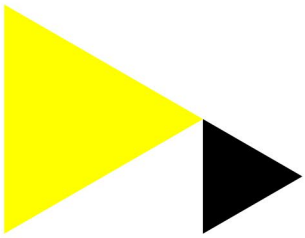
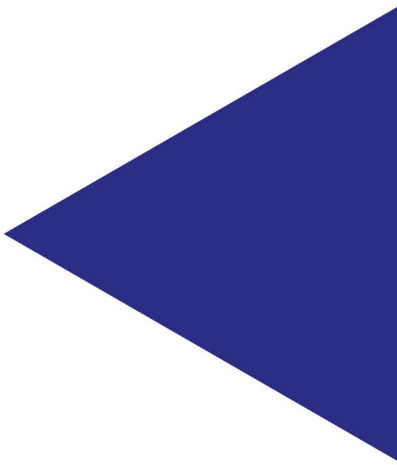
Misschien een domme vraag, maar wat is AI eigenlijk





“Misschien een domme vraag, maar wat is AI eigenlijk?”

Nanda Piersma, Lisanne Henzen

Centre of Expertise Applied Artificial Intelligence, Hogeschool van Amsterdam

Juni 2024

Samenvatting

Er is weinig aandacht voor het precies definiëren van kunstmatige intelligentie, ook wel artificiële intelligentie (AI). Door het bestaan van verschillende interpretaties en definities van AI is niet helemaal duidelijk wat een AI nu wel of niet kan. Met het woord intelligentie in de naam, en de indrukwekkende nieuwe systemen zoals ChatGPT die in 2022 in de markt kwam, ontstaat het beeld dat AI menselijke eigenschappen heeft. Het gevolg is dat AI wordt gebruikt als co-pilot, als digitaal vriendje of zelfs als alwetende. De mens of de AI, wie bepaalt nu eigenlijk? Of is AI als een collega met wie je goed kunt samenwerken? Tijd om nog eens goed te beschrijven wat AI is, hoe AI wordt ingezet en wat nodig is om AI een meerwaarde te geven en verantwoord te maken.

1. Introductie

Kunstmatige Intelligentie, ook wel Artificial Intelligence (AI), is snel opkomend in veel computersystemen en verandert razendsnel. Massaal worden AI-gedreven systemen gebruikt in diverse sectoren, van gezondheidszorg tot financiële diensten. De vraag is of het gehele concept van AI bij iedereen duidelijk is. Laatst vroeg iemand in een vergadering: “misschien een domme vraag, maar wat is AI eigenlijk?” De zogenaamde experts schutterden met woorden als “lerende algoritmes”, “neurale netwerken” en andere technische termen die geen duidelijke uitleg gaven en het antwoord ook juist lastiger kunnen maken.

De definitie van een kunstmatig intelligent algoritme kan variëren afhankelijk van wie ernaar gevraagd wordt. Dit maakt AI tot een complex en vaak verwarrend concept. Bovendien zijn mensen soms sceptisch over AI, omdat het voor velen nog onbekend terrein is. Deze onzekerheid roept belangrijke vragen op, zoals: Wat is de rol van AI? Wat is de meerwaarde van het gebruik van AI? Hoe kan AI verantwoord worden ingezet? Omdat AI nu al, en vrijwel zeker in de toekomst, een grote rol gaat spelen is het belangrijk voor mensen om vertrouwen te hebben in AI-systemen en om meer duidelijkheid te krijgen. Het zal niet voor iedereen van belang zijn om de technische specificaties van AI te kennen om de werking van AI te kunnen begrijpen. In dit artikel wordt ingegaan op wat AI eigenlijk is, zonder in te gaan op de technische specificaties.

De voornaamste boodschap is dat AI een slim gebouwde technologie is die helpt om een nieuwe situatie in te schatten op basis van data over bekende soortgelijke situaties. Wij als mens zijn verantwoordelijk voor de beslissingen van de te nemen acties voor nieuwe situaties. Omdat ook het beslissingsproces geautomatiseerd kan worden lijkt het of een AI-systeem de beslissing “neemt”.

Eerst wordt ingegaan op eerdere definities van AI om tot een aangepaste definitie te komen voor dit artikel. Vervolgens wordt ingegaan op de veronderstelde intelligentie van AI, wat betekent intelligentie van AI? Hiervoor worden de verschillende rollen van AI-systemen voor menselijke taken behandeld en worden daarna de verschillende manieren waarop mensen om (kunnen) gaan met AI

besproken. Vervolgens wordt ingegaan op wat verantwoord gebruik van AI betekent. Als laatste staat de meerwaarde van het gebruik van AI centraal.

2. Definities van Artificiële Intelligence

Er zijn vele definities te vinden over Artificiële Intelligence. Deze definities kunnen erg van elkaar verschillen, maar wat ze wel gemeen hebben is dat ze vrij technisch beschreven worden. Voordat wordt ingegaan op eerder gedefinieerde definities, is het belangrijk om deze bespreking te starten met de realisatie dat een kunstmatig intelligent algoritme een methode is om data te raadplegen en met de kennis daaruit tot een "beslissing" te komen. Waar vroeger vaak werd gevraagd naar de mening van een expert over een bepaald vraagstuk, wordt nu eerder gevraagd: "Wat zegt de data?"

Data vormen een weergave van de werkelijkheid in getallen en gegevens, die met een computer geanalyseerd kunnen worden om inzichten te verkrijgen. Het gaat zelfs verder dan het verkrijgen van inzichten. Met behulp van algoritmes kunnen beslissingen voor vraagstukken geautomatiseerd worden genomen of worden voorzien van een geïnformeerd advies.

Misschien is het daarom goed om te beschrijven wat een algoritme is. Algoritmes zijn computercodes die deterministisch een uitkomst bepalen op basis van data door middel van "als-dan" relaties of beslisregels. Deterministisch betekent dat wanneer een algoritme wordt herhaald, dezelfde uitkomst zal worden geproduceerd, omdat steeds dezelfde stappen worden uitgevoerd. Hierbij wordt een afwegingsproces of inschattingstaak die traditioneel door mensen wordt uitgevoerd, geautomatiseerd. De uitkomst van het algoritme kan bijvoorbeeld een beslissing zijn, of een voorspelling, een aanbeveling, of een planning.

Een kunstmatig intelligent algoritme is ook een algoritme dat gebaseerd is op data-analyse om een uitkomst te bepalen. AI maakt gebruik van een kansmodel om voor nieuwe situaties een uitkomst te berekenen die lijkt op bekende uitkomsten voor vergelijkbare situaties in de data. Dit gebeurt door de AI te trainen met data over situaties en de bijbehorende uitkomsten.

Wat "intelligent" nou precies betekent is een lastig begrip. In eerder definities van Artificial Intelligence wordt het woord 'intelligentie' vergeleken met een persoon, met menselijke intelligentie.

Coppin [1] geeft in 2004 de volgende definitie van AI: "Het vermogen van machines om zich aan te passen aan nieuwe situaties, om te gaan met opkomende situaties, problemen op te lossen, vragen te beantwoorden, apparaat plannen uit te voeren en verschillende andere functies uit te voeren die een bepaald niveau van intelligentie vereisen dat typisch is voor mensen".

In 2018 en 2020 wordt AI door Cioffi et al. [2] en Huang en Rust [3] gedefinieerd als de vaardigheid van machines en programma's om menselijke mentale ervaringen en hun oefenpatronen te simuleren, zoals het vermogen om te leren, te oordelen en te reageren op situaties die niet in de machines zijn geprogrammeerd.

In de dikke Van Dale staat voor kunstmatige intelligentie de definitie: "is de wetenschap die zich bezighoudt met het creëren van een artefact dat een vorm van intelligentie vertoont".

Het rapport Opgave AI van de wetenschappelijke raad voor het regeringsbeleid (WRR, [4]) stelt dat "AI een imitatie of simulatie betreft van iets dat we zelf nog niet volledig begrijpen: menselijke intelligentie". Zij wijzen op de paradox van Moravec: "bepaalde zaken die voor mensen heel moeilijk zijn, zoals schaken of geavanceerde calculus, zijn voor computers vrij gemakkelijk. Maar zaken die

voor ons mensen heel eenvoudig zijn, zoals de perceptie van objecten en motorische vaardigheden als het doen van de afwas, blijken voor computers juist heel moeilijk."

De door de EU[1] gebruikte definitie van artificiële intelligentie (of kunstmatige intelligentie) is: "AI is de mogelijkheid van een machine om mensachtige vaardigheden te vertonen - zoals redeneren, leren, plannen en creativiteit. AI maakt het voor technische systemen mogelijk om hun omgeving waar te nemen, om te gaan met deze waarnemingen en problemen op te lossen om een specifiek doel te bereiken. De computer ontvangt data - reeds voorbereid en verzameld via eigen sensoren, zoals een camera - verwerkt deze en reageert erop. AI-systemen zijn in staat om hun gedrag in zekere mate aan te passen, door het effect van vorige acties te analyseren en autonoom te werken."

Interessant is het woordgebruik bij de beschrijvingen van wat een AI-systeem nu precies doet. AI wordt gezien als een systeem dat in staat is om menselijke taken uit te voeren, en om dat te doen zonder menselijke interventie. Wat is het verschil tussen kunstmatige en menselijke intelligentie? Een AI-systeem zal bijvoorbeeld niet zelf het initiatief nemen om een probleem te gaan oplossen en alleen het probleem dat wordt gegeven gaan oplossen. Het is een rationeel proces, een technologie die wordt gebruikt om een probleem op te lossen, zoals hamers worden gebruikt om een spijker in te slaan, of een Excel sheet wordt gebruikt voor berekeningen. De uitkomst is een inschatting van de input situatie in de vorm van een score, gebaseerd op de meest waarschijnlijke relatie tussen input en de gegeven uitkomst. Net zoals de vraag of een duikboot kan zwemmen[2] niet relevant is, zo is de vraag of AI intelligent is niet relevant. Het is technologie waarmee een cognitieve taak kan worden uitgevoerd.

De OESO heeft in 2023 een (aangepaste) definitie van AI gepubliceerd [5] die de AI als een geautomatiseerd system definieert: *"een op machines gebaseerd systeem dat, voor expliciete of impliciete doelstellingen, afleidt, uit de input die het ontvangt, hoe de output zoals voorspellingen, inhoud, aanbevelingen of beslissingen moet genereren die fysieke of virtuele omgevingen kunnen beïnvloeden. Verschillende AI-systemen variëren in hun mate van autonomie en aanpassingsvermogen na de implementatie/toepassing (deployment) ervan."*

De OESO-definitie benadrukt dus de automatisering door een AI-systeem.

Het kenmerk dat een AI een kansmodel is, die met een bepaalde kans een specifieke uitkomst genereert en waarbij deze kansen zo goed mogelijk worden berekend uit heel veel data, komt niet in de definitie voor. De uiteindelijke definitie van AI die is dit artikel wordt voorgesteld, omvat zowel de automatisering van het geven van een uitkomst met AI als de interpretatie van een kansmodel:

"Kunstmatige intelligentie is een geautomatiseerd computersysteem dat, voor een strikt gedefinieerd probleem, een uitkomst (beslissing, aanbeveling, voorspelling) geeft. De uitkomst is gebaseerd op een kansmodel, getraind op grote hoeveelheden data met bekende "input – uitkomst" relaties, vastgelegd in datagegevens (tekst, afbeeldingen, video, audio, sensordata, enzovoorts). Het systeem traint en berekent voor een nieuwe situatie een zinvolle, bruikbare uitkomst die met de grootste waarschijnlijkheid lijkt op bekende uitkomsten voor gelijksoortige situaties die in de trainingsdata voorkomen."

[1] https://www.europarl.europa.eu/topics/nl/article/20200827STO85804/wat-is-artificiele-intelligentie-en-hoe-wordt-het-gebruikt

[2] Maarten Sukel, AI boek, waarin hij Edgar Dijkstra citeert

3. De veronderstelde intelligentie in AI

Het is belangrijk om na te gaan waar de veronderstelde intelligentie van AI ligt. Waarom heet het Artificiële *Intelligentie* en heeft het geen andere benaming? Intelligentie kan op verschillende manieren worden bekeken en gedefinieerd.

Ten eerste kan een AI-algoritme uitkomsten genereren voor nieuwe input die afwijkt van de trainingsdata, voor situaties die nog niet eerder zijn voorgekomen. Bijvoorbeeld, het bepalen van de beste beslissing voor het wel of niet verstrekken van een hypotheek aan iemand die nog nooit een hypotheek heeft aangevraagd. Mensen kunnen moeilijk voorspellen of deze persoon de hypotheek zal aflossen of in gebreke zal blijven. AI gebruikt een kansmodel om een risicoscore te geven, gebaseerd op eerdere aanvragen met gelijkenissen. Welke gelijkenissen invloed hebben op de risicoscore wordt bepaald door training van het kansmodel. Er is geen deterministische “als-dan” relatie die een beslissing exact bepaalt, AI-uitkomsten geven alleen een inschatting van het risico. De intelligentie van AI ligt in het vermogen om, gegeven de onzekerheid, toch een betrouwbare voorspelling van het risico te geven.

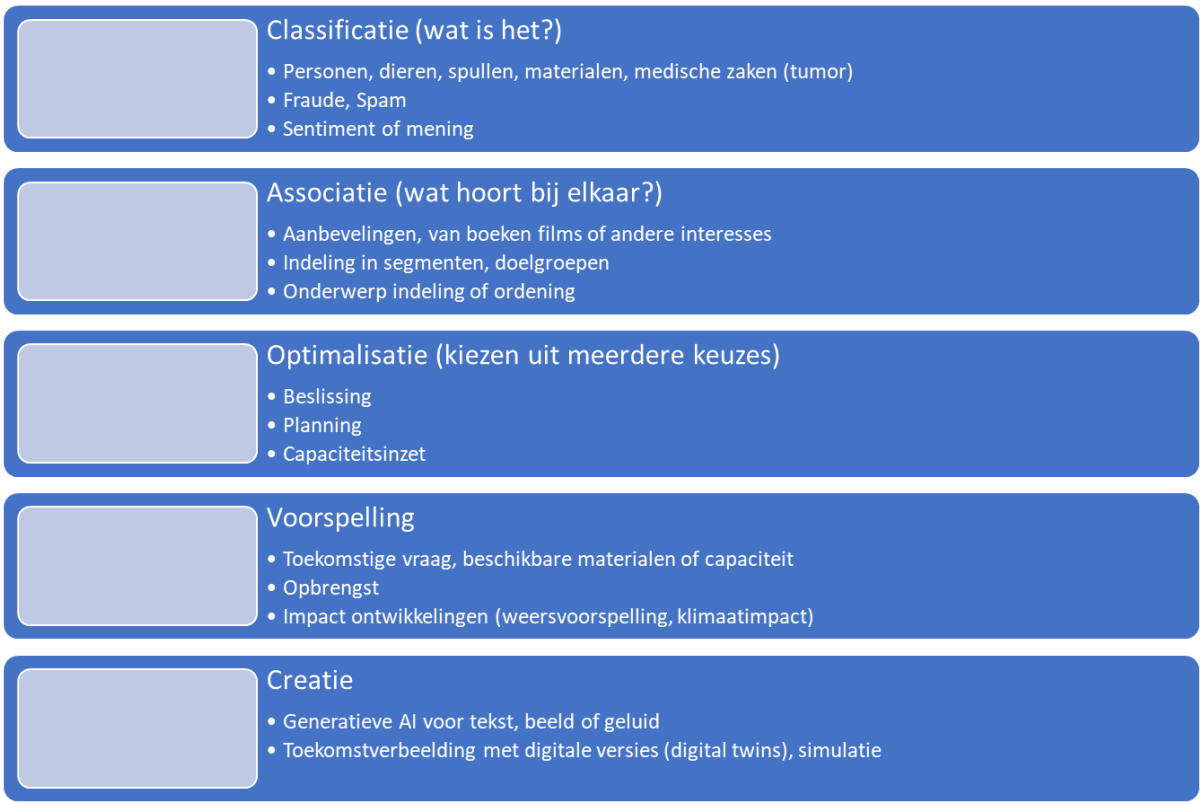
Ten tweede kan een AI-algoritme in grote hoeveelheden data patronen herkennen die mensen niet altijd kunnen vinden en herkennen. Hoe herkennen mensen eigenlijk het verschil tussen een appel en een sinaasappel op een foto, waar kijkt men naar? Misschien is het de kleur van de schil of de vorm van het fruit? Door een algoritme op grote hoeveelheden voorbeeldfoto’s te trainen worden zulke patronen in fruitvormen en -kleuren gemeten en herkend, die met een bepaalde kans bij een appel of een sinaasappel horen. Het trainen van een algoritme op fruit zonder te bepalen naar welke patronen moeten worden gezocht, maakt de gevonden patronen onverwacht, en het algoritme daarmee ogenschijnlijk intelligent.

Een derde intelligentiekenmerk van AI-technologie is de toepassing van een kansmodel op bepaalde mentale taken die mensen kunnen uitvoeren. AI-algoritmes die een nieuwe, originele tekst produceren, of een muziekstuk produceren, dat is een mentale kundigheid die tot nu toe alleen aan mensen werd toegeschreven. Dat geeft een ander gevoel dan technologie die wordt ingezet voor fysieke taken, zoals een ladder om mee omhoog te klimmen, of een fiets om sneller te bewegen. Fysieke hulpmiddelen kunnen heel slim, goed bedacht of technisch vernuftig zijn, maar niet intelligent. Mentale hulpmiddelen kennen we enigszins met rekenmachines, of in veel computerprogramma’s en apps. Daartoe zijn deterministische regels vastgelegd in code, die je als mens kunt narekenen. We begrijpen de regels (code) en zien daarmee dat het een hulpmiddel is. Met AI is dat anders.

Het laatste intelligentiekenmerk van AI is eigenlijk meer de manier waarop met AI-systemen wordt omgaan, en hoe erover wordt gepraat. Er wordt over AI-algoritmes gesproken met technische termen die voor een niet-ingewijde moeilijk te begrijpen zijn, zoals (un)supervised learning en deep learning. Het klinkt allemaal heel ingewikkeld en zal dus wel intelligent moeten zijn.

4. De rol van AI-systemen bij menselijke taken

Om AI goed te kunnen duiden, is het belangrijk om niet te praten over de wiskundige modellen en technische termen, maar over de functie van het AI-algoritme in relatie tot menselijke taken. Met de functie wordt de menselijke taak bedoeld, of het proces dat wordt geautomatiseerd met een kunstmatig intelligent algoritme. Deze indeling staat dicht bij de taken van de mens en is gebaseerd op De Silva en Alahakoon [6], aangevuld met ontwikkelingen rondom generatieve AI voor creatie.



Figuur 1: Functies van AI-tools.

4.1. Voorbeelden van de categorieën

De categorieën geven een beschrijving van het soort taak waarvoor AI kan worden ingezet.

Classificatie

Een herkenningsalgoritme heeft een uitkomst die iets herkent, classificeert met een bepaalde kans. De uitkomst is bijvoorbeeld een selectie van dossiers met de grootste kans op fraude, niet van dossiers met bevestigde fraude. Het kan ook een voorwerp zijn dat wordt herkend op een camera of foto (een fiets, een vuilniszak), een situatie (gevaar, fraude, allerlei ongewenste situaties in het bewakingsdomein), een uitleg of interpretatie van een toestand of mening (zoals een sentiment over een product of marktintroductie, een zogenaamd 'trending topic' (actueel thema)), of de tevredenheid over de metroservice. Er wordt veel gedacht aan beeldmateriaal bij herkenningsalgoritmes, maar het kan ook gebaseerd zijn op teksten (bijvoorbeeld een sentiment herkennen in social mediateksten), of zelfs gebaseerd op sensordata (bijvoorbeeld kwaliteit meten van goederen die worden vervoerd).

Associatie

AI-algoritmes voor associatie worden veelvuldig in relatie gebracht met aanbevelingen. In de sector (online) retail en online producten zijn mensen hier al ruim mee bekend. Aanbevelingen zijn vaak gepersonaliseerd, de ene persoon krijgt andere films te zien in het Netflix-aanbod dan een ander. De gevonden associaties hebben de grootste kans om een goede aanbeveling te zijn. Automatisering van andere taken die associëren gebeurt al veel langer: veel mensen zijn bijvoorbeeld al bekend met het op volgorde zetten van gegevens in een Excelsheet, daar is geen AI voor nodig. Het groeperen van items die op de een of andere manier een relatie hebben (en welke relatie precies) is een

onderwerp van data-analysetechnieken. AI-systemen kunnen getraind worden in het bepalen van een doelgroep die het beste bij een inzamelingsactie passen. Hierbij kunnen veel verschillende kenmerken worden gebruikt die het individuele menselijke associatievermogen te boven gaat. Hetzelfde geldt voor allerlei soorten ordeningen over meer dan één dimensie, die niet meer in een Excelsheet kunnen worden uitgevoerd.

Optimalisatie
Optimalisatietaken bestaan uit elk soort beslissing of planning in het werkproces. Organisatieprocessen zijn meestal gericht op (kosten) efficiency of winstmaximalisatie, economische doelstellingen die optimalisatietaken vereisen. Stel je voor, een transportbedrijf heeft een vloot van vrachtwagens en moet beslissen welke vrachtwagens naar welke locatie moeten gaan om goederen af te leveren. Dit moet zo snel en efficiënt mogelijk gebeuren. AI kan dit soort processen optimaliseren, doordat het systeem traint op ervaringen van eerdere leveringen. Dit geldt voor werkprocessen als planning, procesbesturing en capaciteitsinzet.

Voorspelling
Voorspelmodellen worden gebruikt voor scenario's in de toekomst; welke situatie heeft de meeste kans om te gaan gebeuren? Deze inschattingen worden ondersteund met geautomatiseerde processen die uit historische data voorspellen hoe toekomstige situaties kunnen ontwikkelen. Daarop kan vooruit worden gepland of beslissingen worden genomen om toekomstbestendig te zijn. Als voorbeeld: AI kan huizenprijzen voorspellen op basis van data met kenmerken zoals het aantal slaapkamers, badkamers en de buurt waarin het huis staat. Het algoritme vergelijkt en analyseert deze gegevens met soortgelijke situaties, om te voorspellen wat de prijs zal zijn.

Creatie
Creativiteit van mensen is gebaseerd op kennis en vaardigheden die zijn verkregen met training en ervaring. Met grote datasets van voorbeelden kan geautomatiseerd met AI-systemen nieuwe tekst, beeld en geluid worden gecreëerd. Content creators in allerlei soorten (docenten, marketeers, communicatiemedewerkers, accountants, musici, schrijvers) zien dat hun belangrijkste taken op behoorlijk niveau kunnen worden vervangen door AI-tools.

Er bestaat een overlap tussen het classificeren, associëren, optimaliseren, voorspellen of creëren, als menselijke taak. Zo kan een planner optimaliseren op efficiency om een taak uitgevoerd te krijgen, maar ook rekening houden met arbeidsvreugde van de geplande werknemers. Deze planner herkent het belang van arbeidsvreugde (sentiment), optimaliseert de planning, maar voorspelt eigenlijk ook dat er onrust komt als in de planning geen rekening wordt gehouden met de wens om koffie te drinken met de groep. De categorieën zijn bedoeld om een indeling te geven van de taken die kunnen worden geautomatiseerd (met AI). Een AI-systeem is altijd een algoritme of automatisering van een taak die mensen met meerdere doelstellingen kan uitvoeren. Een algoritme kan de planner vervangen wat optimaal rooster betreft, of om sentiment te herkennen, maar voor de combinatie wordt het vaak niet ingezet. Of er is geen data die wordt gebruikt om het algoritme te trainen op beide doelstellingen. Een indeling in de categorieën is daarom gebaseerd op de inzetpotentie van een AI-systeem, als tool voor classificatie, associatie, optimalisatie, voorspelling of creatie.

5. De mens in omgang met AI

Er zijn meerdere (emotionele) menselijke overwegingen bij het omgaan met de uitkomst van een computersysteem: van negeren (uitzetten, niet gebruiken), een AI-systeem gebruiken als adviseur (waarbij de mens de beslissing autonoom neemt), een AI-systeem als co-pilot gebruiken (zoals men samen met een collega een beslissing bespreekt en neemt), tot het gebruik als alwetende die de beslissing neemt voor de mens.

Het kan zijn dat AI-modellen worden getraind met data, met daarin situaties van eerdere (menselijke) beslissingen die niet altijd de juiste beslissing bleken. Bijvoorbeeld, een scan waarop een radioloog geen tumor zag, bleek later toch een tumor te bevatten. De inschatting dat er geen tumor is omdat er geen tumor te zien is, bleek niet correct. Een getraind algoritme met beeldherkenningstechniek kan de kans berekenen dat een verdachte plek een tumor is, gebaseerd op honderden vergelijkbare gevallen, ook de instanties met de verkeerde inschatting. Met goed databeheer wordt de uitkomst van de scan gecorrigeerd voordat deze wordt gebruikt voor de training van AI-modellen. Hierdoor ondersteunt de AI de radioloog door data te leveren die de menselijke inschatting aanvult met heel veel andere correcte voorbeelden. De radioloog neemt uiteindelijk de beslissing op basis van zowel eigen expertise als de AI-informatie. De AI fungeert meer als co-pilot dan alleen als aanbeveling.

Hoe gaan we als mensen om met de verschillende situaties en belangen rondom de inzet van AI-algoritmes? Hebben we een keuze en kunnen we zelf beslissen, of drijft de technologie ons voort? In het geval van de scan draait het om expertise en ervaring, waarbij extra expertise wordt toegevoegd. Zolang we zelf weten dat we feitelijk een database aan het raadplegen zijn, zoals we een Google zoekvraag intypen, is AI een fijne interface om als mens een database te doorzoeken. De generieke kennis uit de database wordt nu ingezet voor een nieuwe situatie die zelf niet voorkomt in de database, maar die er wel op lijkt.

AI-systemen vervangen dus menselijke inschattingen bij beslissingen. Een algoritme kan bijvoorbeeld een hypotheekaanvraag afwijzen op basis van een hoge risicoscore, de AI bepaalt daarbinnen alleen de risicoscore. Kennelijk is er ook een drempelwaarde bepaald waarboven het risico onaanvaardbaar wordt geacht en de aanvraag wordt afgewezen. De risicoscore wordt bepaald door gebruik te maken van de kennis uit een database van eerdere beslissingen voor een hypotheekaanvraag. Stel dat er een aanvraag is voor een Tiny house in de tuin van kinderen om hun ouders in huis te nemen. Dit is een tweede hypotheek, maar verschilt sterk van een aanvraag voor een vakantiewoning in Spanje. Hoe beïnvloedt deze unieke situatie de risicoscore van de aanvrager, is er wel data van Tiny House situaties in de data waarop de risicoscore is gebaseerd?

Een derde voorbeeld is het gebruik van taalmodellen als een persoonlijke raadgever of digitaal vriendje. Steeds meer mensen onderhouden online contacten, zowel met vrienden uit de fysieke wereld als met nieuwe vrienden die ze uitsluitend online ontmoeten. Deze contacten spelen een grote rol in het sociale leven. Wanneer AI-systemen worden geraadpleegd over gevoelsleven, angsten, onzekerheden of levensvragen, ontstaat een nieuwe situatie: de AI wordt een digitaal vriendje. In plaats van advies aan mensen te vragen, wordt een beroep gedaan op een algoritme, getraind op een database met eerdere vraag-antwoord situaties. Dit kan veilig en anoniem zijn voor gevoelige vragen zoals:

- "Kan het zijn dat ik zwanger ben? Moet ik abortus overwegen?"
- "Ik voel me eenzaam, wat kan ik doen?"

Waar komt de data vandaan met het antwoord op zo'n persoonlijke vraag? Er zijn meerdere voorbeelden waarbij dit soort vragen gesteld worden aan ChatGPT. Nu is ChatGPT gebaseerd op een tekstcreatie algoritme [7], het creëert teksten op een vraag, een prompt, en is getraind op heel veel teksten waarin dezelfde woorden en woordcombinaties voorkomen. Het geeft dus een tekst terug dat het meest lijkt op de data (teksten) waarin dezelfde woordcombinatie voorkomen, de tekst is een inschatting van een antwoord.
Komt de data van uitgeschreven therapiesessies, of is het data uit romans, van internetwebsites, of van allerlei andere soorten bronnen? Hoe geloofwaardig is het antwoord, dat is gebaseerd op de grootste kans op een uitkomst die het meeste lijkt op de data waarop het algoritme is getraind? En voor levensvragen geldt: Hoe geloofwaardig kan een antwoord op dit soort vragen eigenlijk zijn?

Mensen zullen zich niet alleen moeten bezighouden met de uitkomst van de AI zelf, maar ook met de data waarmee de AI-systemen worden getraind. De manier waarop de AI wordt ingezet, én hoe deze wordt gebruikt is van groot belang. De menselijke taak van inschatten en beslissen wordt geautomatiseerd met algoritmes en data. We kunnen dus concluderen dat de AI-software een score geeft om het risico, de onzekerheid van een situatie weer te geven, niet de beslissing zelf.

6. Verantwoorde AI

Er zijn veel voorbeelden van ongewenste situaties door AI-toepassingen, zoals discriminatie in uitkomsten, hoge energieverbruik voor dataopslag en training en algoritmisch management dat leidt tot slechte werkomstandigheden. Veel AI-verrijkte apps verzamelen en verkopen persoonlijke data voor commerciële doeleinden. Deze problemen hebben geleid tot terughoudendheid tegenover gehypte AI-toepassingen die weinig service bieden, en veel nadelen hebben die vaak pas na introductie aan het licht komen.

Hier worden drie dimensies van AI besproken. Deze kunnen helpen om de nadelen te verklaren:

1. **Data**: AI is gebaseerd op representaties van de fysieke wereld in data. Vaak is de data niet representatief voor bepaalde groepen mensen (zoals vrouwen, kinderen, ouderen of niet-witte personen) en bevat deze discriminerende beslissingen. Culturele waarden beïnvloeden wat als een gewenste uitkomst wordt gezien. Veel ongewenste situaties met geautomatiseerde systemen zijn al beschreven. De vraag is of, en hoe, deze problemen opgelost kunnen worden.
2. **Automatisering versus menselijke inschatting:** Algoritmes houden vaak geen rekening met de context. Bijvoorbeeld, een medewerker die steeds te laat inlogt kan ten onrechte lager scoren in functioneren door een algoritme, terwijl een menselijke manager context kan toevoegen en passende afspraken kan maken zonder een lagere beoordeling. We kunnen automatisch laten registreren dat de medewerker te laat inlogt, het inzicht waarom is mensenwerk.
3. **Kansmodel van AI**: AI-systemen kunnen in sommige gevallen nauwkeuriger zijn dan menselijke inschattingen, zoals bij het detecteren van tumoren op CT-scans. AI kan tot 70% nauwkeurigheid bereiken, terwijl menselijke inschattingen soms minder dan 50% nauwkeurig zijn. De nauwkeurigheid van AI hangt af van de kwaliteit van de data waarop het is getraind. Goede beslissingen in de data leiden tot betrouwbare uitkomsten, maar slechte data leidt tot ongewenste resultaten.

Hoe AI wordt ingezet en op welke data het is gebaseerd, is cruciaal. Verantwoorde AI-systemen geven inzicht in hun beperkingen en presenteren uitkomsten niet als absoluut waar. De samenwerking tussen mens en algoritme moet zorgen voor gewenste uitkomsten en ongewenste situaties voorkomen. En: veel beslissingen willen mensen van andere mensen horen, niet van machines.

Er zijn meerdere ethische frameworks en checklists beschikbaar om een AI-toepassing voor de introductie te testen op ongewenste consequenties.

De hogescholen van Amsterdam, Rotterdam en Utrecht doen langjarig onderzoek naar verantwoorde AI-systemen in een programma dat Responsible Applied Artificial Intelligence (RAAIT) heet. Het RAAIT onderzoeksteam geeft de volgende definitie van verantwoorde AI:

"Wij verstaan onder Responsible Applied AI het ontwerpen, ontwikkelen en implementeren van AI-technologieën in de praktijk waarbij rekening wordt gehouden met ethische en sociaal-maatschappelijke kwesties. Dat kan op twee vlakken:

1. *De randvoorwaarden waar een AI-systeem aan moet voldoen, bijvoorbeeld dat het voor mensen nog te controleren of te begrijpen is, en dat bepaalde groepen niet worden uitgesloten of gediscrimineerd (ook wel Trustworthy AI genoemd), en*
2. *Het doel of probleem waarvoor AI wordt ingezet, bijvoorbeeld voor het verbeteren van onderwijs of gezondheidszorg, het tegengaan van klimaatverandering, of andere toepassingsdomeinen die pogen de wereld te verbeteren (ook AI for Good genoemd). 'AI for good' wordt vaak geassocieerd met het inzetten/ontwikkelen van AI om de Sustainable Development Goals (SDG's) te behalen* [8] [9]*."*

In de uitwerking wordt het lastig om te bepalen of de doelstelling om een verantwoord (AI) systeem te ontwerpen kan worden bereikt. Welke randvoorwaarden worden wel en welke worden niet meegenomen? Wie moet precies op welk niveau een AI-systeem begrijpen, en welke consequenties heeft een AI-systeem dat nieuw wordt geïntroduceerd op situaties die niet zijn meegenomen in de training?

Er zijn ook technische uitwerkingen hoe verantwoorde AI-ontwerpen zouden kunnen worden gebouwd. Deze zijn vaak nog theoretisch of conceptueel.

Het boek Responsible Artificial Intelligence van Virginia Dignum [10] gaat over geautomatiseerde ethische beslissingsprocessen: Hoe ontwerp je geautomatiseerde systemen die menselijke waarden en ethische principes kunnen meenemen in het geprogrammeerde beslissingsproces? Daarbij wordt de ethische theorie in drie onderdelen verdeeld in benaderingen van een technisch ontwerp:

- Consequenties: de resultaten (consequenties) doen ertoe, niet de acties zelf. Een actie is juist als het de meest wenselijke consequenties heeft;
- Deontologie: mensen zijn altijd het doel en nooit het middel. Een actie is juist als het voldoet aan bestaande morele (menselijke) principes;
- Waarde-ethiek: Benadruk het karakter van degene die de actie onderneemt. Een actie is juist als deze wordt uitgevoerd door een virtuoos persoon.

In het boek "The Ethical Algorithm: The Science of Socially Aware Algorithm Design" van Michael Kearns en Aaron Roth [11] wordt het principe van "fairness" besproken. Ze stellen dat om bias of discriminatie te verminderen, een algoritme soms minder nauwkeurig moet zijn, dus minder lijkt op de uitkomsten van de trainingsdata. Bij het trainen wordt geprobeerd te voorkomen dat het algoritme leert van data met oneerlijke beslissingen. Soms is een grotere foutmarge beter dan een

oneerlijke uitkomst. Het begrip "unfair" is echter moeilijk te kwantificeren en in een wiskundige formule te vatten.

In "Responsible AI" werken Agarwal en Mishra een wiskundige formule uit die ervoor zorgt dat een algoritme uitkomsten genereert met gelijke kansen voor verschillende groepen [12]. Een voorbeeld is een hypotheekaanvraag systeem dat geen onderscheid maakt tussen mensen van verschillende religies, waarbij de goedkeuringskans gelijk is voor mensen met dezelfde kenmerken, behalve religie. De identificatie van deze groepen gebeurt door de ontwerper van het algoritme en moet bekend zijn tijdens het ontwerp.

De technische modellen om meer verantwoorde systemen te ontwerpen kunnen een verbetering opleveren voor vooraf geïdentificeerde mogelijk ongewenste consequenties van een AI-systeem. De uitwerking van “verantwoord” is helaas zowel technisch als sociaal moeilijk te vatten in richtlijnen of regels, in wetten of in ontwerp. Shannon Vallor stelt in “The AI Mirror” dat de pogingen om AI-veiligheid te bouwen niet gaan werken, omdat menselijke waarden juist een de wortel van het probleem ten grondslag liggen (pagina 9) [13]. Het is niet mogelijk om een verantwoord systeem te bouwen op menselijke waarden uit het verleden, AI is niet gelijnd met onze huidige waarden.

De essentie van een kunstmatig intelligent systeem is dat er met een bepaalde kans een bepaalde uitkomst is. Bij herhaling, of voor een nieuwe situatie, kan de uitkomst weer anders zijn. Een “goed” of “juist” antwoord is er niet, wel een meest gewenst antwoord gebaseerd op eerdere data. Als we het niet eens kunnen worden over de sociale wenselijkheid en als het ook erg moeilijk blijkt om wenselijkheid technisch uit te werken, dan wordt het des te belangrijker om goed na te denken wanneer AI-systemen van meerwaarde kunnen zijn. Grote vraag daaronder is of we overtuigd zijn dat de data waarop de beslissingen zijn bepaald representatief zijn voor de toekomstige situaties en leiden tot gewenste, faire uitkomsten.

7. De meerwaarde van het gebruik van AI-systemen

Het populaire frame rondom het gebruik van AI-systemen is dat we moeten innoveren, dat we als maatschappij en bedrijfsleven mee moeten en dus AI moeten inzetten en bouwen. Er klinkt ook een roep voor meer computerkracht, om meer AI-systemen te kunnen trainen en gebruiken. Gegeven de moeilijkheid om verantwoorde AI-systemen te realiseren is deze haast bevreemdend.

Ook vervangt AI vaak andere technologische systemen en heeft daardoor soms geringe meerwaarde. Als we een taak al hebben geautomatiseerd, waarom zou dat systeem dan vervangen moeten worden door een andere technologie? Voorbeelden zijn camerasystemen in verzorgingshuizen die met AI observeren of iemand valt in een kamer. Voorheen waren er al sensorsystemen die konden registreren of iemand onder een bepaalde hoogte terechtkomt (dus valt) en een signaal sturen naar de verpleegpost. Een camerasysteem met AI doet eigenlijk hetzelfde, maar is wel een registratie van de bewegingen via beeldmateriaal van mensen in de kamers, ook als iemand helemaal niet valt. Een ander voorbeeld is een ingangshek bij een supermarkt die vervangen wordt door een camera. Door het ingangshek te passeren wordt geteld hoeveel mensen er in de supermarkt zijn. Een camera registreert het aantal mensen dat de deur passeert, met een AI-systeem die herkent wanneer mensen waarschijnlijk een beweging maken de supermarkt in.

Eigenlijk is het raar dat we een deterministisch systeem (iemand passeert met zekerheid het ingangshek) vervangen door een kansmodel (het beeld toont iemand die met 88% de kans de supermarkt ingaat).

AI-systemen worden vaak al als bèta-versie in de markt gezet, er lijkt een enorme haast te zijn om als eerste een model te trainen en aan te bieden. Analisten (Jesse Weltevreden, CMI HvA) wijzen erop dat er momenteel tientallen creatie algoritmes met AI worden aangeboden, bijvoorbeeld om het maken van een PowerPointpresentatie te automatiseren. Het is onwaarschijnlijk dat ze allemaal zullen resulteren in een volwaardig businessmodel. Het gevolg is wel dat de modellen slecht worden doorontwikkeld, niet worden getest, een slecht design hebben en een onbekende uitwerking hebben na introductie op de markt. Een voorbeeld is een slimme beademingsmachine die mensen op de Intensive Care alleen zuurstof geeft als een patiënt het nodig heeft en die geen vast toedieningsritme heeft. Omdat de machine niet aangeeft of deze aan of uit staat, bijvoorbeeld met een lichtje die aangeeft of de machine wel werkt, is het niet duidelijk of de beademingsmachine nu slim of stuk is. We hebben gezien dat ChatGPT en andere taalmodellen gingen hallucineren en onzin teksten produceerden. Geautomatiseerde handhavingsprocessen hadden discriminerende en onverwachte effecten, zoals in Nederland de toeslagen affaire liet zien. De geautomatiseerde controle op fouten in de aanvraagformulieren van toeslagen bleek culturele achtergrond als indicator aan te wijzen voor mogelijke fraude.

Daarmee komen we op een afwegingsproces voor de meerwaarde van AI-systemen.

AI-modellen worden veel gebruikt om betere beslissingen te kunnen nemen waarbij de algoritmes risico's inschatten. Bij een samenwerking tussen menselijk inschattingen en AI-inschattingen kunnen betere beslissingen worden genomen op de volgende kenmerken:

- Juistheid - meer correcte beslissingen
- Tijdig - eerder een correcte beslissing maken (denk aan onderhoud aan een machine nog voordat deze kapot gaat)
- Volledig - beslissingen voor meerdere deelnemers
- Vertrouwen - beslissingen erkennen, accepteren en vertrouwen

Het volledig automatiseren van een beslissing betekent dat de kansmodellen van een AI-systeem worden omgezet in beslissingen.

Vaak heeft de toepassing van een AI-systeem een handhavings- of controledoelstelling. Het risico op fraude wordt met een AI-systeem ingeschat aan de hand van de door mensen aangeleverde data. De kwaliteit van de data is belangrijker dan de efficiëntie van het algoritme.

Kunnen we in plaats van handhaven, juist het aanleveren van de gegevens van een aanvrager automatiseren (zoals al gebeurt bij belastingaangiftes)? Mensen hoeven dan alleen te bevestigen dat de gegevens kloppen en aan te geven of ze de toeslag willen aanvaarden, zou dat fraude niet veel effectiever kunnen voorkomen? Het ontwerp vraagstuk is dan omgedraaid: niet machines controleren mensen, maar mensen controleren machines.

Concluderend moeten mensen met systemen leren omgegaan die niet perfect zijn en die uitkomsten geven met een bepaalde kans of nauwkeurigheid. Het is belangrijk om onderscheid te maken tussen menselijke inschatting, menselijke inschatting met een AI-co-pilot, en volledig geautomatiseerde systemen (met AI). Dit vormt feitelijk een ontwerpvraagstuk voor een maatschappij die streeft naar betere beslissingen, waarbij de samenwerking tussen mensen en geautomatiseerde systemen centraal staat.

8. Conclusie

Kunstmatige intelligentie ontwikkelt zich ontzettend snel. Veel mensen kunnen nog sceptisch zijn over AI, omdat het ook een vrij complex onderwerp is. Toch is het belangrijk voor mensen om hiermee om te kunnen gaan, omdat AI een grote rol gaat spelen in de samenleving. Daarom is het verstandig om kunstmatig intelligente technologie goed te definiëren en om de tijd te nemen om de werking en consequenties van het gebruik van AI-systemen te doordenken.

Als eerste is een definitie van Artificial Intelligentie ontwikkeld waarbij de focus lag op een definitie die zowel voor een ontwikkelaar als een gebruiker van AI te begrijpen is. De definitie luidt als volgt:

"Kunstmatige intelligentie is een geautomatiseerd computersysteem dat, voor een strikt gedefinieerd probleem, een uitkomst (beslissing, aanbeveling, voorspelling) geeft. De uitkomst is gebaseerd op een kansmodel, getraind op grote hoeveelheden data met bekende "input – uitkomst" relaties, vastgelegd in datagegevens (tekst, afbeeldingen, video, audio, sensordata, enzovoorts). Het systeem traint en berekent voor een nieuwe situatie een zinvolle, bruikbare uitkomst die lijkt op bekende uitkomsten voor gelijksoortige situaties die in de trainingsdata voorkomen."

Vervolgens is gekeken naar waar de veronderstelde intelligentie van AI zit. AI-systemen geven inschattingen van nieuwe situaties op basis van risicomodellen en patronen die zijn herkend in grote hoeveelheden data. Dit vermogen geeft AI de schijn van intelligentie, vooral in mentale taken zoals het schrijven van teksten of componeren van muziek, die voorheen uitsluitend door mensen werden uitgevoerd. AI lijkt intelligent door het vermogen om patronen te herkennen en uitkomsten te geven, maar het functioneert fundamenteel anders dan menselijke intelligentie. Deze eigenschappen en de presentatie van AI-systemen dragen bij aan hun perceptie als intelligent, ondanks de verschillen met menselijke intelligentie.

Daarnaast is het belangrijk om te weten waar AI voor kan worden ingezet. Technische specificaties van AI-modellen, de wiskundige modellen waaruit de AI-systemen bestaan, zijn niet nodig om het gebruik van AI te kunnen begrijpen. Voor het begrijpen en duiden van AI kan het helpen om de functie van het AI-algoritme te beschrijven in relatie tot menselijke taken. Dit omvat classificatie, associatie, optimalisatie, voorspelling en creatie. Deze functies automatiseren specifieke menselijke taken, met overlappingen in de manier waarop automatisering kan worden toegepast. Het is belangrijk om AI-systemen te evalueren op basis van het vermogen om deze taken te automatiseren en op de potentie als hulpmiddel voor menselijke beslissingsprocessen. Hierdoor kan AI effectief worden ingezet en worden begrepen in de context van menselijke activiteiten en processen.

Omgekeerd is het niet alleen de vraag waar AI voor kan worden ingezet, maar ook de vraag hoe mensen om kunnen gaan met AI. De omgang met AI-systemen vereist complexe overwegingen, waarbij systemen met AI als adviseur, copiloot, of autonome beslisser kan dienen. AI-systemen worden getraind op data en kunnen menselijke inschattingen aanvullen of vervangen, de betrouwbaarheid hangt af van de kwaliteit van de trainingsdata. Het is belangrijk te begrijpen hoe AI menselijke beslissingen beïnvloedt en de rol die AI speelt in menselijke taken, vooral in gevoelige kwesties zoals medische diagnoses en persoonlijke advies. De inzet van AI moet zorgvuldig worden overwogen op basis van de betrouwbaarheid en kwaliteit van de onderliggende data.

Het verantwoord gebruik van AI is een belangrijk onderwerp voor het verhogen van het vertrouwen in deze technologie. AI-toepassingen kunnen gepaard gaan met ethische uitdagingen, zoals discriminatie, energieverspilling en slecht (algoritmisch) management. Maar ook de werking van het systeem is niet altijd goed, zoals te zien is bij hallucinerende generatieve AI-systemen. Deze

problemen hebben geleid tot voorzichtigheid tegenover AI-toepassingen die vaak gebrekkige diensten leveren. De discussie rond AI wordt ondersteund door drie dimensies: de kwaliteit van data, de complexiteit van automatisering versus menselijke inzichten, en de nauwkeurigheid van AI-(kans)modellen. Het ontwikkelen van verantwoorde AI-systemen vereist inzicht in de beperkingen, transparantie over gebruikte data en een afgewogen samenwerking tussen mens en machine om ongewenste situaties te vermijden.

Als laatst is besproken dat, ondanks de hype om AI-systemen te omarmen en te ontwikkelen, bij implementatie de plank wordt misgeslagen. We beschrijven vervanging van bestaande technologieën zonder duidelijke meerwaarde, slecht ontworpen modellen, en onverwachte negatieve effecten zoals discriminatie en fouten. Het is nodig om strak te sturen op meerwaarde van AI-systemen, waarbij samenwerking tussen menselijke inzichten en AI-resultaten kan leiden tot betere besluitvorming.

Bibliografie

[1] B. Coppin, Artificial Intelligence Illuminated, Jones & Bartlett Learning, 2004.

[2] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo en F. De Felice, „Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions," *Sustainability,* p. 492, 2020.

[3] M.-H. Huang en R. T. Rust, „Artificial Intelligence in service," *Journal of service research,* pp. 155-172, 2018.

[4] Wetenschappelijke Raad voor het Regeringsbeleid, „Opgave AI. De nieuwe systeemtechnologie, WRR-Rapport 105," WRR, Den Haag, 2021.

[5] S. Russel, K. Perset en M. Grobelnik, „Updates to the OECD's definition of an AI system explained," OECD, 29 November 2023. [Online]. Available: https://oecd.ai/en/wonk/ai-system-definition-update. [Geopend 13 Mei 2024].

[6] D. S. Daswin en D. Alahakoon, „An Artificial Intelligence life cycle: From conception to production," Patterns, 2022.

[7] OpenAI, „GPT-4 Technical Report," 2023.

[8] R. Vinuesa, H. Azizpour, I. Leite en e. al., „The role of artificial intelligence in achieving the Sustainable Development Goals," *Nature,* 13 Januari 2020.

[9] L. Floridi, J. Cowls, T. King en e. al., „How to Design AI for Social Good: Seven Essential Factors," *Springer,* 3 April 2020.

[10] V. Dignum, responsible artificial intelligence: how to develop and use AI in a responsible way, Springer, 2019.

[11] K. Michael en A. Roth, Kearns, Michael, and Aaron Roth. The ethical algorithm: The science of socially aware algorithm design, Oxford University Press, 2019.

[12] S. Agarwal en S. Mishra, Responsible AI, Springer International Publishing, 2021.

[13] S. Vallor, The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking, Oxford University Press, 2024.