Ethiek voor kunstmatige intelligentie

Artificiële intelligentie (AI) is een van de snelst groeiende en meest besproken digitale technologieën van dit moment.

AI wordt inmiddels in alle sectoren van de samenleving gebruikt. We kunnen over artificiële (of kunstmatige) intelligentie spreken wanneer digitale artefacten taken kunnen uitvoeren die intelligentie vereisen als ze door mensen worden uitgevoerd. Kunnen deze systemen ook echt denken of iets begrijpen? De beroemde Nederlandse computer wetenschapper Edsger Dijkstra zei daar ooit eens over: “de vraag of computers kunnen denken, is net zo interessant als de vraag of onderzeeërs kunnen zwemmen.” AI zal aanzienlijke effecten hebben op mensen en op de samenleving. Daar moeten we onze aandacht nu primair op richten. Zo kunnen mensen in steeds meer soorten werkzaamheden worden vervangen door AI systemen. Er is veel discussie over hoe groot die werkgelegenheidseffecten in de toekomst zullen zijn en hoe we daar als samenleving mee moeten omgaan. AI wordt ingezet op tal van zeer gevoelige maatschappelijke domeinen. AI kan het moment van overlijden voorspellen van patiënten op de intensive care, tumoren herkennen, verkeersstromen coördineren, verdachte financiële transacties identificeren, de beste match tussen vacature en kandidaat vinden. Sommige van deze dingen doet AI nu al aantoonbaar beter dan menselijke experts. Kortom: overal waar wordt geclassificeerd, gerubriceerd, herkend, gecategoriseerd, waar patronen moeten worden ontdekt, geoptimaliseerd, gelabeld, geoordeeld, zal AI worden toegepast.

Helaas gaat er daarbij ook het nodige mis. In het Verenigd Koninkrijk werd enige tijd geleden ontdekt dat een algoritme grote aantallen leerlingen lagere scores voor hun eindexamen had toekend dan gerechtvaardigd was. Aanvragen voor vervroegde vrijlating van gedetineerden in de VS werden afgewezen op basis van een AI systeem dat door rechters werd gebruikt en dat een behoorlijke raciale *bias* bleek te hebben. In de Amazon HR systemen die *high potentials* zochten bleek een flinke gender *bias* te zitten. Intelligente risico indicatiesystemen bij de Belastingdienst in Nederland hebben tot onterechte verdachtmakingen geleid. Deze en andere problemen zullen toenemen als we beslissingen steeds vaker *outsourcen* aan intelligente systemen en algoritmische benadering van problemen.

Inmiddels leven er dan ook vele vragen over de ethische beoordeling van algoritmes en slimme software door bedrijven en overheden, over verantwoordelijkheid voor autonome robots, over aansprakelijkheid voor zelfrijdende auto’s en over rekenschap voor de inzet van autonome dodelijke wapens. Welke toepassingen van AI zouden we categorisch willen verbieden? Antwoorden blijven nog uit, evenals samenhangende visies en richtinggevende kaders. Als ze worden gegeven zijn ze vaak alarmistisch of juist naïef optimistisch. Hoe innoveren we op verantwoorde wijze met AI?

Jeroen van den Hoven is universiteitshoogleraar en hoogleraar ethiek en techniek aan de TUDelft. Hij is wetenschappelijk directeur van het *Delft Design for Values Institute*. Hij is *Founding Editor in Chief van Ethics and Information Technology* (Springer Nature).
Hij publiceerde onder meer Evil Online (Wiley Blackwell, 2018), Designing in Ethics (Cambridge University Press, 2017), Information Technology and Moral Philosophy (Cambridge University Press, 2008. Hij is permanent lid van de *European Group on Ethics* van de Europese Commissie die de president van de Commissie adviseert over technologie en ethiek. In die capaciteit was hij de rapporteur van een advies over de ethiek van AI, robotica en autonome systemen. Hij is door de Europese Commissie benoemd als onafhankelijk waarnemer in de *High Level Expert Group on AI* en maakt deel uit van het grootste EU AI project in het komende decennium: *Humane AI*. Hij is ook Nederlands contact van *CLAIRE* (Confederation of AI Labs in Europe). Zie voor overzicht van publicaties en activiteiten www.jeroenvandenhoven.eu

Misschien wel een van de belangrijkste uitgangspunten van verantwoord gebruik van AI zou moeten zijn dat we met het gebruik ervan onze eigen verantwoordelijkheid niet ondermijnen of uithollen.

Een van de meest in het oog springende ethische problemen is de ondoorzichtigheid van AI systemen. Als we niet weten wat we precies aan het doen zijn als we AI systemen gebruiken is het moeilijk verantwoordelijkheid te dragen of te nemen. *Deep learning* toepassingen ontbreekt het aan begrijpelijkheid en verklaring van hun uitkomsten. Ze vormen een *black box*. Niet alleen voor leken en gebruikers, maar ook voor de ontwikkelaars en informatici die er aan werken. We kunnen vaak wel zien of de output correct is of niet (zeker als we daar wat tijd voor kunnen nemen), maar we kunnen niet precies uitleggen hoe het systeem aan een uitkomst komt. Net als met de zogenoemde 'autocomplete functie' van een internet zoekmachine put een AI systeem uit wat het op onnavolgbare wijze heeft geleerd uit de historische data waarop het is getraind. Bevat die dataset een eenzijdig of bevooroordeeld beeld dan wordt dat beeld door het systeem gewoon gereproduceerd. Een chatbot die wordt losgelaten op Twitter in een vrouwvijandelijke, racistische omgeving zal zelf binnen de kortste keren ook misogyne en racistische taal uitslaan. Als het systeem een onderscheid moet maken tussen wolven en huskies op basis van talloze voorbeeld foto's kan het zijn dat de sneeuw op de achtergrond van de wolven het meest opmerkelijke verschil tussen beide hond-achtigen in de aangeboden foto's blijkt te zijn, zodat het systeem een wolf in het groene gras voor een husky aanziet, of een husky in de sneeuw als wolf rubriceert. Dit is een grappig voorbeeld maar als het om mensen en belangrijke beslissingen in hun leven gaat, dan is het zorgelijk en kunnen we ook niet wachten totdat is gebleken is dat een cohort patiënten jarenlang is onderbehandeld, een lichting leerlingen een lager cijfer heeft gekregen dan ze verdienden, of hele postcodes ten onrechte van fraude zijn verdacht.

Er wordt nu met man en macht gewerkt aan *explainable AI*, waarmee we de kracht van *deep learning* zouden kunnen combineren met het verkrijgen van inzicht in de manier waarop resultaten worden bereikt. Dat is een goede ontwikkeling in de richting van het verantwoord gebruik van AI. Maar het is zeker niet voldoende.

We moeten in hoog tempo de institutionele omgeving creëren die het mogelijk maakt om verantwoord te innoveren met AI.

Op andere terreinen hadden we daar meer tijd voor. Schepen en vliegtuigen, voedsel en medicijnen zijn veilig gemaakt en voldoen nu aan onze eisen, niet omdat we deze zaken op hun beloop hebben gelaten, maar omdat we actief hebben geleerd van ongelukken en problemen. We hebben relevante kennis en skills opgebouwd over een lange reeks van jaren. We hebben standaarden, certificaten, controle en audit mechanismen, speciale autoriteiten, methodes, inspectie, keuring, toezicht, wet- en regelgeving, governance, aansprakelijkheid ter zake ingevoerd. We zullen AI en big data eco-systemen, met bijbehorende instituties, mechanismen en socio-technische 'systems of systems' moeten ontwerpen.

De ethiek van AI zal in ieder geval stelling moeten nemen tegen de tendens om menselijke verantwoordelijkheid te relativeren nu onze omgeving steeds slimmer, autonomer, zelflerend en complexer wordt: alleen de mens is moreel verantwoordelijk voor technologie, hoe slim en ingewikkeld de technologie ook lijkt te zijn. Morele termen zoals 'verantwoordelijkheid', 'plicht', 'recht', 'deugd' zijn slechts van toepassing op mensen van vlees en bloed. Wij hebben reeds eeuwen geëxperimenteerd om andere zaken dan mensen verantwoordelijk te houden: voorwerpen en dieren bijvoorbeeld. In het oude Athene konden verdachte dingen voor het gerecht worden gebracht. Eeuwenlang kon de kerk objecten schuldig verklaren. Ratten, honden en ezels en sprinkhanen konden in de middeleeuwen – en nog tot in de 18e eeuw – worden berecht. We zijn met deze praktijken gestopt. En dat is heel begrijpelijk, want we kunnen er totaal geen betekenis aan geven.

De mens is en blijft verantwoordelijk en moet zich niet kunnen verschuilen achter slimme en zelf lerende dingen.

Als we verantwoord willen innoveren met AI dan zullen we ervoor moeten zorgen dat aan de condities voor menselijke verantwoordelijkheid kan worden voldaan. In de 21 eeuw is verantwoordelijkheid een brede ontwerpopgave.