INTERVIEW MET AUTEUR EN DOCENT BAREND LAST OVER KUNSTMATIGE INTELLIGENTIE

‘DAT IS TOCH GAAF?’

AI-gegenereerde afbeelding met als zoekterm 'Chatten met Napoleon'

De eerste editie van *Chatten met Napoleon** verscheen in september. In het boek gingen docenten Barend Last en Thijmen Sprakel in op het gebruik van generatieve artificiële intelligentie (AI) in het onderwijs. Inmiddels ligt een herziene tweede editie in de winkel. Aanleiding voor *M&P* om met Barend Last te spreken.

IVO PERTIJS

‘In december kwam ik erachter dat ik zoveel nieuwe en relevante inzichten had… Ik sluit niet uit dat dit over een half-jaar weer het geval is’, vertelt Barend Last over het zo snel verschijnen van een nieuwe editie van *Chatten met Napoleon*. De ontwikkelingen gaan volgens hem onverminderd door. Last: ‘Soms zeggen mensen dat de hype voorbij is en de bubbel gaat barsten. Ik vraag mij dan af onder welke steen zij liggen. De AI-wereld gaat harder dan ooit. Elke dag verschijnen er nieuwe modellen en onderzoeken. Deze ontwikkeling is ongekend. We zitten nu wel in een tussenfase, denk ik. De nieuwe *tools* zijn absoluut indrukwekkend, maar nog verre van perfect. Maar toch, slechter dan dit wordt het niet.’

Wat was het afgelopen jaar de grootste verandering?

‘Het speelveld veranderde. ChatGPT was lange tijd dé *tool*. Zo zie je bijvoorbeeld dat de *tool* Claude sinds kort beter uit de benchmark komt dan ChatGPT. Google verscheen ook op het toneel. Daarnaast zijn er nieuwe inzichten voor het schrijven van *prompts*. We weten nu dat dat je slechtere antwoorden krijgt als je onbeleefd tegen AI bent. Dat vind ik absurd! Je moet dit weten om het maximale eruit te halen. Een andere update is dat we er eerst vanuit gingen dat iedereen vooral ChatGPT gebruikte, maar inmiddels zijn er veel spelers op de markt. Bovendien zijn veel van deze *tools* inmiddels multimodaal: ze kunnen niet alleen tekst genereren, maar ook beeld, geluid, documenten, en in sommige gevallen zelfs video’s. Daarom moesten we het boek wel herzien.’

Hoe belangrijk is kennis van de techniek achter generatieve AI voor docenten die hiermee aan de slag gaan?

‘Ik vind basale kennis voorwaardelijk als je de grenzen van deze *tools* goed en verantwoord wilt kunnen gebruiken. Zo weet je wat de beperkingen zijn en wanneer ze juist kansen bieden. Door mij te verdiepen in het proces van het stellen van een vraag en de totstandkoming van het uiteindelijke antwoord, merkte ik dat ik beter begreep hoe ik een goede vraag moest formuleren en het antwoord kon interpreteren. Je hoeft de wiskundige formules niet te kennen, maar met wat basiskennis kun je generatieve AI echt beter voor je laten

‘CHATGTP GENEREERT EEN VERHAAL OP BASIS VAN DATA, PATROONHERKENNING EN KANSBEREKENING OM MENSELIJK OVER TE KOMEN, MAAR HET BEKOMMERT ZICH NIET OF HET WAAR IS OF NIET.’

werken. Je begrijpt bijvoorbeeld waarom het systeem doet wat het doet, zoals dat het altijd een ander antwoord genereert, *biases* kan hebben of hallucineert. Zie het alsof je op een paard zit: als je de teugels vasthebt, maar niet goed weet hoe die teugels werken en hoe je op dat paard moet zitten, dan doet het paard waar het zin in heeft. Zo kom je nooit op de plek waar je wilt komen. Maar als je leert hoe je zo’n paard kunt beteugelen, dan doet het wat je vraagt en kun je het sturen. Zo is het ook met een generatieve AI-tool.’

Kun je een voorbeeld van een hallucinerend systeem geven?
‘Hallucineren is jargon voor een generatief AI-systeem dat om menselijk over te komen inhoud genereert die onjuist is. Ik gebruikte AI eens voor een pub-quiz. Ik vroeg welke van de tien paddenstoelen - de namen waren in het Latijn geschreven - eetbaar waren en welke niet. Ik kreeg een prachtig antwoord van ChatGTP, heel geloofwaardig, maar zeven van de tien antwoorden bleken fout. Het klonk wel plausibel en waar. Hallucineren beschrijft dus de situatie wanneer een taalmodel informatie verzint op basis van statistiek zonder dat dit voor ons “waar” is. Ik zet “waar” hier tussen haakjes, omdat waarheid ook subjectief is. Ik noem dan ook vaak de term *bullshit*, waarbij je iets zegt zonder er erg in te hebben of het waar is of niet. Donald Trump kan dat steengoed. ChatGTP doet eigenlijk precies hetzelfde: het genereert een verhaal op basis van data, patroonherkenning en kansberekening om menselijk over te komen, maar het bekommert zich niet of het waar is of niet. Dat kan ook helemaal niet, want daar is het simpelweg niet voor gemaakt.’

In je boek benadruk je onder meer het belang van demystificatie.
‘Ja, als je beter begrijpt hoe de techniek werkt, dan wordt AI stukken minder bedreigend. Media blazen AI weleens op tot mythische proporties. Natuurlijk, we kennen voortdurend menselijke eigenschappen toe aan iets dat niet menselijk is, maar het is in essentie niets meer dan een vernuftig rekenmodel dat werkt op basis van elektriciteit, algoritmen en computerchips. Als je AI ziet voor wat het is en het demystificeert, dan heb je minder angst voor het onbekende, want AI vervangt docenten niet, maar ze moeten zich wel degelijk aanpassen en dat lukt vooral als ze weten hoe de techniek werkt.’

‘ALS JE BETER BEGRIJPT HOE DE TECHNIEK WERKT, DAN WORDT AI STUKKEN MINDER BEDREIGEND.’

Wat is het effect als een docent aardig tegen het systeem is?
‘Ik vind het een paradox. Aan de ene kant zeggen we dat AI geen mens is en we het als een gereedschap moeten zien. Aan de andere kant moeten we er om er het maximale uit te halen zo menselijk mogelijk mee omgaan. Onderzoeken in experimentvorm tonen bijvoorbeeld aan dat als je het systeem vraagt om beter zijn best te doen, omdat dit belangrijk voor je carrière is, je beter resultaat krijgt. Onderzoekers kunnen nog niet precies verklaren hoe dit precies werkt, maar zo’n AI-model heeft zo ongeveer het hele internet aan tekst bestudeerd en daar zitten nu eenmaal menselijke kenmerken in die hun weg naar de output lijken te vinden. Het voelt dus soms gek en een beetje tegenstrijdig, maar het is geen gek idee om aardig tegen een robot te zijn. Lange tijd was AI voorbehouden aan een select clubje ICT’ers. Nu niet meer. De interactie met apparaten wordt steeds taliger, en dus ook menselijker.’

Waar moeten docenten maatschappijleer en hun leerlingen vooral op letten bij het invoeren van *prompts*?
‘Op de eerste plaats het privacy-aspect. We onderscheiden *tools* die werken in de *cloud* die je gratis of met een licentie gebruikt en taalmodellen die lokaal draaien. In lokaal draaiende taalmodellen kun je alles stoppen, maar die zijn niet zo goed als *cloud*-modellen. De gratis *cloud*-modellen zijn echter niet echt gratis, want jij bent het product. Alles wat je erin stopt kan worden verkocht of voor training van nieuwe modellen worden gebruikt. Bij een betaalde versie van ChatGPT of CoPilot kun je dit echter uitzetten. Toch zijn er docenten die denken dat alles wat zij erin zetten per definitie kwijt is. Dat klopt dus niet; het hangt af van welke *tool* je gebruikt en in welke hoedanigheid. Daarnaast is er onderzoek gedaan naar de politieke voorkeur van verschillende modellen. Een onderzoek toonde aan dat ChatGPT bij standaardantwoorden een linksliberale voorkeur heeft. Dat heeft te maken met de data waarop het is getraind, zoals het plukken van artikelen van bepaalde journalistieke platforms met een bepaalde politieke voorkeur. Als je naar een standaardantwoord vraagt, dan kun je verwachten dat het een bepaalde richting opgaat. Als je specifiek naar een rechts standpunt vraagt, dan krijg je dat natuurlijk wel - al zitten er in die systemen soms flinke beveiligingen die voorkomen dat er ongewenste teksten uit komen. Bij informatie over maatschappelijke vraagstukken is het dus heel belangrijk om goed naar de vraagstelling te kijken en de output kritisch onder de loep te nemen. Tot slot is er een risico op vooringenomenheid, de *biases*. Mensen zijn niet neutraal, taal is niet neutraal, cultuur is niet neutraal, dus dat vindt zijn weg terug naar de data. Je kunt de systemen niet volledig biasvrij maken, want daarvoor moet de maatschappij biasvrij zijn. Artificiële intelligentie fungeert zo ook als een spiegel van onze eigen waarden en normen. Als gebruiker heb je dus ontzettend veel verantwoordelijkheid om de juiste vraag te stellen en het antwoord

WERKVORM	VOORBEELDPROMPT
Ideeën voor opdrachten genereren	'Ik ben een docent in [CONTEXT] en verzorg onderwijs aan de volgende doelgroep: [DOELGROEP]. Ik wil voor de volgende leeruitkomsten ideeën uitwerken voor opdrachten: [LEERUITKOMSTEN]. Werk in bullets tien suggesties uit voor de volgende soorten opdrachten: [DISCUSSIE/BRAINSTORM/ ICEBREAKER/ETC.]. Licht per suggestie toe wat de voor- en nadelen zijn.'
Een opdracht in meer detail laten uitwerken	'Werk het idee voor een opdracht over [NAAM IDEE] uit in meer detail. Let daarbij op: [UITGANGSPUNTEN].'
De instructie voor een opdracht uitschrijven voor de leerlingen of studenten	'Schrijf de instructie voor [OPDRACHT] uit. Mijn doelgroep bestaat uit: [DOELGROEP). Houd rekening met de volgende uitgangspunten: [UITGANGSPUNTEN).'

Voorbeeld uit *Chatten met Napoleon* bij de werkvorm 'Bedenk activerende werkvormen'.

kritisch te evalueren. Een mooi voorbeeld is deze: als je aan vertaal*tool* DeepL vraagt om "*I am a business person*" te vertalen, dan vertaalt het systeem dit als "Ik ben een zakenman", maar een "*bossy business person*" wordt een "bazige zakenvrouw". Dat is toch bizar?! Bij een dialoog tussen een arts en een verpleger wordt de arts vaak een man en de verpleger een vrouw. Je moet van goeden huize komen om dat soort problemen eruit te vissen, en dus moeten we zowel docenten als leerlingen leren om hier bewust mee om te gaan.'

Zijn er ook wat rechtsere taalmodellen?
'Niet uitgesproken rechts. Aanvankelijk was ChatGPT getraind op het volledige internet, inclusief artikelen van bijvoorbeeld *NRC* en *The New York Times*. Dit leverde een discussie over het auteursrecht op, waarover het laatste woord - zelfs tot in de rechtszaal - nog niet is gezegd. OpenAI, het bedrijf achter ChatGPT, biedt nu de mogelijkheid om een bestandje op je website te plaatsen dat het gebruik van data voor AI-modellen inperkt. Veel nieuwsplatforms doen dit inmiddels, en laat dit nu net de meer linkse nieuwssites zijn. In volgende iteraties waarin het model wordt getraind is er dus een risico dat steeds minder links-liberale teksten worden gebruikt, en daarmee dus ook meer rechtse. Je begrijpt het risico. Toekomstige modellen geven wellicht rechtsere antwoorden. Maar enige geruststelling: het veld bouwt allerlei veiligheidsmaatregelen in en er zijn steeds meer nieuwe technieken om modellen te ontwikkelen. Bijvoorbeeld, als een taalmodel een antwoord produceert, dan controleert een tweede taalmodel dit op politieke richting en normen om zo voor neutralere antwoorden te zorgen.'

Het klinkt alsof het vak maatschappijleer een voorlopersrol kan spelen.
'100 procent. Bij het vak maatschappijleer ligt een grote kans om dit soort aspecten aan taal, cultuur en *biases* te koppelen. Alles komt samen. Het vak maatschappijleer kan bij uitstek een goede invalshoek bieden voor kritische denkvaardigheden.'

Het boek heet *Chatten met Napoleon*. Kun je ook met eigentijdse politici chatten?
'In essentie kan het, maar dit hangt van het taalmodel af. Sommige taalmodellen imiteren geen levend persoon. Als je een taalmodel lokaal draait, dan kun je dit zo voor elkaar boksen. De prompt die je daarvoor schrijft noem je een systeemprompt. Deze prompt is net iets anders dan een basisprompt. Als ik een vraag aan ChatGPT stel, dan zit daar een systeemprompt boven. In die prompt staat hoe ChatGPT zich moet gedragen. Als je Mark Rutte wilt nabootsen, dan moet je een systeemprompt schrijven waarin je duidelijk aangeeft dat de chatbot de rol van Mark Rutte aanneemt, dat het zich aan bepaalde richtlijnen moet houden en wat het wel of niet mag zeggen. Zo'n systeemprompt kan soms wel een of twee pagina's lang worden. Maar onthoud: leerlingen gaan dat systeem testen, dus daar moet je als docent slim mee omgaan. In mijn boek zeg ik daarom dat je eerst klein moet beginnen. Maak een Mark Rutte-bot en experimenteer ermee. In constante iteratie test je dit verder en pas je de prompt net zolang aan totdat je een prompt hebt die je veilig in de klas kunt inzetten. Dat is toch gaaf?' ◆

* De tweede druk van *Chatten met Napoleon. Werken met generatieve AI in het onderwijs* verscheen in april bij Boom Hoger Onderwijs.