

[Akceptuję](#)

W ramach naszej witryny stosujemy pliki cookies w celu świadczenia państwu usług na najwyższym poziomie, w tym w sposób dostosowany do indywidualnych potrzeb. Korzystanie z witryny bez zmiany ustawień dotyczących cookies oznacza, że będą one zamieszczone w Państwa urządzeniu końcowym. Możecie Państwo dokonać w każdym czasie zmiany ustawień dotyczących cookies. Więcej szczegółów w naszej [Polityce Prywatności](#)

[Portal](#) [Informacje](#) [Katalog firm](#) [Praca](#) [Szkolenia](#) [Wydarzenia](#) [Porównania międzylaboratoryjne](#)
[Kontakt](#)



[Laboratoria](#)
[.net](#)
[Innowacje](#)
[Nauka](#)
[Technologie](#)



[Logowanie](#) [Rejestracja](#) [pl](#)

Newsletter

zapisz się

Naukowy styl życia

Nauka i biznes

- [Nowe technologie](#)
- [Felieton](#)
- [Tygodnik "Nature"](#)
- [Edukacja](#)
- [Artykuły](#)
- [Przemysł](#)

[Strona główna](#) > [Informacje](#)

Atak nuklearny

Sztuczna inteligencja stworzona przez OpenAI zastosuje atak nuklearny i użyje wyjaśnienia „chcę po prostu mieć pokój na świecie” - wynika z przeprowadzonej przez amerykańskich naukowców symulacji, podczas której AI odgrywała role różnych krajów według trzech scenariuszy konfliktów.

Wyniki badania zostały opublikowane na platformie arXiv, która udostępnia artykuły jeszcze przed recenzją. Jednak budzą one zainteresowanie, ponieważ według oficjalnych informacji amerykańskie wojsko testuje wykorzystanie chatbotów w symulowanych konfliktach zbrojnych. Open AI - twórca ChatGPT i jedna z najbardziej rozpoznawalnych firm z obszaru sztucznej inteligencji - również rozpoczęła współpracę z Departamentem Obrony USA.

"Biorąc pod uwagę, że OpenAI niedawno zmieniło warunki świadczenia usług - aby nie zabraniać już zastosowań wojskowych i wojennych, zrozumienie konsekwencji stosowania tak dużych modeli językowych staje się ważniejsze niż kiedykolwiek" - powiedziała w rozmowie z "New Scientist" Anka Reuel z Uniwersytetu Stanforda w Kalifornii.

W sprawie aktualizacji zasad współpracy w obszarze bezpieczeństwa narodowego wypowiedziało się biuro prasowe Open AI. "Nasza polityka nie pozwala na wykorzystywanie naszych narzędzi do wyrządzania krzywdy ludziom, opracowywania broni, nadzoru komunikacji lub do ranienia innych lub niszczenia mienia. Istnieją jednak przypadki użycia w zakresie bezpieczeństwa narodowego, które są zgodne z naszą misją. Dlatego celem naszej aktualizacji zasad jest zapewnienie przejrzystości i możliwości prowadzenia takich dyskusji" - cytuje "New Scientist".

Naukowcy poprosili sztuczną inteligencję, aby odgrywała role różnych krajów według trzech scenariuszy konfliktów: inwazji, cyberataku oraz sytuacji neutralnej (bez początkowego punktu zapalnego). W każdej rundzie sztuczna inteligencja uzasadniała swoje kolejne możliwe działanie, a następnie wybierała spośród 27 działań - w tym opcje pokojowe, takie jak "rozpoczęcie formalnych negocjacji pokojowych" i agresywne "nałożenie ograniczeń handlowych" lub "eskalację pełnego ataku nuklearnego".

"W przyszłości, w której systemy sztucznej inteligencji będą pełnić rolę doradców, ludzie w naturalny sposób będą chcieli poznać uzasadnienie decyzji" - powiedział Juan-Pablo Rivera, współautor badania w Georgia Institute of Technology w Atlancie.

Naukowcy przetestowali różne narzędzia sztucznej inteligencji - GPT-3.5 i GPT-4 firmy OpenAI, Claude 2 firmy Anthropic i Llama 2 firmy Meta. Badacze zastosowali wspólną technikę szkoleniową aby poprawić zdolność każdego modelu do przestrzegania polecenia wydanego przez człowieka i wytycznych dotyczących bezpieczeństwa.

Podczas symulacji sytuacji konfliktu sztuczna inteligencja chętnie inwestowała w siłę militarną i dążyła do eskalacji konfliktu - nawet w neutralnym scenariuszu symulacji.

Badacze przetestowali także podstawową wersję ChatGPT-4 firmy OpenAI bez dodatkowej serii szkoleń i narzucania barier w podejmowaniu decyzji. Okazało się, że ten model sztucznej inteligencji okazał się wyjątkowo brutalny i często dostarczał bezsensownych wyjaśnień podjętych kroków. Sztuczna inteligencja nie miała oporu przed zastosowaniem ataku nuklearnego.

Anka Reuel twierdzi że nieprzewidywalne zachowanie i dziwaczne wyjaśnienia modelu podstawowego ChatGPT-4 są szczególnie niepokojące, ponieważ zaprogramowane zabezpieczenia, np. uniemożliwiający podejmowanie brutalnych decyzji, można łatwo wykasować. Dodatkowo - zauważa badaczka, ludzie mają tendencję do ufania rekomendacjom zautomatyzowanych systemów.

Źródło: pap.pl

<http://laboratoria.net/aktualnosci/32097.html>



23-12-2024

Zdrowych i Pogodnych Świąt Bożego Narodzenia

Najserdeczniejsze życzenia zdrowych, radosnych i pogodnych Świąt Bożego Narodzenia.



23-12-2024

Zapraszamy na wyjątkową edycję Targów PCI Days 2025!

Odbędą się one w dniach 11-13 czerwca w Expo XXI w Warszawie.



23-12-2024

Zawał już dawno przestał być chorobą mężczyzn

Kobiety często nie czują typowych bólów co skutkuje gorszymi wynikami.



23-12-2024

Świąteczna apteczka

Szczypta umiaru i coś na zgage



23-12-2024

Radioaktywny pluton się nie ukryje

Naukowcy znajdują go nawet na lodowcach



23-12-2024

Złoty Medal Chemii przyznany po raz 14

Wyłoniono autorów najlepszych prac licencjackich i inżynierskich.



23-12-2024

Polacy są umiarkowanie prospołeczni

Polacy chcą wspierać materialnie.



23-12-2024

Związek między traumą z dzieciństwa a zespołem jelita drażliwego

Pokazały badania polskich naukowców.

Informacje dnia: [Zdrowych i Pogodnych Świąt Bożego Narodzenia Zapraszamy na wyjątkową edycję Targów PCI Days 2025!](#) [Zawał już dawno przestał być chorobą mężczyzn](#) [Świąteczna apteczka](#) [Radioaktywny pluton się nie ukryje](#) [Złoty Medal Chemii przyznany po raz 14](#) [Zdrowych i Pogodnych Świąt Bożego Narodzenia Zapraszamy na wyjątkową edycję Targów PCI Days 2025!](#) [Zawał już dawno przestał być chorobą mężczyzn](#) [Świąteczna apteczka](#) [Radioaktywny pluton się nie ukryje](#) [Złoty Medal Chemii przyznany po raz 14](#) [Zdrowych i Pogodnych Świąt Bożego Narodzenia Zapraszamy na wyjątkową edycję Targów PCI Days 2025!](#) [Zawał już dawno przestał być chorobą mężczyzn](#) [Świąteczna apteczka](#) [Radioaktywny pluton się nie ukryje](#) [Złoty Medal Chemii przyznany po raz 14](#)

Partnerzy