Responsible AI: wat maakt AI belangrijk in onze samenleving?

Kunstmatige of artificiële Intelligentie (AI) is vandaag de dag niet meer weg te denken, omdat het een relevant onderwerp voor de hele samenleving is geworden. Bart Verheij is al geruime tijd bezig met AI, en signaleert: 'De aard van de kunstmatige Intelligentie of AI is door de jaren heen erg veranderd.' Hoe ziet hij de relatie met verantwoordelijkheid?

Prof. dr. Bart Verheij is hoofd van de afdeling Kunstmatige Intelligentie aan de Rijksuniversiteit Groningen. Hij werkt aan de theoretische, computationele en empirische verbanden tussen kennis, data en redeneren, als bijdrage aan verantwoorde kunstmatige intelligentie (AI). Hij doet dit vanuit een argumentatie-perspectief, geïnspireerd door de domeinen van recht en bewijs.

Verheij studeerde wiskunde vanaf 1985 op de universiteit van Amsterdam. In die tijd was AI verbonden met 'logica', wat in feite gaat over het correct redeneren. "Waar het destijds om ging was de formele geldigheid van een intelligent denkproces. Je moest proberen vast te leggen wat de regels waren van rationaliteit. Wiskundige logica was de belangrijkste theoretische basis van AI en ook van de informatica. De werkwijze van computers komt voort uit deze logica."

Vrijwel alles was gebaseerd op logica en werd daardoor beschouwd als verantwoord en betrouwbaar, aldus Verheij. "Als een modellen incorrect bleek te zijn, had dat altijd een aanwijsbare oorzaak. Je moest soms wel goed zoeken, maar het kon, want elke stap was expliciet en had betekenis."

Verheij refereert aan John McCarthy, de bedenker van de term 'artificial intelligence' in de jaren 1950. McCarthy was heel invloedrijk in het uitdragen van het perspectief dat logica de basis van AI is en moet zijn. Deze voorgeschiedenis speelde ruimschoots vóór de datarevolutie. "Wat we nu zien in de AI is heel anders dan logica, en wel een voorzetting van technieken die er al voor de datarevolutie waren. Vooral omdat er nu veel meer data en rekenkracht beschikbaar zijn, kan er veel meer worden gedaan, zoals machine learning. Wat een interessant gegeven is, is dat de achterliggende techniek aanvankelijk niet zo populair was, zelfs verguisd werd door de logica-georiënteerde AI-professionals."

Verheij stelt vast dat door de komst van het internet, ontwikkelingen een enorme vlucht hebben genomen. AI-toepassingen zijn voor een groot deel een datagedreven context geworden en dit heeft een multiplierreffect wat tot steeds meer AI leidt. "We leven nu in een tijd met veel meer data en computers met veel rekenkracht. Dit maakt de AI-technieken populair. Het is een voortzetting van technieken en de mantra van John McCarthy dat alles op logica gebaseerd moet zijn, is langzaam en stilletjes vervallen."

Verheij zegt: "We verloren daarmee wel de 'responsibility' en de 'explainability'. Voorheen kon men in het model herleiden waar een bepaalde conclusie op was gebaseerd. Dat is nu verdwenen want een getraind neuraal netwerk (een belangrijke techniek in de datagedreven AI) wordt gezien als een 'black box'. We kunnen niet meer vanuit de binnenkant goed kijken, welke stappen er naar de conclusie zijn gezet. Het kan wel, maar niet op een manier die veel betekenis heeft voor mensen. Technisch gezien kun je van binnenuit kijken naar de technische geheimen maar de menselijke interpreteerbaarheid van wat er gebeurt in een getraind neuraal netwerk, is er niet. Daar wordt aan gewerkt. Dat heet interpretability als probleem en explainability als probleem."

Wat betekent 'Responsible AI' in uw optiek?
Verheij verwijst hierbij naar zijn collega Cor Steging die een proefschrift over het responsibility probleem met data heeft geschreven. "De kern van wat hij aangeeft is dat niet alleen de conclusie relevant is maar óók de redenering ernaartoe is belangrijk. De onderbouwing van een conclusie is nu eenmaal onontbeerlijk in sommige domeinen. In het juridische domein bijvoorbeeld geldt dat het stappenplan erg relevant is. De jurist wil kunnen onderbouwen waarom een conclusie wordt getrokken en die moet ook correct zijn naar de juridische maatstaven. Dit gaat niet vanzelf met moderne data gedreven AI, machine learning."

Verheij benadrukt dat de menselijke betrokkenheid nodig blijft om voldoende betrouwbaarheid te bereiken. "De garantie van de logica kun je nu niet bereiken maar je wilt zulke correctheid wel benaderen. Je wilt zoveel mogelijk correct zijn en een deelbare onderbouwing van standpunten hebben. Dat is 'responsible'. Het gaat om de correcte redenering en daarnaast ook om 'responsible zijn' in de ethische zin. Dit laatste is een diepere laag waarbij geldt dat het vaak niet vast staat wat de correcte ethische randvoorwaarden zijn. Daar is altijd discussie over mogelijk."

DATAGEDREVEN AI IS KANSLOOS OM EEN SERIEUZE PARTNER TE ZIJN BIJ EEN ETHISCH GESPREK

Hij vervolgt: "Er is eigenlijk maar één fysisch beschikbaar systeem dat ethische randvoorwaarden kan opstellen en dat zijn wijzelf: de mens. Het zijn uiteindelijk de mensen die de ethische randvoorwaarden opstellen. De mensen zijn als enigen in staat om hierover te overleggen om tot een verstandig perspectief te kunnen komen, en soms worden we het zelfs in moeilijke situaties eens. Er is geen andere bron van ethische randvoorwaarden dan wij. Hier kan AI niet voor zorgen. Datagedreven AI is kansloos om een serieuze partner te zijn bij een ethisch gesprek. De middelen om ethisch te trainen bestaan gewoonweg nu niet."

Waar staat Nederland als het gaat om responsible AI?
"Nederland heeft een speciale positie als het gaat om 'responsible AI' maar dat idee is wel een beetje gekleurd doordat ik een Nederlandse onderzoeker ben." Nederland doet academisch gezien serieus mee in het nadenken over 'responsible AI'. Verheij verwijst in dit verband naar een interessant landelijk project, het Hybrid Intelligence project. Dat is een groot, tienjarig landelijk project waar 'responsible AI' één van de onderwerpen is. Het is nu ongeveer halverwege.

Het idee van het project is dat meerwaarde te bereiken is wanneer er samenwerking is met mensen van verschillende disciplines. Dat is niet zo makkelijk. Bij het ontwerp van AI-systemen is het betrekken van mensen noodzakelijk. Als je mensen en machines samenbrengt kun je iets bijzonders krijgen, namelijk meer dan de som van de delen, het wonder van hybride intelligentie." Nederland doet volgens Verheij zijn best om 'responsible' te werken. "Er wordt nadrukkelijk onderzocht om de human alignment te proberen te sturen, het afstellen voor wat voor mensen belangrijk is. Het betrekken van de mensen die met het systeem gaan werken of het systeem gaan bouwen én dus ook onderwerp van het systeemontwerp zijn, is noodzakelijk. Dit aspect moet zeer breed bekeken worden, want ontwerpen is zelf een ethische taak. Ook ontwerpprocessen kleuren de wereld en zijn daarmee ethisch.
AI-systemen 'doen' iets, soms zelfs veel. Het zijn wat wel heet 'handelende agenten' en hun ontwerp is daarmee ethisch geladen."

Wat is de grootste uitdaging als het gaat om 'responsible AI'?
"Ten eerste is dat het correct redeneren, zodat je weet dat iets klopt. Ten tweede is een grote uitdaging het aspect van de ethische randvoorwaarden. Voor data gestuurde AI is dat op dit moment een hopeloze zaak. Machine learning kan niet correct redeneren en het is ondoenlijk om het zich aan ethische randvoorwaarden te laten houden."

Wat is het belangrijkste juridische vraagstuk als het gaat om responsible AI?
"Kan het recht met AI worden ondersteund? Dan is meteen de vraag of je de onderbouwing en standpuntbepaling op een rechtvaardige wijze kunt laten ondersteunen met AI. Dit is heel lastig. Iets kan theoretisch gezien interessant zijn maar kan hat vanuit het perspectief van de jurist juist helemaal niet zijn. Wat juridisch relevant werk is, zit juist in wat níet in het model zit. Juristen zoeken altijd naar wat anders is en gaan dus buiten het model om. De jurist is niet bezig met de eenvoudige zaken maar de jurist is altijd bezig met de uitzondering. Met het onderscheid en wat nieuw is. Als een vraagstuk al een duidelijk antwoord had, zou de jurist zich er niet mee bezig hoeven houden." Verheij is van mening dat AI zou moeten faciliteren, zoals juristen in de praktijk echt werken. "We zijn bezig met het analyseren van het denkproces op een manier, zodat het je kan ondersteunen met computers."

Kan jurisprudentie hier iets in betekenen want bij jurisprudentie kan je ergens naar verwijzen?
"Bij jurisprudentie gaat het om een voorbeeld dat je kunt navolgen. Er is zelfs een term hiervoor in de AI: casus-gebaseerd redeneren. Je redeneert aan de hand van andere casussen, precedenten. Stel dat een precedent exact past bij een specifieke situatie, dan kun je dat navolgen en is de vraag makkelijk te beantwoorden. Juristen zoeken ook hier juist naar de verschillen. Zit het niet tóch net anders? Precedenten ondersteunen dus wel en kunnen het denken versnellen. Het is een AI-techniek in ontwikkeling die nu wordt uitgeprobeerd."

Verheij legt uit: "Het gaat om twee aspecten. Aan de ene kant is er wetgeving en die is behulpzaam.
Het past bij de AI uit het verleden, die op logica gebaseerd is. Kennis, je hebt een regel en die pas je toe. Precedenten zijn ook behulpzaam, namelijk je kunt precedenten navolgen, met als aanname dat het precedent moet passen bij het huidige geval. Dit zijn bestaande methoden in de AI maar die laten tegelijkertijd zien dat de echte wereld lastiger is. Regels hebben namelijk altijd uitzonderingen. Juristen zullen dit juist altijd erkennen. Dit is de kern waar intelligentie een rol speelt. Argumentatie, duiding, uitleggen, het goede gesprek voeren en luisteren zijn de belangrijke competenties hierin. Je moet kunnen verantwoorden of een specifiek geval een uitzondering is, of een betoog en de onderbouwing ervan klopt of dat het toch anders is."

CHATGPT HEEFT HET HALLUCINATIEPROBLEEM, HET VERZINT MAAR WAT

Welke rol ziet u weggelegd voor actuarissen als het gaat om 'responsible AI'?
"Actuarissen die iets willen met AI, zouden liefst zelf actief moeten meedenken en meebouwen. Het zijn ook `stakeholders' in het ontwerpproces. Op deze wijze kunnen zij de eigen belangen verwoorden, uitleggen en overbrengen. Actief op tafel brengen van de eigen lessen en van wat je echt van belang vindt. Denk proactief na over wat AI voor je vak kan betekenen. Ga met elkaar in gesprek en sluit aan. Dan krijg je de samenwerking. Denk ook vooral niet te snel dat AI het allemaal toch wel kan. Voorbeeld: Taalgebruik in modellen (ChatGPT) in gebruik is wonderbaarlijk. Bedenk wel dat het nu vooral gebruik van taal is, en dat betekenis nog steeds lastig is in de taal-modellen. Je kunt het gebruiken maar de AI-systemen kunnen het niet begrijpen. Dat is nogal een verschil: gebruik en begrip. Wij begrijpen, zij niet. ChatGPT heeft het hallucinatieprobleem, het verzint maar wat, het heeft geen besef van zijn eigen correctheid, ook niet als het systeem zelf zegt dat iets goed of fout is." ■