Grootste risico's van AI liggen in de zachte sectoren: zorg, onderwijs, defensie

Posted on 17 oktober 2024 by Rik Smits

Vorige week werd de Nobelprijs voor natuurkunde 2024 toegekend aan de Amerikaan John Hopfield en de Brits-Canadese Geoffrey Hinton, voor hun baanbrekende werk op het gebied van kunstmatige intelligentie. Tegelijk waarschuwt Hinton op dramatische toon voor de gevaren van AI, tot het einde van onze soort toe. Maar wat zijn die gevaren dan, en is het echt zo erg?

Met kunstmatige intelligentie, inmiddels ook hier beter bekend als AI, het acroniem van het Engelse Artificial Intelligence, bedoelen we tegenwoordig vooral computersystemen die op basis van een neuraal netwerk kunnen leren door ervaring, en zo 'vanzelf' beter worden in wat ze moeten doen. Het idee erachter dateert al van rond 1950: boots de beste leermachine na die we kennen, ons eigen brein.

Een zelflerende computer

In beginsel zijn onze hersenen een kluwen van miljarden met elkaar verbonden neuronen. Onder invloed van elektrochemische signalen vanuit andere delen van die kluwen, het zenuwstelsel en de rest van ons lichaam ontwikkelt het brein zich door nieuwe verbindingen te leggen of bestaande juist op te ruimen, maar vooral door verbindingen meer of minder doorlaatbaar te maken voor signalen, alsof je kraantjes verder open- of dichtdraait. Zo ontstaan in de hersenen min of meer stabiele patronen die we 'kennis', 'woorden' of 'herinneringen' noemen – dat is, op het meest basale niveau, natuurlijke intelligentie.

Zo'n neuraal netwerk zou je ook best kunnen bouwen van elektronische onderdelen, en dan had je dus een zelflerende computer – kunstmatige intelligentie. Goed idee, maar er was nog heel wat technisch en theoretisch kunst- en vliegwerk voor nodig om tot een concreet werkend apparaat te geraken, terwijl het nog ruim een halve eeuw zou duren voordat er voldoende rekenkracht beschikbaar kwam voor serieuze toepassingen, en voldoende gedigitaliseerde data om zo'n apparaat te trainen. Al die tijd bleef het bij bescheiden prototypen die zich bewogen binnen piepkleine, strak gereguleerde en superoverzichtelijke wereldjes als schaken en het spel go.

Veel van dat voorbereidende werk werd gedaan door de nu gelauwerde Hopfield en Hinton, waarbij vooral Hinton ging beseffen dat we misschien wel te maken hebben met de bezem uit 'De tovenaarsleerling' van componist Paul Dukas. Die tovenaarsleerling moet het atelier van zijn baas aanvegen, maar hij is lui en zodra de tovenaar van huis is, bezielt hij met een toverspreuk de bezem. En ja hoor, het ding gaat braaf aan het werk. Maar het wordt steeds enthousiaster en blijkt al gauw niet meer te stoppen, het veegt alles aan diggelen. Wat als op neurale netwerken gebaseerde toestellen en installaties ook aan onze controle zouden ontsnappen? Wat als ze voor zichzelf zouden beginnen, hun eigen zin gaan doen, slimmer worden dan wij?

Daarover werd ook al lang geleden nagedacht door Isaac Asimov, die de drie wetten van de robotica (lees: AI) formuleerde:

1. Een robot mag een mens niet schaden, ook niet door nalatigheid;
2. Een robot dient opdrachten van een mens te gehoorzamen, tenzij hij daardoor de eerste wet overtreedt;
3. Een robot moet zijn eigen voortbestaan bevorderen voor zover hij daardoor

niet de eerste of tweede wet overtreedt.

LAWS

Zo lang dat allemaal sciencefiction bleef, was er weinig aan de hand. Maar inmiddels zijn dankzij AI robots met verregaande, soms levensbedreigende vermogens in rap tempo werkelijkheid aan het worden. Vandaar dat Hinton al sinds een jaar of zeven ageert voor een wereldwijd verbod op zogeheten LAWS, 'lethal autonomous weapons systems', oftewel dodelijke autonome wapensystemen, die op eigen houtje doelen zoeken en aanvallen.

Zo'n verbod heeft weinig kans van slagen en nog minder zin. Kijk maar hoe Oekraïne op dit moment als proeftuin fungeert voor steeds dodelijker en steeds zelfstandiger opererende oorlogsdrones. Ook zijn er heel wat ontwikkelaars die lak hebben aan Asimovs edelmoedige wetten. Bovendien gaat het tovenaarsleerlingprobleem over veel meer dan wapensystemen alleen. Het maakt immers niet uit of je omkomt door een eigenwijze LAWS of door een losgeslagen robotauto.

Hoe bedreigend ook, toch zijn dit nog de gemakkelijkst op te lossen problemen. Het gaat om concrete apparaten, fysieke techniek die je met techniek kunt bestrijden. Een killer-drone kan ook afgericht worden voor de jacht op vijandelijke drones en andere oorlogsrobots. Veel lastiger, want ongrijpbaarder, is AI die voor surveillance en manipulatie van burgers wordt ingezet. De meerwaarde van neurale netwerken is dat ze uitblinken in het werken met onafzienbare hoeveelheden gegevens, iets waar wij mensen helemaal niet goed in zijn. Maar de keerzijde is dat repressieve overheden en naargeestige werkgevers een ongekende grip op onschuldige burgers kunnen krijgen, erger dan in Orwells *1984*.

Een eerste voorbeeld is het sociale-kredietsysteem dat China heeft opgetuigd, met behulp van onder meer gezichtsherkenning. Loop bij wijze van spreken drie keer door een rood voetgangerslicht, en je krijgt geen hypotheek meer. Daartegen helpt zelfs geen autonome killer-robot.

Bizarre onzin

Maar het geniepigst van alles is de fundamentele onbetrouwbaarheid van AI. Het

punt is dat we bij traditionele, algoritmische computerprogramma's heel precies weten wat de computer doet. Gaat er iets mis, dan kunnen we de fout dus in principe altijd opsporen en corrigeren. Bij neurale netwerken kan dat niet. Een neuraal netwerk is een hermetisch gesloten zwarte doos waar je een vraag instopt, waarna er een antwoord uitkomt, maar zonder dat je enig idee hebt hoe het apparaat daarbij komt. Je hebt daarom ook geen enkele garantie dat AI op dezelfde vraag een volgende keer hetzelfde antwoord zal geven.

Wie weleens met ChatGpt of een soortgelijk systeem gespeeld heeft, weet wat voor bizarre onzin daar ineens uit kan komen. Dat is als we ons ermee amuseren niet erg, maar wel als het bijvoorbeeld om politieke of historische kennis gaat. Of om medische diagnostiek, om het nakijken van schoolwerk of het op afstand beoordelen van de emotionele staat en intenties van mensen (ja, dat gebeurt echt, nu al).

Gaat het om toepassingen op een gebied dat behoorlijk strak en precies omschreven is, zoals wis- of natuurkunde, dan zullen experts serieuze rariteiten nog wel opmerken. Maar hoe 'zachter' het werkterrein, hoe meer er afhangt van menselijke interpretatie. En hoe belangrijker de rol van interpretatie, hoe groter de kans dat AI gevoed wordt met al dan niet opzettelijke nepkennis. Hoe kleiner bovendien de kans dat fouten en onzin worden gezien. En daar is ten principale niets aan te doen. In die zachte sectoren, zoals de zorg, het sociaal werk en het onderwijs, maar ook defensie en de beleggingswereld, liggen de allergrootste risico's.

Leren als een peuter

Hintons vrees dat AI onze intelligentie zal overtreffen en zich dan tegen ons keert om ons te onderwerpen of uit te roeien, komt voort uit zijn overtuiging dat neurale netwerken 'leren als een peuter'. Maar daar klopt gelukkig weinig van. Neurale netwerken leren door patronen te vinden in onmetelijke hoeveelheden gegevens. Daarbij bepaalt de kwaliteit van die gegevens rechtstreeks de kwaliteit van wat het netwerk oplevert.

Maar, zoals de grote linguist Noam Chomsky al rond 1955 ontdekte: zo 'werken' taallerende peuters helemaal niet. Die verwerven de grammatica van hun moedertaal allemaal vrijwel perfect op basis van niet meer dan wat ze in een paar jaar om zich heen horen. Dat is relatief heel weinig, en dat weinige is dan ook nog

ernstig vervuild door allerlei uitspraakfouten, valse starts, onderbrekingen en zinnen die halverwege ineens een andere kant uitgaan.

Geen opstand der digitale horden

Hinton lijkt de menselijke geest te bezien door de bril van een naïeve behavioristische psycholoog uit 1950: als een *tabula rasa*, een onbeschreven blad, dat uitsluitend door de invoer van gegevens naar believen gevormd kan worden. Maar het bestaan en de onmisbaarheid van allerlei aangeboren zaken als driften, behoeften, en de basisvoorwaarden voor taal, motoriek, procedureel denken en duizend andere dingen valt echt niet te ontkennen. Dus al zou AI in enig opzicht intelligenter worden dan wij, dan nog hoeven we voor een opstand der digitale horden niet te vrezen. Daar zijn wil, motivatie, begeerte en gevoelens van macht voor nodig, allemaal dingen die een neuraal netwerk vreemd zijn.

Rik Smits *is taalkundige en wetenschapsjournalist.*

Wynia's Week verschijnt nu drie keer per week! De groei en bloei van Wynia's Week is te danken aan de donateurs. **Doet u al mee? Doneren kan op verschillende manieren. Kijk *HIER*. Hartelijk dank!**