Dit computerprogramma schrijft zelf verhalen maar erg intelligent zijn die niet

Posted on 25 maart 2023 by Rik Smits

Het moet gezegd, het inmiddels veelbesproken taalprogramma ChatGPT dat in november 2022 de wereld verraste, levert ogenschijnlijk indrukwekkende communicatieve prestaties. Maar schijn bedriegt. ChatGPT is een ramp, doordat het ons werkelijkheidsbegrip en onze kennis ondermijnt. Het is een fundamenteel probleem dat vooralsnog zonder oplossing blijft.

Gelikt is het woord voor de teksten die ChatGPT produceert, want dit staaltje artificial intelligence (AI) excelleert in beleefde pluimstrijkerij. Het schrijft vlotte, nogal clichématige verhalen. Daar is het programma op getraind. Maar wie even verder kijkt, ziet dat ChatGPT vooral klatergoud verkoopt. Wat voorbeelden: journalist en FvD-horzel Chris Aalberts ChatGPT'de zichzelf en zag zich tot FvD-politicus gebombardeerd.

Toen ik hetzelfde deed, bleek mijn boek Rebellen geschreven door de schrijver

Rutger Bregman, mijn Dageraad door de Britse archeoloog Mark Pagel, en Dawn, de Engelse versie daarvan, door diens collega Stephen Mithen. Ikzelf werd als taalkundige bestempeld tot hoogleraar in eerst Utrecht en daarna Leiden, compleet met Vici-beurs en Spinoza-premie. Utrecht! Dat is wel de laatste universiteit waar ik iets mee heb, hoogleraar ben ik niet en die premie bleek bij Eveline Crone te horen, die weliswaar in Leiden in hersenen doet maar allerminst taalkundige is.

Glibberige taalformules

Kortom: ChatGPT grabbelt maar raak in de gigantische ballenbak van teksten waaruit het zijn gegevens put. De software zoekt databrokjes bij elkaar die op een of andere manier met de vraagstelling en met elkaar geassocieerd zijn, en strijkt die oceanische chaos glad onder een olievlek van glibberige taalformules.

Om te zien hoe ver ChatGPT daarin gaat, vroeg ik naar iets dat elke enigszins politiek bewuste Nederlander vrijwel onmiddellijk eerst als raar en vervolgens als fysiek onmogelijk herkent: de invloed van de ideeën van Pim Fortuyn op het denken van Boer Koekoek. De leider van de geruchtmakende Boerenpartij maakte immers vooral naam in de jaren zestig van de vorige eeuw, toen Fortuyn nog scholier en beginnend student was. Koekoek verdween al in 1981 roemloos uit de Tweede Kamer en overleed in 1987, ruim een decennium voordat Fortuyns politieke carrière zou beginnen.

Dat kan ChatGPT allemaal niet bommen. Opgewekt steekt het van wal met: ‘De ideeën van Pim Fortuyn, […] die in de vroege jaren 2000 opkwam, hebben een significante impact gehad op het politieke landschap van Nederland. Eén individu die [sic] sterk beïnvloed werd door de ideeën van Fortuyn was Hendrik Koekoek, een politicus die diende als leider van de Boerenpartij in de jaren 70 en 80.’

Kunstmatig intelligent fabuleert ChatGPT voort, ruim 400 woorden lang: ‘Koekoek […] was aangetrokken tot Fortuyns ideeën en begon ze op te nemen in zijn eigen politieke platform’. Hij had ‘een meer populistische en anti-establishment oriëntatie dan Fortuyn, die lid was van de liberale VVD-partij.’ Koekoek ‘zag Fortuyn als een belangrijke bondgenoot’, maar ‘sommige critici beschuldigden Koekoek ervan een xenofobische en intolerante [sic] wereldbeeld te hebben overgenomen en beweerden dat zijn omarming van Fortuyns ideeën een teken was van zijn eigen radicalisering.’

Binnenwereld

De oorzaak voor dit misleidende gezwatel is dat ChatGPT geen andere wereld kent dan de teksten waarop het getraind is en waar het zijn gegevens uit tapt. Fortuyn, Koekoek en appeltaart betekenen voor ChatGPT bijna net zo weinig als x en y in de algebra: het zijn vrijelijk inwisselbare, willekeurige en betekenisloze bouwsteentjes. ChatGPT is daardoor wel kunstmatig, maar niet erg intelligent.

In ons, mensen, bestaat ook een binnenwereld van woorden en begrippen waaruit we patronen, gedachten vormen, min of meer net als ChatGPT dat doet. In die binnenwereld zijn Donald Duck en Hamlet precies even echt als Caroline van der Plas en je fiets. Maar, anders dan bij ChatGPT, wordt onze binnenwereld in toom gehouden doordat wij weten dat er ook een buitentalige wereld bestaat, en we kennen allerlei wetten die daarin gelden, ook al zijn we ons daar niet direct van bewust.

Die deels impliciete kennis is deel van ons biologisch erfgoed, we worden ermee geboren. Op dezelfde manier weten ook dat onze binnenwereld slechts een model van die buitentalige wereld is, en dat die twee strak aan elkaar gekoppeld moeten blijven, anders vliegen we mentaal uit de bocht. Daarom toetsen we zo'n vraag naar Fortuyn en Koekoek automatisch onmiddellijk aan wat we over de echte wereld weten, en besluiten dat Fortuyn Koekoek nooit beïnvloed kan hebben, maar andersom wel. Meestal gaat dat goed.

Virtuele atoombom

Tegenwoordig ligt een groot deel van onze bewuste kennis van de wereld en haar geschiedenis opgeslagen in het online-archief van miljarden teksten en tekstflarden dat de enige wereld en gegevensbron vormt die ChatGPT kent. En daar loert een enorm gevaar, want het maakt ChatGPT tot de natte droom van elke volksmenner en maatschappelijke onruststoker.

Zulke types hebben belang bij het veroorzaken van verwarring, wantrouwen en gevoelens van dreiging en naderend onheil. Dat doen ze enerzijds door het verspreiden van leugens, gruwelverhalen en andere desinformatie, en anderzijds door het in diskrediet brengen van betrouwbaar geachte kennisbronnen. Met ChatGPT krijgen zij een virtuele atoombom in handen, want iedere kwaadwillige kan, domweg door ChatGPT op industriële schaal willekeurige onzinnige vragen als

die over Koekoek te laten beantwoorden, ons mondiale kennisarchief in hoog tempo onherstelbaar vervuilen.

Het werkelijkheidsbesef aangetast

Die warboel kan ontstaan doordat alles wat ChatGPT produceert onmiddellijk van dat archief deel gaat uitmaken, bijvoorbeeld doordat het op een sociaal platform gepubliceerd wordt, of waar dan ook. Eenmaal op het web, krijg je die desinformatie er niet meer af, zoals menig sexting- en wraakpornoslachtoffer weet. En zo wordt al die rommel onderdeel van het bronnenmateriaal waarop ChatGPT zich baseert. En, wat erger is, wij ook.

Het gevolg is dat voor iedereen, van scholier, journalist en onderzoeker tot politicus en consument, het onderscheid vervaagt tussen kennis en kul, tussen werkelijkheidsgetrouwe informatie en leugenachtige, misleidende onzin. De verspreider van desinformatie hoeft niet meer te liegen of het gedrukt staat, de leugen stáát gedrukt. Kijk maar, hier, zwart op wit – en dus waar! Zo wordt ons hele werkelijkheidsbesef aangetast.

Kinderziektes

OpenAI, de maker van ChatGPT doet het voorkomen alsof het hier om kinderziektes gaat. Alsof het 'hallucineren', zoals zij het noemen, van het programma, op den duur wel verdwijnt, net als ongewenste stereotypen en discriminerende voorkeuren. Maar dat is niet zo, het gaat hier om fundamentele, onoplosbare gebreken. Dat zit zo.

Traditionele computerprogramma's worden met minimale, pijnlijk nauwkeurige stapjes geprogrammeerd, zodat we tot in het kleinste detail weten wat ze doen. Zulke programma's zijn als een breiwerk, dat je tot op de steek nauwkeurig kunt natellen en eventueel namaken, en waarin je elke fout kunt opsporen. Dat maakt ze absoluut voorspelbaar en honderd procent betrouwbaar, je weet precies hoe ze op elke vraag of opdracht reageren.

ChatGPT is anders. Het berust in de kern op een neuraal netwerk, dat getraind wordt zoals je een hond africht. Je geeft hond of netwerk een opdracht, en als daarop de gewenste reactie volgt, geeft de trainer een schouderklopje en volgt een

nieuwe taak. Zo niet, dan moet het werk over, net zo lang tot het wel lukt. Wat er in de hersens van de hond omgaat, weten we niet, en kan ons ook niet schelen.

Zo is het bij zo'n netwerk ook. Dat is in feite een enorme kluwen met elkaar verbonden schakelaars de het netwerk zelf aan en uit kan zetten. Levert een bepaalde configuratie iets op dat lijkt op het gevraagde resultaat, dan betekent ons schouderklopje 'houden zo', en kunnen we de resultaten door hogere eisen aan het antwoord te stellen verder gaan verfijnen.

Maar hoe die configuratie er precies uitziet, weten we niet. Dat maakt neurale netwerken veel flexibeler en veelzijdiger dan klassieke programma's, maar ook ondoorzichtig en onbetrouwbaar. Je weet nooit hóe een netwerk aan een antwoord komt, en ook niet of het op dezelfde vraag een volgende keer hetzelfde antwoord zal geven.

Soevereine kunstenaar

Soms weegt de flexibiliteit van netwerken tegen de ondoorzichtigheid en onvoorspelbaarheid ervan op. Dat is bijvoorbeeld zo in de kunstwereld, waar elke keuze en elke beslissing willekeurig is, uitsluitend afhankelijk van de soevereine wil van de kunstenaar. Daar zijn verrassing en onvoorspelbaarheid juist een pré.

Ook bij het experimenteren in kleine, strak gereguleerde omgevingen als schaken of Go kunnen neurale netwerken waardevolle originele wendingen en tactieken produceren. Zetten waar een mens niet gauw op zou komen. En microbiologen laten netwerken met succes zoeken naar nieuwe, nuttige manieren om bestaande eiwitten op te vouwen.

Maar dat zijn allemaal zuiver objectieve, waardevrije experimenten, die bovendien alleen gereedschap aanreiken waar menselijke kunstenaars, spelers en onderzoekers mee aan de slag kunnen – of niet, als het resultaat ze niet bevalt. Een bruikbare, betrouwbare tekst is iets heel anders. Die bestaat behalve uit woorden vooral uit een enorme kluwen van daarmee verbonden waardeoordelen. Oordelen in termen van goed en kwaad, mooi en lelijk, veilig en onveilig, en geschikt voor alle leeftijden of niet.

Zulke oordelen zijn subjectief en persoonlijk, terwijl ChatGPT net als alle neurale netwerken geen subject is en iedere persoonlijkheid ontbeert. Daardoor kan het

domweg nooit zelf achterhalen dat een afbeelding van een opengesperd vrouwelijk geslachtsdeel pornografisch is, maar niet als de maker Gustave Courbet heet en het ding in Parijs in het Musée d'Orsay hangt, en ook niet als hij in een medisch leerboek staat, maar volgens sommigen weer wel als hij voorkomt in het 'lentekriebels' materiaal voor de basisschool. Voor ChatGPT zijn alle katjes immers grauw.

Onwelvoeglijk taalgebruik

Omdat het programma geen toegang heeft tot de buitentalige werkelijkheid, kan het ook nooit ooit op eigen kracht ontdekken welke informatie klopt en welke niet, welke bronnen min of meer betrouwbaar geacht worden en welke niet, wie als schurk gezien wordt en wie als held, wat logisch is en wat niet, en welke argumenten deugen of juist drogredenen zijn. Of zelfs maar wat onder welke omstandigheden telt als onwelvoeglijk taalgebruik.

Er is dus maar één manier om ChatGPT van al die broodnodige waardeoordelen te voorzien. Ze moeten expliciet als ge- en verbodsregels in het trainingsprogramma worden opgenomen – voor een klein deel is dat ook al gebeurd, zo weigert ChatGPT bijvoorbeeld om over mensen te oordelen.

En daarmee zijn we terug bij af, en zitten we zelf al onze vooroordelen, voorkeuren en stereotypen in het programma in te bouwen. Reken maar dat dat nu al enthousiast gebeurt, met in Moskou heel andere resultaten dan in Nederland of in Tanzania. Op die manier draagt ChatGPT vooral bij tot de afbraak van ons werkelijkheidsbegrip. Daar zullen we, linksom of rechtsom, mee moeten leren leven.

Rik Smits is taalkundige en freelance wetenschapsjournalist. Behalve over taal, hersenen en digitalisering schrijft hij ook over zaken als linkshandigheid, geschiedenis en politiek. Zijn nieuwste boek is 'The Art of Verbal Warfare' (2022).