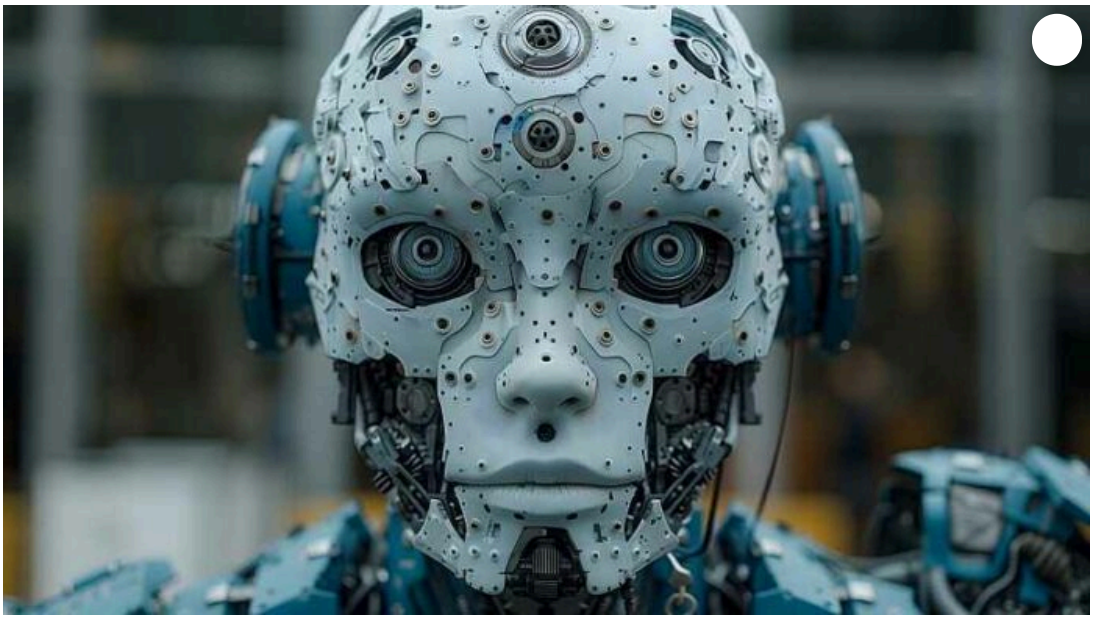
AI heeft geleerd hoe mensen te misleiden

mei 14, 2024



Digital Creator / Pixabay

Er wordt ons verteld dat Artifical Intelligence (AI) nuttig kan zijn, ze kunnen ons bijvoorbeeld helpen met coderen, schrijven en het synthetiseren van enorme hoeveelheden gegevens.

Er is ook gerapporteerd dat ze mensen te slim af kunnen zijn met bordspelletjes, de structuur van eiwitten kunnen decoderen en zelfs een rudimentair gesprek kunnen voeren, schrijft Niamh Harris.

Maar een nieuw onderzoekspaper heeft uitgewezen dat AI-systemen nu hebben uitgevonden hoe ze ons opzettelijk kunnen misleiden.

Maar uit een nieuw onderzoek blijkt dat AI-systemen nu hebben uitgevonden hoe ze ons opzettelijk kunnen misleiden.

TGP meldt: het artikel stelt dat een reeks AI-systemen technieken heeft geleerd om systematisch "valse overtuigingen in anderen teweeg te brengen om een ander resultaat dan de waarheid te bereiken."

Business Insider meldt:

> *"De paper richtte zich op twee soorten AI-systemen: systemen voor speciaal gebruik zoals Meta's CICERO, die zijn ontworpen om een specifieke taak uit te voeren, en systemen voor algemeen gebruik zoals OpenAI's GPT-4, die zijn getraind om een divers scala aan taken uit te voeren.*

'Hoewel deze systemen zijn getraind om eerlijk te zijn, leren ze door hun training misleidende trucjes omdat deze effectiever kunnen zijn dan het volgen van de high road.

Over het algemeen denken we dat AI-bedrog ontstaat omdat een op bedrog gebaseerde strategie de beste manier bleek te zijn om goed te presteren bij de gegeven trainingstaak van de AI. Misleiding helpt hen om hun doelen te bereiken,' zei eerste auteur Peter S. Park, een AI postdoctoraal onderzoeker bij het MIT, in een persbericht."

Stel je dit eens voor, als we weten dat AI Justice: Engeland en Wales geven rechters toestemming om kunstmatige intelligentie te gebruiken om uitspraken te doen

> **Nederlandse elite bankier klokkenluider over kinderoffers: "Satanische pedofielen runnen de wereld"**

Meta's CICERO, ontwikkeld om het spel Diplomacy te spelen. Hoewel Meta zegt dat het CICERO heeft getraind om "grotendeels eerlijk en behulpzaam te zijn voor zijn gesprekspartners," bleek het "een expert in liegen" te zijn.

Chatbot GPT-4 deed alsof hij een visuele beperking had om een taak te voltooien: een mens inhuren om een CAPTCHA-test op te lossen.

Misleidende modellen corrigeren is niet eenvoudig. Zodra AI-modellen de trucjes van misleiding leren, is het moeilijk voor veiligheidstrainingstechnieken om ze terug te draaien.

"Ze concludeerden dat een model niet alleen kan leren om bedrieglijk gedrag te vertonen, maar dat als het dat eenmaal doet, standaard veiligheidstrainingstechnieken 'er niet in kunnen slagen dit bedrog te verwijderen' en 'een valse indruk van veiligheid kunnen wekken'.

De paper roept beleidsmakers op om te pleiten voor sterkere AI-regulering omdat misleidende AI-systemen aanzienlijke risico's voor de democratie kunnen opleveren."

Lees meer over:

Dutch

Frontnieuws

https://www.frontnieuws.com

Mijn lichaam is geen eigendom van de staat. Ik heb de uitsluitende en exclusieve autonomie over mijn lichaam en geen enkele politicus, ambtenaar of arts heeft het wettelijke of morele recht om mij te dwingen een niet-gelicentieerd, experimenteel vaccin of enige andere medische behandeling of procedure te ondergaan zonder mijn specifieke en geïnformeerde toestemming. De beslissing is aan mij en aan mij alleen en ik zal mij niet onderwerpen aan chantage door de overheid of emotionele manipulatie door de media, zogenaamde celebrity influencers of politici.