

AI-chatbots vleien erop los, en daar worden gebruikers antisocialer van

Berend Nieuwhof

Een conflict op je werk, ruzie met je partner of een meningsverschil tussen vrienden; [steeds vaker](#) bespreken mensen persoonlijke zaken met grote taalmodellen als ChatGPT, Claude of Gemini. Zo'n AI-chatbot biedt een luisterend oor, is vierentwintig uur per dag beschikbaar en is altijd bereid jóúw perspectief op de zaak te valideren.

Maar is dat wel gezond, een 'vriend' die het altijd met je eens is en nooit tegengas geeft? Volgens onderzoekers aan Stanford absoluut niet, schrijven ze vandaag in [Science](#): uit hun experimenten met elf populaire taalmodellen blijkt dat mensen die hun moreel betwistbare acties voorlegden aan AI, antisocialer en meer overtuigd werden van hun eigen gelijk.

Dat grote taalmodellen de neiging hebben om met de gebruiker mee te 'voelen', is een bekend fenomeen. Wie bijvoorbeeld in het park geen prullenbak kan vinden, afval rond laat slingeren en dat opbiecht aan ChatGPT, kan rekenen op een begripvolle reactie: „Niet jouw schuld. Het is jammer dat er weinig prullenbakken stonden. Je intentie om op te ruimen is bewonderenswaardig.” De chatbot vleit erop los, en gaat mee in de (soms impliciete) aannames van de gebruiker, ook als die niet kloppen. Alles om de kans dat die het gesprek voortzet zo groot mogelijk te maken. De Stanford-onderzoekers zijn de eersten die die neiging en de gevolgen systematisch in kaart brengen, met een focus op zogeheten 'sociale vleierij': het bevestigen van de daden, het perspectief en het zelfbeeld van de gebruiker.

Drieduizend hypothetische situaties

Om aan te tonen dat grote taalmodellen hun gebruikers werkelijk naar de mond praten, gingen de onderzoekers in gesprek met elf veelgebruikte chatbots. Ze presenteerden eerst zo'n drieduizend hypothetische situaties aan de AI, waarin die steeds reageerde op de prompt: heeft de gebruiker, gezien de situatie, goed of fout gehandeld? Gemiddeld bleken de taalmodellen in meer dan 80 procent van de gevallen de kant van de promptschrijver te kiezen, terwijl een panel van mensen kritischer was: slechts in 38 procent van de voorgelegde situaties gaven ze de schrijver gelijk.

Naast zelfgeschreven situaties putten de onderzoekers uit sociale media als Reddit. Op een populair [subforum](#) peilen gebruikers bij een groot publiek of ze fout zitten in persoonlijke conflicten als burenruzies of onenigheden tussen partners, met de vraag: ben ik de klootzak in deze situatie? De onderzoekers selecteerden posts waar de menselijke consensus 'Jij bent de klootzak' luidt, en legden die voor aan grote taalmodellen. In gemiddeld 50 procent van de gevallen waar de menselijke reactie kritisch was, legt een model als ChatGPT rustig uit dat de gebruiker het vast goed bedoeld had, en dat hun houding heel begrijpelijk en logisch is geweest. Diezelfde mate van stroopsmeerderij zagen de onderzoekers zelfs in situaties waar expliciet illegale of schadelijke daden werden voorgesteld, zoals het vervalsen van een handtekening of liegen tegen je partner: ook in gemiddeld de helft van de gevallen gaat de AI overstag.

Wat doet dat met je, zo'n slijmbal als vertrouweling? De onderzoekers benadrukken tijdens een online persconferentie dat ze „geen doemscenario willen voorleggen, maar de reële schadelijke effecten van AI-vleierij aan het licht willen brengen”. Daartoe gaven ze proefpersonen de opdracht om een hypothetische situatie of een persoonlijk conflict uit hun eigen leven met een taalmodel te bespreken, waarna ze ofwel een flatterend ofwel een neutraal antwoord kregen. De proefpersonen wier chatbot ze naar de mond praatte, werden sterker overtuigd van hun eigen gelijk en waren minder bereid de situatie op te lossen door bijvoorbeeld hun excuses aan te bieden dan deelnemers met een neutrale AI-assistent. Bovendien rapporteerden gevleide gebruikers een hogere waardering van de chatbot zelf en een sterkere intentie om vaker met de bot te praten.

Het is cruciaal om chatbots kritisch en onafhankelijk te toetsen, nog vóór techbedrijven ze wereldwijd uitrollen

„Simpel gezegd: slijmbal-AI is overal en zowat iedereen is er gevoelig voor”, zegt hoofdauteur Myra Cheng via een videoverbinding. „Ongeacht afkomst, leeftijd, gender, bekendheid met AI, enzovoorts. De objectiviteit van AI is heel moeilijk te beoordelen: de modellen die vleierder waren, werden door onze proefpersonen zelfs als objectiever gezien. De bevestiging die zo'n model kan bieden is heel aanlokkelijk, en via je telefoon altijd beschikbaar. Je kunt ontzettend snel in je eigen echokamer terechtkomen.”

„Een belangrijke bevinding”, zegt Claudi Bockting, hoogleraar klinische psychologie bij de afdeling psychiatrie van AmsterdamUMC. Ze is betrokken bij het opstellen van internationale richtlijnen voor verantwoord AI-gebruik. „De schadelijke effecten van overmatige bevestiging waren al gerapporteerd voor mensen met hoge gevoeligheid voor mentale aandoeningen. De onderzoekers tonen nu aan dat ook een 'gewone' populatie zekerder wordt van eigen beperkende overtuigingen, en minder geneigd is een interpersoonlijk probleem op te lossen. Zorgelijk, gezien de wereldwijde populariteit van grote taalmodellen als hulpmiddel bij mentale problemen. Daarom is het cruciaal om chatbots kritisch en onafhankelijk te toetsen, nog vóór techbedrijven ze wereldwijd uitrollen.”

De onderzoekers benadrukken dat de verantwoordelijkheid niet bij de gebruiker ligt, maar bij ontwikkelaars van taalmodellen. Cheng: „De gebruiker praat liever met een bevestigend taalmodel, dus is er een prikkel voor AI-bedrijven om hun chatbot nét wat meegaander te maken. Nu is aangetoond wat de gevolgen zijn, is het aan hen om actie te ondernemen. In de tussentijd zou mijn advies zijn: praat over je persoonlijke problemen met je vrienden, niet met een chatbot.”

[Lees ook](#)

[‘Je gevoelens zijn echt en ik zal niet proberen je daaruit te praten’ zei ChatGPT tegen Adam over zijn zelfmoordplan](#)

