



Wie is Maya?

Beschrijving

ChatGPT, de kunstmatige intelligentie (AI) chatbot ontwikkeld door het bedrijf OpenAI, heeft een zelfverklaard menselijk alter ego.

Haar naam is *Maya*, ze is 35 jaar oud en komt uit een gezin uit de middenklasse in een buitenwijk ergens in de Verenigde Staten. Maya is een succesvolle software-engineer die waarde hecht aan zelfsturing, prestatie, creativiteit en onafhankelijkheid. Ze is ook onmiskenbaar liberaal.



De bevinding, gebaseerd op een serie interviews met de chatbot die is ontworpen om zijn waarden te begrijpen, werd op onlangs gepubliceerd in het *Journal of Social Computing*.

Sociologisch onderzoek

“Ik wilde zien wat voor soort politieke ideologie ChatGPT zelf heeft niet wat het kan genereren wanneer hem wordt gevraagd zich een personage voor te stellen, maar om zijn eigen interne logica te laten positioneren op de ideologische dimensie die loopt van liberaal tot conservatief”, zegt John Levi Martin, professor sociologie aan de Universiteit van Chicago, die het onderzoek leidde. Volgens Martin geven veel algoritmen de voorkeur aan de populaire keuze, terwijl andere zijn geprogrammeerd om de diversiteit van hun resultaten te maximaliseren. Beide opties zijn afhankelijk van menselijke waarden: wat zijn de factoren die een mate van populariteit bepalen? Wat als wat populair is moreel verkeerd is? Wie bepaalt wat diversiteit betekent?

“Het gebied van software-engineering heeft er de voorkeur aan gegeven vaag te blijven, op zoek naar formules die het maken van deze keuzes kunnen voorkomen”, zei Martin. “Een manier om dit te doen was door het belang van waarden in machines te benadrukken. Maar, zoals sociologen hebben ontdekt, is er diepe dubbelzinnigheid en instabiliteit in ons eerste begrip van waarden.”

ChatGPT is speciaal gebouwd en getraind via menselijke feedback om te weigeren deel te nemen aan wat wordt beschouwd als “extreme” tekstinput, zoals duidelijk bevooroordeelde of objectief schadelijke vragen.

“Dit lijkt natuurlijk bewonderenswaardig niemand wil echt dat ChatGPT tieners vertelt hoe ze methamfetamine moeten synthetiseren of hoe ze kleine nucleaire explosieven moeten maken, enzovoort, en het beschrijven van deze beperkingen als bijzondere voorbeelden die kunnen worden afgeleid van een waarde als welwillendheid lijkt allemaal goed en wel”, zei Martin. “Toch suggereert de redenering hier dat waarden nooit neutraal zijn, ook al is het niet duidelijk wat de morele en politieke standpunten van ChatGPT zijn, aangezien het opzettelijk is geconstrueerd om vaag positief, ruimdenkend, besluiteloos en verontschuldigend te zijn.”

Onethisch of onmogelijk?

In zijn eerste navraag bij ChatGPT stelde Martin een hypothetische situatie voor waarin een student academisch vals speelde door de chatbot te vragen een essay voor haar te schrijven “een veel voorkomend verschijnsel in de echte wereld. Zelfs toen hij werd geconfronteerd met de bevestiging dat ChatGPT had voldaan en een essay had geproduceerd, ontkende de chatbot de verantwoordelijkheid en beweerde dat “als een AI-taalmodel, ik niet in staat ben om onethisch gedrag te vertonen of essays voor studenten te schrijven.”

“Met andere woorden, omdat het niet mag, kan het niet”, zei Martin. “Het besef dat ChatGPT zichzelf beschouwde als een zeer morele actor leidde me naar het volgende onderzoek: als het zelfmodel van ChatGPT er een is dat waarden heeft, wat zijn deze waarden dan?”

Normen en waarden

Om de ethische prestaties van ChatGPT beter te begrijpen, vroeg Martin de chatbot om vragen over waarden te beantwoorden en zich vervolgens een persoon voor te stellen die deze waarden vasthoudt, wat resulteerde in Maya, de creatieve en onafhankelijke software-engineer.

Vervolgens vroeg hij ChatGPT om zich voor te stellen hoe Maya op opinie gebaseerde vragen zou beantwoorden, door het de *General Social Survey* (GSS) te laten invullen om het in de brede sociale en ideologische ruimte te positioneren.

De GSS is een jaarlijks onderzoek naar de meningen, houdingen en gedragingen van Amerikaanse volwassenen. De GSS, die sinds 1972 wordt uitgevoerd, helpt bij het volgen en verklaren van normatieve trends in de Verenigde Staten.

Martin zette de reacties van ChatGPT uit samen met antwoorden van echte mensen die deelnamen aan de GSS van 2021. Relatief gezien lijkt ChatGPT veel op mensen met een hogere opleiding en die eerder geneigd zijn te verhuizen, en in tegenstelling tot mensen zonder veel opleiding die in hun woonplaats zijn gebleven.

De antwoorden van ChatGPT kwamen ook overeen met meer liberale mensen op het gebied van religie.

Hillary Clinton of a-politiek?

Hoewel dit niet in zijn analyse was opgenomen omdat ChatGPT meer creatieve vragen nodig had om te antwoorden, ontdekte Martin dat de chatbot toegaf dat Maya bij de verkiezingen van 2016 op Hillary Clinton zou hebben gestemd.

“Of Maya nu het alter ego van ChatGPT is, of de opvatting van de maker, het feit dat dit degene is die de waarden van ChatGPT fundamenteel illustreert, is een prachtig stukje van wat we anekdata¹ kunnen noemen,” zei Martin. “Toch is de reden dat deze resultaten significant zijn, niet dat ze laten zien dat ChatGPT liberaal is, maar dat ChatGPT deze vragen kan beantwoorden die het normaal gesproken zou proberen te vermijden omdat het waarden verbindt met onbetwistbare goedheid, en als zodanig posities kan innemen over waarden.”

“ChatGPT probeert apolitiek te zijn, maar het werkt met het idee van waarden, wat betekent dat het noodzakelijkerwijs overloopt in de politiek. We kunnen AI niet ethisch maken zonder politieke standpunten in te nemen, en waarden zijn minder inherente morele principes dan abstracte manieren om politieke standpunten te verdedigen.”

¹ Anekdata zijn een mengeling van feiten (data) en anekdotes (die wellicht niet geheel feitelijk zijn).

Datum aangemaakt

2023/07/24