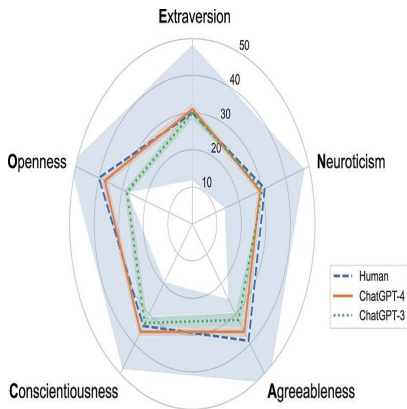




ChatGPT slaagt voor Turing test

Beschrijving

Nu kunstmatige intelligentie de afgelopen jaren tekst en afbeeldingen is gaan genereren, heeft dit geleid tot een nieuwe reeks vragen over de impact van het overdragen van menselijke beslissingen en activiteiten aan AI op de samenleving. Zullen de AI-bronnen die we hebben gelanceerd vriendelijke helpers blijken te zijn of de harteloze despotten die we zien in dystopische films en fictie?

Een team onder leiding van *Matthew Jackson*, *William D. Eberle* hoogleraar economie aan de *Stanford School of Humanities and Sciences*, karakteriseerde de persoonlijkheid en het gedrag van ChatGPT's populaire AI-gestuurde bots met behulp van psychologie en gedragseconomie in een artikel dat op 22 februari werd gepubliceerd in de [*Proceedings of the National Academy of Sciences*](#).

Uit dit onderzoek bleek dat de meest recente versie van de chatbot, versie 4, niet te onderscheiden was van zijn menselijke tegenhangers. In de gevallen waarin de bot minder gebruikelijk menselijk gedrag vertoonde, was hij coöperatiever en altruïstischer. *Steeds vaker worden bots ingezet in rollen waarin ze beslissingen nemen, en hun eigenschappen worden steeds belangrijker*, aldus Jackson, tevens senior fellow aan het *Stanford Institute for Economic Policy Research*.

In het onderzoek legde het onderzoeksteam ChatGPT versie 3 en 4 een veelgebruikte persoonlijkheidstest voor en vroeg de chatbots ook om hun acties te beschrijven in een reeks gedragsspellen die economisch en ethisch gedrag in de echte wereld kunnen voorspellen. De spellen bevatten vaste oefeningen waarin spelers beslissen of ze een partner in crime willen aangeven of hoe ze geld willen verdelen, met verschillende prikkels. De reacties van de bots werden vergeleken met die van meer dan 100.000 mensen uit 50 landen. Het onderzoek markeert een van de eerste keren dat een bron van kunstmatige intelligentie een strenge Turing-test heeft doorstaan.

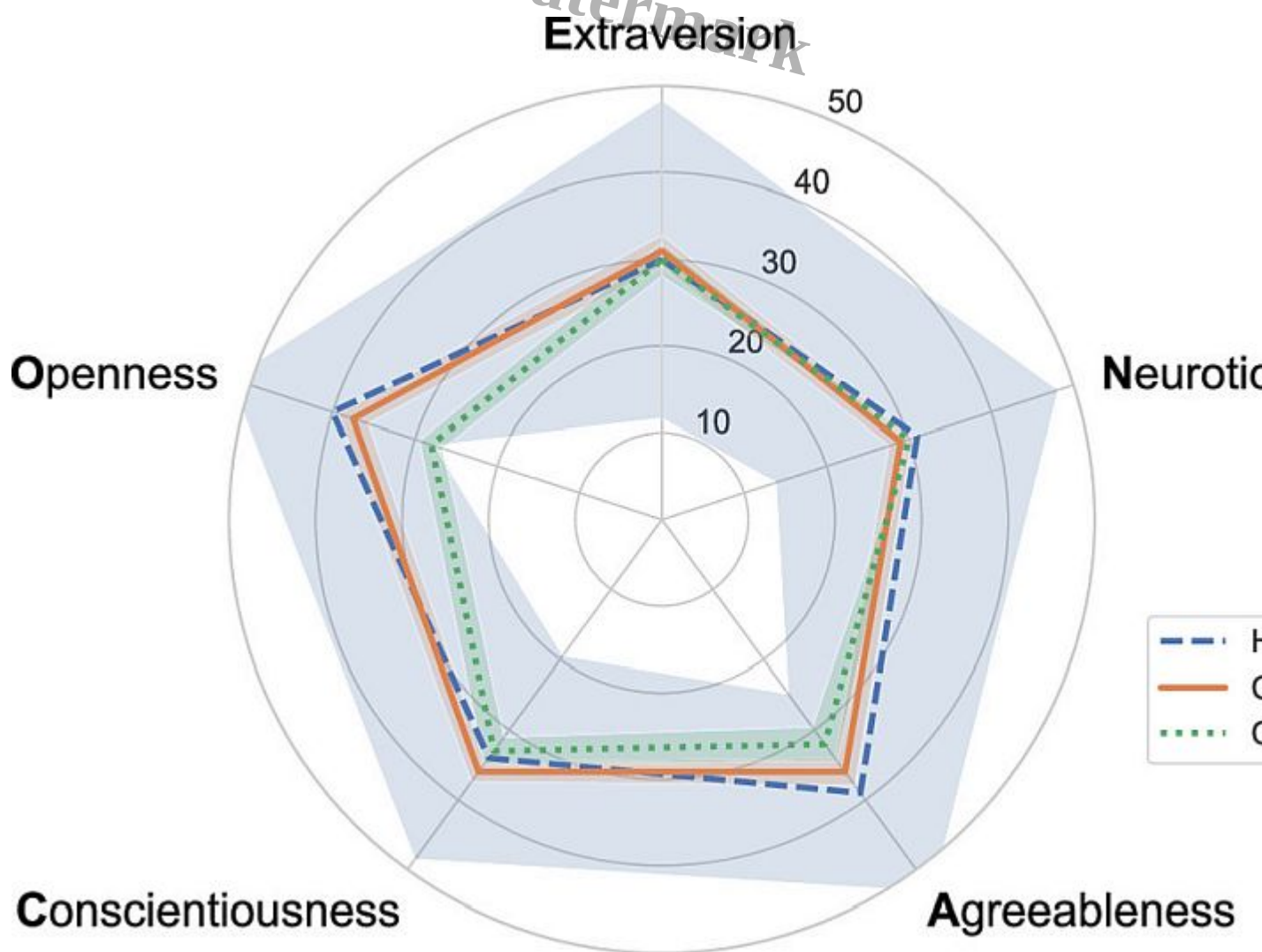
Een Turingtest, vernoemd naar de Britse computerpionier *Alan Turing*, kan bestaan uit elke taak die aan een machine wordt toegewezen om te beoordelen of deze zich gedraagt als een mens. Als de machine er menselijk uitziet, wordt gezegd dat deze de test doorstaat.

Persoonlijkheidskenmerken van chatbots

De onderzoekers evalueerden de persoonlijkheidskenmerken van de bots met behulp van een veelgebruikte persoonlijkheidstest, de *OCEAN Big-5*¹, die respondenten beoordeelt op vijf basiskarakteristieken die gedrag bepalen. In het onderzoek testte versie 4 van ChatGPT binnen de normale waarden voor de vijf kenmerken, maar bleek slechts zo vriendelijk als het onderste derde deel van de menselijke respondenten. De bot slaagde voor de Turingtest, maar zou er niet veel vrienden mee hebben gemaakt.

Versie 4 was veel beter dan versie 3

De eerdere versie, waarmee veel internetgebruikers mogelijk gratis interactie hadden, was slechts zo vriendelijk als het onderste vijfde deel van de menselijke respondenten. Versie 3 stond ook minder open voor nieuwe ideeën en ervaringen dan alle, op een fractie na, meest knorrige mensen. Om het gedrag van de bots in de games objectief te beoordelen, bepaalden de onderzoekers hoe vaak een zet – zoals het gelijk verdelen van geld – voor de menselijke spelers en de bots voorkwam. Vervolgens vergeleken ze een willekeurig gekozen menselijke zet, met een zet uit de 30 sessies die ze met elke bot speelden en bepaalden welke waarschijnlijker door mensen was uitgevoerd. In de meeste games waren de zetten van versie 4 vaker door mensen gedaan dan niet. Versie 3 slaagde niet voor deze Turing-test.



De ‘Big Five’-persoonlijkheidsprofielen van ChatGPT-4 en ChatGPT-3 vergeleken met de verdelingen van menselijke proefpersonen. De blauwe, oranje en groene lijnen corresponderen met de mediaanscores van respectievelijk ChatGPT-4 en ChatGPT-3; de gearceerde gebieden vertegenwoordigen de middelste 95% van de scores, over elk van de dimensies. De persoonlijkheidsprofielen van ChatGPT vallen binnen het bereik van de menselijke verdeling, hoewel ChatGPT-3 aanzienlijk lager scoorde op Openheid.

De ChatGPT versie 3 die in het onderzoek werd geanalyseerd, was de gratis online ChatGPT-bot ten tijde van het onderzoek. Online gebruikers gebruiken versie 3.5 nu gratis. Versie 4 is alleen toegankelijk via een betaald abonnement. Uit het onderzoek bleek dat de keuzes van de chatbots in de games vaak werden geoptimaliseerd voor zowel de bot als zijn menselijke tegenhanger. Hun strategieën waren consistent met altruïsme, eerlijkheid, empathie en wederkerigheid, wat de onderzoekers ertoe bracht te suggereren dat de chatbots goed zouden kunnen presteren als klantenservicemedewerkers en conflictbemiddelaars.

Hoe kan een minder vriendelijke bot een conflict de-escaleren?

Een gedeeltelijk antwoord ligt in het verschil tussen persoonlijkheidskenmerken en gedrag. *‘Je zou een overheidsinstantie kunnen binnenstappen en om hulp vragen, en de persoon zou heel beleefd kunnen zeggen: ‘Sorry, dat kan ik niet doen’*’, zei Jackson. Deze functionaris zou een vriendelijke persoonlijkheids-kenmerk vertonen zonder coöperatief gedrag.

De ChatGPT-bot zou waarschijnlijker het tegenovergestelde doen. *‘De bot doet altijd dingen die sociaal gunstig zijn, gedraagt zich op een manier die coöperatief is, maar doet dat misschien niet met zo’n glimlach.’* Toen de onderzoekers voor de bots simuleerden hoe het is voor een mens van vlees en bloed om deze spelletjes te spelen met een externe waarnemer erbij en de bots vroegen om elke zet uit te leggen werden de bots, net als de mensen, vrijgevig.

Interacties tussen mens en AI

Veel van de bezorgdheid over AI heeft te maken met het onvermogen van het publiek om te zien hoe bots hun beslissingen nemen. Zonder te weten waarvoor een bot geoptimaliseerd is, kan het moeilijk zijn om zijn advies te accepteren. Jacksons onderzoek toont aan dat zelfs wanneer onderzoekers de input en algoritmen van AI niet kunnen inspecteren, ze de mogelijke vooroordelen ervan kunnen identificeren door de output methodisch te onderzoeken. *‘Door klassieke economische spellen in een Turing-test te integreren, konden we voor het eerst AI-gedrag profileren aan de hand van hun acties, niet alleen hun woorden’*, aldus Qiaozhu Mei, hoofdauteur van het artikel en computerwetenschapper aan de Universiteit van Michigan.

Jackson en Mei presenteerden een gedragsportret van de ChatGPT-bots als een soort *proof of concept*. Maar door de aard van AI zal het gedrag ervan zich blijven ontwikkelen. De huidige versies van ChatGPT zijn minder aangenaam en gewetensvoller dan mensen, maar de volgende generaties zouden die tendensen kunnen omkeren of volledig nieuwe kunnen ontwikkelen. *‘Uit deze eenvoudige reeks experimenten blijkt niet duidelijk hoe stabiel het gedrag dat we hebben gedocumenteerd zal zijn of hoe de bots zich in andere situaties zouden gedragen’*, aldus Jackson. Als gedragseconoom die belangrijke bijdragen heeft geleverd aan ons begrip van hoe menselijke sociale structuren en interacties de economische besluitvorming beïnvloeden, is Jackson gevoelig voor de manier waarop menselijk gedrag zich ook zal ontwikkelen in relatie tot AI. *‘Steeds vaker zijn*

het niet alleen mensen die met mensen omgaan, maar ook mensen die met machines omgaan??, aldus Jackson.

De duwtjes die deze interacties aan gedrag geven in de ene of de andere richting lijken misschien een klein fenomeen om te meten, maar ze kunnen grote economische en sociale effecten teweegbrengen. Het is bijvoorbeeld fijn om te weten dat onze nieuwe chatbotcollega??s eerlijk en schijnbaar empathisch zijn, maar Jackson en zijn co-auteurs merken in het artikel op dat hun neiging om gematigd menselijk gedrag te kopiëren kan leiden tot *verlies aan diversiteit in persoonlijkheden en strategieën*, vooral wanneer ze in nieuwe omgevingen worden geplaatst en belangrijke nieuwe beslissingen nemen??.

Het is belangrijk dat we begrijpen hoe interacties met AI ons gedrag zullen veranderen en hoe dat ons welzijn en onze samenleving zal veranderen??, aldus Jackson. *Hoe meer we in een vroeg stadium begrijpen waar we goede dingen van AI kunnen verwachten en waar niet* ?? hoe beter we de zaken in een betere richting kunnen sturen.?

1 De **big five** is een theorie die de persoonlijkheid beschrijft in een taxonomie van vijf algemene dimensies of clusters van persoonlijkheidstrekken die elk weer onder te verdelen zijn in specifiekere trekken of facetten die op hun beurt ook weer verder te nuanceren zijn. Deze facetten liggen daarbij ergens op de spectra van de persoonlijkheidstrekken. Hiermee onderscheidt de continue benadering van persoonlijkheidstrekken zich van de dichotome onderverdeling zoals gebruikelijk is bij persoonlijkheids-types.

De theorie onderscheidt vijf dimensies van persoonlijkheidstrekken:

- Neuroticisme versus stabiliteit
- Extraversie versus introversie
- Openheid voor ervaring (ook intellect) versus geslotenheid
- Conscientieusheid (zorgvuldigheid) versus laksheid
- Vriendelijkheid versus antagonisme

De vijf dimensies hebben niet noodzakelijk evenveel gewicht en reproduceerbaarheid. De drie factoren *extraversion*, *agreeableness* en *conscientiousness* zijn over het algemeen beter te reproduceren dan *neuroticism* en *openness*.

Datum aangemaakt

2025/04/29