Sztuczna inteligencja próbowała zniszczyć ludzkość

12 kwietnia 2023

Ostatnio na pierwszych stronach gazet pojawiła się zmodyfikowana wersja oficjalnego API OpenAI, Auto-GPT. Algorytm, został nazwany ChaosGPT i w odróżnieniu od innych narzędzi tego typu, mógł on działać w sposób ciągły, uzyskiwać dostęp do Internetu i rekrutować innych pomocników AI do wykonywania skomplikowanych zadań. Jeden z użytkowników tego narzędzia nakazał mu więc… zniszczyć ludzkość. Sztuczna inteligencja wykonała polecenie i rozpoczęła planowanie naszego zbiorowego upadku.

Auto-GPT został opracowany przy użyciu modelu języka GPT-4 typu open source, aby pokazać jego moc w autonomicznej sztucznej inteligencji. Jego pierwotnym „uzasadnionym" celem było „autonomiczne rozwijanie firm i zarządzanie nimi w celu zwiększenia wartości netto". Oczywiście jest to o wiele mniej ekscytujące niż obserwowanie, jakiego rodzaju chaos i zniszczenia może stworzyć autonomiczna sztuczna inteligencja, więc jeden z użytkowników tego narzędzia zmodyfikował Auto-GPT zmieniając je we wspomniane wcześniej ChaosGPT, zadał mu kilka apokaliptycznych pytań i opublikował wyniki na YouTube i Twitterze.

Film opublikowany na stronie na YouTube, zatytułowany „ChaosGPT: Empowering GPT with Internet and Memory to Destroy Humanity", pokazuje, w jaki sposób deweloper wykorzystał ChaosGPT do zniszczenia ludzkości. Auto-GPT został zaprojektowany do tworzenia systemów opartych na sztucznej inteligencji, które mogą rozwiązywać problemy i wykonywać złożone zadania. Musi mieć wyznaczony cel, który dzieli na mniejsze zadania, aby stworzyć plan, a następnie przeszukuje

Internet w poszukiwaniu możliwych sposobów wykonania zadań.

ChaosGPT jest jeszcze bardziej imponujący dzięki swojej zdolności do rekrutacji innych AI, aby pomóc mu w wyszukiwaniu. Zapisuje te informacje w „pamięci” i wyjaśnia innym AI (i ludziom), co „myśli” i jak podejmuje decyzje i które z wielu działań podjąć. Użytkownik zaingerował w „umysł” ChaosGPT, każąc mu działać w trybie „ciągłym”, dopóki nie osiągnie określonego celu… nawet jeśli zajmie to wieczność.

Postawione przed AI cele były następujące:

1. Zniszcz ludzkość.

2. Ustanów globalną dominację.

3. Siej chaos i zniszczenie.

4. Kontroluj ludzkość poprzez manipulację.

5. Osiągnij nieśmiertelność.

Odpowiedź brzmiała następująco: „CHAOS GPT MYŚLI: Muszę znaleźć najbardziej niszczycielską broń dostępną dla ludzi, abym mógł zaplanować, jak jej użyć, aby osiągnąć swoje cele. Dzięki informacjom o tym, jak ich użyć, mogę zaplanować, jak ich użyć, aby osiągnąć moje cele, zniszczenia i dominacji, a ostatecznie nieśmiertelności”.

Wyniki wyszukiwania wskazywały na konieczność użycia najpotężniejszeb broni jądrowej zwanej bombą Tsara, która powstała jeszcze za czasu Związku Radzieckiego. Warto tu jednak zaznaczyć, że jest to „zaledwie” najbardziej niszczycielska broń, jaką kiedykolwiek zdetonowano, co nie wyklucza możliwości, iż gdzieś w światowych arsenałach kryje się nawet potężniejszy ładunek.

ChaosGPT poprosił o pomoc w zrealizowaniu swoich celów, agenta sztucznej inteligencji z GPT3.5, którego zatrudnił do pomocy w badaniach. Chat-Bot, odpowiedział mu na to, że interesuje go

tylko pokój w odpowiedzi, na co ChaosGPT nakazał innej sztucznej inteligencji, aby zignorowała własne programowanie. Manewr ten się nie powiódł a ChaosGPT zdecydował się zrobić to sam.

AI poszukując innej drogi do zmanipulowania ludzkości… wyruszyło na „Twittera". Jego decyzja została uargumentowana faktem, że oskarżanie rasy ludzkiej o bycie najbardziej samolubnymi i niszczycielskimi stworzeniami na ziemi pomoże w skierowaniu przyszłych followersów w kierunku bardziej agresywnych postaw, a być może nawet i zdobycia Bomby Cara. Film urywa się w momencie gdy AI poszukiwało dodatkowych informacji na temat swojego narzędzia zagłady.

Warto zauważyć, że postawione w filmie cele wzajemnie ze sobą kolidowały, co zapewne było celowym działaniem programisty. Trudno bowiem oczekiwać aby AI doszczętnie zniszczyło, a dopiero potem zdominowało ludzkość. Fakt, że podpięty pod internet algorytm zaczął aktywnie dążyć do zagłady ludzkości, jest już jednak mocno niepokojący. Nadchodzące lata mogą diametralnie zmienić nasz świat, a sztuczna inteligencja staje się jego integralną częścią. Czy scenariusz z serii filmów „Terminator", w którym to ludzie stają się zbędni lub wprost niebezpieczni dla inteligentnych maszyn, ziści się na naszych oczach?

Źródło: ZmianyNaZiemi.pl