REEKS ZIENERS (1)

‘Ik heb de gevaren van AI niet ernstig genoeg genomen’

Hij is een van de peetvaders van artificiële intelligentie, maar vandaag maakt Yoshua Bengio zich grote zorgen over zijn uitvinding. Over tien jaar kunnen de systemen al slimmer zijn dan wij, gelooft hij. ‘Ik heb altijd gedacht dat we de menselijke intelligentie moesten trachten te imiteren. Nu denk ik dat dat een grote vergissing zou zijn.’

Dominique Deckmyn

Zaterdag 1 juli 2023 om 03 00

Bezorgd is Yoshua Bengio al langer. In 2017 al was hij een van de initiatiefnemers van de *Declaratie van Montreal*, die pleitte voor meer ethische artificiële intelligentie. Bengio is samen met Geoffrey Hinton en Yann LeCun een van de grondleggers van AI: in 2019 kreeg het drietal de Turingprijs, zowat de Nobelprijs voor computerwetenschap, voor hun rol in de ontwikkeling van *deep learning*, de tak van AI die sinds 2010 tot nooit geziene doorbraken leidde. ‘Zo’n vijf jaar geleden begonnen grote bedrijven als Facebook en Google elkaar te beconcurreren bij het opkopen van AI-start-ups’,

vertelt hij. ‘Het werd toen duidelijk dat we van spoor moesten veranderen. Dit was niet langer een puur intellectuele verkenning voor de wetenschap, we moesten echt nadenken over hoe de technologie gebruikt zou worden.’

Maar de stroomversnelling waarin de technologie dit jaar is terechtgekomen, maakt Bengio ronduit bang. *De Standaard* praatte via videoverbinding met hem over de verbijsterende ontwikkelingen van de afgelopen maanden en zijn groeiende bezorgdheid.

Wat is uw grootste zorg?
‘De onzekerheid en de omvang van de schade die aangericht kan worden als we machines bouwen die slimmer zijn dan wijzelf. De technologie kan ook misbruikt worden of op een onverstandige manier gebruikt. Beide scenario’s kunnen catastrofale gevolgen hebben.’

Wanneer kreeg u het gevoel dat het allemaal sneller gaat dan u had verwacht?
‘Dat is geleidelijk gebeurd, de voorbije paar maanden, naarmate ik meer vertrouwd raakte met ChatGPT en later GPT-4 en daarover nadacht en praatte met collega’s van over de hele wereld.’

Taalmodellen
Nochtans zijn taalmodellen als GPT-3.5 (het AI-systeem dat ChatGPT aandrijft) en opvolger GPT-4 niet nieuw voor Bengio. Hij was zelf de grondlegger van de technologie, toen hij 20 jaar geleden begon neurale netwerken (computerprogramma’s die geïnspireerd zijn op de werking van hersencellen) te trainen door ze te voeden met tekst. ‘We begonnen met zeer kleine datasets en zeer kleine neurale netwerken, maar de resultaten waren toch al interessant.’ In 2014 werd een nieuw element toegevoegd: een attentiemechanisme, dat maakt dat het taalmodel beter rekening kan houden met context. ‘Toen we dat deden aan de universiteit, werkte het gewoon iets beter dan de bestaande systemen, maar echt indrukwekkend was het niet. Maar Google bouwde een groter neuraal netwerk, trainde dat op meer data en liet er goede engineering op los. Het resultaat was verbluffend. Toen begonnen we het effect van schaalvergroting te zien: we hoeven het recept van die neurale netwerken eigenlijk niet zo veel aan te passen, we maken ze gewoon groter, trainen ze langer en op meer data, en ze worden verbazingwekkend veel beter.’

En dat heeft u dit jaar opnieuw verbaasd?
‘Ja.’

Wat was het dan precies, dat u niet had zien aankomen?
‘Ik dacht dat de systemen, zelfs met zoveel trainingsdata, zouden falen als ze geconfronteerd werden met nieuwe vragen. Maar dat blijkt niet zo te zijn. Je kunt de systemen nog altijd in de fout laten gaan, maar dan moet je daar al echt moeite voor doen. Veel mensen zouden diezelfde vragen ook niet correct beantwoorden. Ik realiseerde me dus dat we een paar echt belangrijke mijlpalen hebben bereikt. We hebben de Turingtest gehaald, het moment waarop je mens en machine niet meer uit elkaar kunt houden als je ermee interageert door tekst in te tikken.’

U denkt dat systemen als ChatGPT voor die test zouden slagen?
‘Officieel is dat niet gebeurd, maar in de praktijk is dat toch wat we zien gebeuren.’

De jongste maanden groeit de bezorgdheid over de ontwikkelingen. Wat loopt er momenteel mis?
‘Wel, er is net iets goeds aan het gebeuren: meer mensen worden zich bewust van de gevaren die misschien in het verschiet liggen. Eerst waren dat de mensen die de technologie ontwierpen en de bedrijven errond oprichtten. Maar ook politici zijn zich er mee gaan bemoeien en dat is heel belangrijk.’

‘Ik ben ervan overtuigd dat er manieren zijn om veilige AI-systemen te bouwen – systemen waarover we de controle niet kunnen verliezen en die de mensheid echt kunnen helpen. Het probleem is: als we weten hoe we veilige AI-systemen moeten bouwen, dan weten we ook hoe we de gevaarlijke kunnen bouwen – dus: hoe verhinderen we dat dat gebeurt? En dat is een politieke vraag, geen technische.’

Europa werkt aan een AI-Act. Zitten we op het juiste spoor?
‘Ik vrees dat het om een erg log instrument gaat, terwijl alles nu heel snel evolueert. We hebben een regelgevend orgaan nodig dat zijn regels snel kan updaten als er nieuwe potentiële gevaren opduiken. Er is ook een interessant Canadees wetsvoorstel, dat de rol van de wet en van het regelgevend orgaan opsplitst op een flexibele manier. Maar deze wetten zijn ontworpen vóór ChatGPT en vóór de huidige discussie over het verlies van controle en de bedreiging van de democratie. Er moeten meer controles op die potentiële gevaren ingebouwd worden. Ik meen te begrijpen dat het comité over AI-richtlijnen van de G7 zal nadenken over alle gevaren en dat is goed.’

Verschillende kopstukken van de AI-industrie, zoals OpenAI-baas Sam Altman, verklaren openlijk dat zij grote gevaren zien en dringen aan op regulering. Hoe oprecht is Altman daarin?
‘Ik ken hem niet goed genoeg om daarop te antwoorden. Hij ziet er oprecht uit, maar anderzijds: verklaren dat er regulering nodig is en dan zeggen dat de Europese wetgeving onmogelijk gerespecteerd kan worden, dat klinkt niet zo coherent.’

OpenAI heeft de technologie achter ChatGPT niet uitgevonden, maar wel als eerste op de markt gebracht. Was dat onverantwoord?
‘We weten eigenlijk niet precies wat er in hun systeem zit. Maar we hebben aanwijzingen dat ze verschillende bestaande elementen hebben samengebracht – zoals wel meer gebeurt. En dat hebben ze opmerkelijk goed gedaan. Ik denk dat ze zeker gestart zijn met een sterke intentie om de risico’s aan te pakken, inbegrepen het risico op verlies van controle. Ik lees dat ze die zorg misschien wat laten verslappen door de commerciële druk en de race waarin ze zich nu bevinden, maar dat kan ik niet met zekerheid zeggen.’

Yann LeCun, die andere peetvader van AI, vindt het zinloos om ons nu al zorgen te maken over superintelligentie, omdat we daar nog mijlenver vanaf staan.
‘Dat is niet wat hij zegt, denk ik. Ik denk dat hij het ermee eens is dat we op een spoor zitten dat zal leiden tot machines die slimmer zijn dan wij. En niet pas over een eeuw. Misschien ligt die horizon voor hem wat verderaf dan voor mij, maar dat is geen fundamenteel meningsverschil. Zoals ik het begrijp, vindt LeCun dat we ons geen zorgen hoeven te maken omdat we het probleem wel zullen kunnen oplossen als het zich aandient.’

En u denkt van niet?
‘Ik denk dat de inzet veel te groot is om niet zo voorzichtig mogelijk te zijn.’

U tekende in maart de open brief die opriep tot een pauze van zes maanden in de ontwikkeling van de meest geavanceerde AI-systemen. Zou dat helpen?
‘Ja, maar het zal niet gebeuren. Nog voor de brief gepubliceerd werd, wist ik dat het nooit zou gebeuren. Maar de brief had een positieve impact: het bracht de discussie op gang. Dat is een indrukwekkende verwezenlijking.’

Dus wat moet er nu gebeuren?
‘Geoff Hinton zei enkele weken geleden: we moeten evenveel uitgeven aan het verbeteren van de algoritmes als aan het maken dat ze veilig zijn en dat we het publiek beschermen. Dat kunnen AI-researchers niet in hun eentje, we moeten werken met juristen, met mensen die de ethische aspecten begrijpen, specialisten in cybersecurity en in nucleaire, biologische en chemische wapens.’

De grote AI-systemen die nu zo snel evolueren, zoals GPT-4, kunnen die veilig gemaakt worden?
‘Het probleem zit niet bij de architectuur, al kunnen we die altijd verbeteren. Vooral de manier waarop de systemen getraind worden, moet veranderen. En daar heb ik enkele ideeën over. Eén manier om de systemen veilig te maken, is garanderen dat ze geen *agency(het vermogen om initiatief te nemen, red.)* hebben, dat ze zelfs geen notie hebben van een eigen doel of plan. Al wat ze doen, is proberen te begrijpen hoe de wereld werkt en ons met die kennis helpen om vragen op te lossen.’

Systemen als GPT-4 leken geen agency te hebben, ze beantwoorden alleen vragen. Maar plots komen er systemen uit als AutoGPT (https://www.standaard.be/cnt/dmf20230421_96062149), dat ChatGPT omtovert in een soort autonoom optredende assistent. Is dat wat u agency noemt?
‘Ja. Het is erg eenvoudig om iets te nemen wat dienst doet als een soort orakel dat alleen vragen beantwoordt, en het om te toveren in een *agent*. De *agent* hoeft immers alleen te weten: van alle mogelijke acties die ik kan ondernemen, welke actie zal mij helpen om mijn doel te bereiken? En als je het antwoord hebt op die vraag, dan onderneem je die actie.’

Dus AutoGPT is gevaarlijk?
‘Niet zoals het nu is. Er is nog niet veel werk in gestopt om het goed te maken. En ChatGPT is nog niet zo slim. Maar stel dat we in kwaliteit nog een stapje boven GPT-4 gaan, laten we zeggen GPT-5 of zoiets, dan kan het gevaarlijk worden.’

Superintelligentie

Bengio betreurt dan ook dat hij niet eerder in actie is geschoten. ‘Ik denk dat er een psychologisch obstakel is dat ons verhindert om iets te zien wat zo in tegenstelling is met het doel waar we naartoe werken. En daarom heb ik niet gedaan wat ik had moeten doen. Deze gevaren zijn al 10 à 20 jaar bekend, maar ze werden niet ernstig genomen omdat mensen als ik dachten: ach, dat ligt nog veel te ver in de toekomst, we hebben zelfs nog geen idee van hoe zulke AI-systemen eruit zullen zien, dus hoe kunnen we daar vandaag al iets aan doen? Maar nu is de situatie anders, nu staat het vlak voor ons.’

Er zijn nog heel wat technische obstakels met de huidige taalmodellen, maar Bengio gelooft dat die in sneltempo opgelost kunnen worden. Dat systemen als ChatGPT om de haverklap feiten verzinnen, de zogenoemde 'hallucinaties', bijvoorbeeld: 'Dat kan ingeperkt worden, daar werken momenteel veel mensen aan. En zelfs in hun huidige staat, met al hun gebreken, kunnen die systemen nuttig zijn. Alle mensen die ik om me heen zie, gebruiken ze.'

'We begrijpen in grote lijnen wat er aan die systemen ontbreekt. Er wordt gewerkt aan algoritmes die dat kunnen verbeteren. Misschien kunnen we alle problemen in drie tot vijf jaar oplossen, en dat is wat mij zorgen baart. Nu, het blijft mogelijk dat er nog iets anders, iets fundamenteels, ontbreekt wat we nu nog niet zien. Dat zou dan eigenlijk een goede zaak zijn, want het zou ons meer tijd geven om ons aan te passen. Maar wat als het juist sneller gaat dan verwacht, zoals het afgelopen jaar het geval was?'

U gelooft nu dat superintelligentie, een AI-systeem dat de mens overtreft, er over 10 jaar al kan zijn?
'Ja, net als veel van mijn collega's.'

Nochtans zeggen andere onderzoekers dat er nog te veel onopgeloste problemen zijn, dat het veel langer zal duren.
'Absoluut. En had je mij die vraag een jaar geleden gesteld, dan had ik 20 tot 100 jaar gezegd. Nu zeg ik: 5 tot 20 jaar.'

Dus hoe kijkt u nu naar uw eigen onderzoek? Vindt u het gevaarlijk?
'Mijn onderzoek heeft zeker geholpen om ons te brengen waar we nu staan. Al die vragen stel ik mezelf elke dag: wat is het beste gebruik van mijn tijd om de risico's te minimaliseren en de voordelen voor de hele mensheid te maximaliseren?'

Het klinkt alsof u daar echt mee worstelt.
'Dat is ook zo. Hier zijn geen gemakkelijke antwoorden op te vinden.'

DS

Uw onderzoek zou het mogelijk maken dat de AI-systemen meer denken zoals wij – maar dat vindt u dus nu eigenlijk geen goed idee meer?

‘Gedurende mijn hele carrière heb ik gedacht dat we inspiratie moesten halen uit de mens, moesten proberen om onze eigen intelligentie te imiteren. Nu denk ik dat dat een vergissing is. Het is belangrijk dat we begrijpen hoe het brein werkt, maar machines bouwen die erg lijken op mensen zou in veel opzichten een grote vergissing zijn. In sciencefiction zien we androïden en AI-systemen die min of meer als mensen zijn, maar dan mechanisch. Zoals Data in *Star Trek*. Dat beeld klopt volgens mij niet. Om te beginnen zijn machines in wezen onsterfelijk, want ze kunnen hun programmacode en hun staat kopiëren. Wij zijn fundamenteel anders. We ontwerpen die machines omdat we instrumenten willen bouwen die ons helpen. Maar misschien scheppen we een nieuwe soort die gevaarlijk is voor ons en zelfs voor het voortbestaan van onze eigen soort. In het verleden zijn vele soorten uitgestorven, gewoonlijk omdat er een slimmere soort kwam. In de laatste paar eeuwen hebben we het uitsterven van zo’n duizend soorten veroorzaakt. Niet omdat we ze dood wilden, maar omdat ze in de weg stonden tussen ons en meer land of meer geld.’

Dus u gelooft op dit moment dat het beter is dat we die superintelligentie nooit bereiken?

‘Dat zeg ik niet. Als we die superintelligentie op een veilige manier bereiken, kan het ons helpen met veel van de uitdagingen waarvoor de mensheid staat: ziekte, klimaatverandering, armoede en economische ongelijkheid – problemen waar we geen of amper vooruitgang in boeken. Maar we moeten het op een veilige manier doen. En op dit moment weten we eigenlijk niet hoe dat moet. En dus moeten we het langzaamaan doen en werken aan maatregelen die de schade kunnen beperken.’

‘Het andere probleem is: zelfs als ik, of iemand anders in een paar landen, beslist om te vertragen, dan zijn er toch mensen in andere landen die dat niet doen. Wat doen we daar dan tegen? Een van de dingen die mijn vriend en collega Yann LeCun zegt, en waar ik het mee eens ben, is dat we waarschijnlijk de hulp van sommige AI-systemen nodig zullen hebben om ons te beschermen tegen mogelijke losgeslagen AI’s. Dus daar staan we dan: we hebben die systemen nodig, maar tegelijk willen we ze niet. En als ze toch ergens opduiken, dan hebben wij er ook nodig om ons ertegen te beschermen. Het is zoals nucleaire wapens, nietwaar? Kernwapens zijn gevaarlijk, maar als iemand anders ze gaat bouwen, dan hebben wij ze ook nodig. We moeten daarmee leren om te gaan.’

Moeten we ons zorgen maken over de toekomst?

‘Wel, je zorgen maken is op zichzelf niet nuttig. We moeten handelen. In de eerste plaats om die kwesties beter te begrijpen, want er is veel wat we nog niet begrijpen. Een goede aanwijzing daarvoor is dat er zoveel onenigheid bestaat tussen onderzoekers. Daarnaast moeten we in regulering voorzien, in internationale akkoorden, om op een pad te raken naar veiligheid en fairness, met de bedoeling het publiek beter te beschermen.’

Hebben overheden nog tijd om op te treden? Er bestaan al open-sourceversies die iedereen zomaar kan downloaden en verder bewerken. Dat kun je toch niet meer stoppen?

‘Nee, de huidige systemen zijn nog te dom om echt gevaarlijk te zijn. Ja, ze kunnen gebruikt worden voor desinformatie. Maar de echt grote gevaren liggen nog een aantal jaren voor ons, dus het is nog niet te laat. Maar we zullen wel veel sneller wetgevende maatregelen moeten opstellen dan we, bijvoorbeeld, gedaan hebben tegen de klimaatverandering. Daarin waren we veel te traag.’

Zieners

Hoe kunnen we de radicale veranderingen begrijpen die onze wereld ondergaat? In de reeks ‘Zieners’ zoeken we houvast en inspiratie bij vooruitziende denkers.

In deze aflevering is dat **Yoshua Bengio**, een van de grondleggers van Artificial Intelligence. Bengio is professor computerwetenschappen aan de Universiteit van Montreal en wetenschappelijk directeur van het Montreal Institute for Learning Algorithms. In 2019 won hij de Turing-prijs voor zijn bijdrage aan de ontwikkeling van AI, samen met Yann Lecun Geoffrey Hinton. Vorig jaar was Bengio de meest geciteerde wetenschapper ter wereld.

Volgende week: wetenschapsfilosofe Vinciane Despret.

Bron artikel: De Standaard, 1 juli 2023, https://www.standaard.be/cnt/dmf20230630_96350956.