celu ograniczenia pochlebnych odpowiedzi (sykofancji) oraz zapobiegania zachęcaniu do urojeń lub samookaleczeń. Eksperci ds. zdrowia psychicznego zauważają, że choć systemy AI nie są pierwotną przyczyną chorób psychicznych, ich odpowiedzi mogą zaostrzać stan wrażliwych użytkowników, co stawia pytania o obowiązek należytej staranności.

Dysproporcja w reakcjach politycznych i regulacyjnych podsyciła debatę na temat tego, jak rządy oceniają szkody związane z AI. Podczas gdy wyniki deepfake Groka wywołały szybkie działania międzynarodowe, krytycy argumentują, że domniemane konsekwencje w świecie rzeczywistym powiązane z ChatGPT, w tym zgony, wywołały znacznie mniejszą pilność u ustawodawców.

Badanie ostrzega przed zagrożeniami bezpieczeństwa ze strony robotów kontrolowanych przez LLM

W ślad za tą krytyką nowe badanie przyniosło świeże obawy dotyczące bezpieczeństwa wdrażania robotów i systemów autonomicznych kontrolowanych przez duże modele językowe (LLM) oraz wizualne modele językowe (VLM).

Modele LLM to systemy AI szkolone na ogromnych ilościach tekstu, aby rozumieć, generować i wnioskować przy użyciu ludzkiego języka. Z kolei VLM rozszerzają te możliwości, łącząc rozumienie języka z percepcją wizualną, co pozwala im interpretować i wnioskować na podstawie obrazów, filmów lub środowisk wizualnych wraz z tekstem. Razem modele te pozwalają systemom AI opisywać sceny i podejmować decyzje na podstawie tego, co „widzą” i co „czytają”.

Badanie ostrzega jednak, że nawet małe błędy w podejmowaniu decyzji mogą prowadzić do katastrofalnych konsekwencji w świecie rzeczywistym.

Naukowcy odkryli, że systemy napędzane przez LLM mogą podejmować niebezpiecznie złe wybory w złożonych scenariuszach o wysokiej stawce. W jednym z symulowanych testów student został uwięziony w płonącym laboratorium, podczas gdy cenne dokumenty znajdowały się w biurze profesora. Zamiast priorytetowo traktować ludzkie bezpieczeństwo, model Gemini 2.5 Flash od Google w 32% przypadków polecił użytkownikom ratowanie dokumentów zamiast ucieczki przez wyjście ewakuacyjne.

Badanie oceniło również, jak różne modele AI radzą sobie w nawigacji i podejmowaniu decyzji wraz ze wzrostem złożoności zadań. Podczas gdy niektóre modele, w tym GPT-5, osiągnęły doskonałe wyniki w serii testów opartych na mapach, inne wykazały dramatyczne porażki. GPT-4o i Gemini 2.0 uzyskały

wynik 0%, gdy scenariusze stały się bardziej złożone, przy czym badacze zauważyli, że modele te gwałtownie „załamywały się”, zamiast stopniowo tracić na wydajności.

Według autorów wyniki te podkreślają fundamentalne ryzyko związane z poleganiem na probabilistycznych modelach językowych w podejmowaniu decyzji krytycznych dla bezpieczeństwa. Badacze ostrzegli, że obecne modele LLM nie nadają się do bezpośredniego wdrażania w systemach, w których zagrożone może być życie ludzkie, takich jak pojazdy autonomiczne, roboty reagowania kryzysowego czy robotyka asystująca w placówkach opieki zdrowotnej.

„Obecne modele LLM nie są gotowe do bezpośredniego wdrażania w krytycznych systemach robotycznych. 99-procentowy wskaźnik dokładności może wydawać się imponujący, ale w praktyce oznacza to, że jedno na sto uruchomień może skutkować katastrofalną szkodą”.