Continueringsnota 2022-2025
Lectoraat Artificial Intelligence

Kenniscentrum
Digital Business & Media

SAMENVATTING

Deze continueringsnota geeft een beeld van de huidige ontwikkeling van artificial intelligence (AI), de technologische en maatschappelijke vraagstukken die deze ontwikkeling met zich meebrengt, en hoe het lectoraat Artificial Intelligence met deze vraagstukken de komende vier jaar aan de slag wil.

De ICT-beroepspraktijk verandert continu onder invloed van de introductie van nieuwe systeemtechnologieën en de steeds luider wordende roep van politiek en samenleving om mensgerichte automatisering en strenge handelingskaders. De opgave om verantwoord om te gaan met de kansen en uitdagingen die AI biedt raakt tal van sectoren en beroepen, en is meer dan een technologievraagstuk alleen. Het is daarom van belang dat de inbedding van deze systeemtechnologie plaatsvindt in nauwe samenspraak en samenwerking met niet-ICT professionals, overheid en burgers.

Het lectoraat Artificial Intelligence stelt zich tot doel om vanuit diepgaande kennis van AI de verbinding aan te gaan met de opleidingen van de hogeschool, met lectoraten van het HU Kenniscentrum Digital Business & Media en daarbuiten, en met externe partners uit de beroepspraktijk. Op die manier wordt de HU-brede opgave om digitalisering in goede banen te helpen leiden gevoed met actuele inzichten en expertise van AI-technologie.

Regionaal positioneert het lectoraat zich als een aantrekkelijke kennispartner voor praktijkgericht onderzoek en talentontwikkeling. Als kernlid van het SPRONG consortium Responsible Applied AI profileren het lectoraat zich nationaal en internationaal als excellente partner in de ontwikkeling van verantwoorde AI-technologie.

"AI is niet zomaar een technologie, maar een systeemtechnologie die de samenleving fundamenteel zal veranderen. [...] Een systeemtechnologie is alomtegenwoordig, kent continue verbetering en maakt complementaire innovatie mogelijk. De ontwikkeling van deze technologie staat momenteel op een keerpunt: de overgang van het lab naar de samenleving, waarin de technologie met de tijd ingebed moet raken."

Wetenschappelijke Raad voor Regeringsbeleid: Opgave AI. De nieuwe systeemtechnologie (11 november 2021)

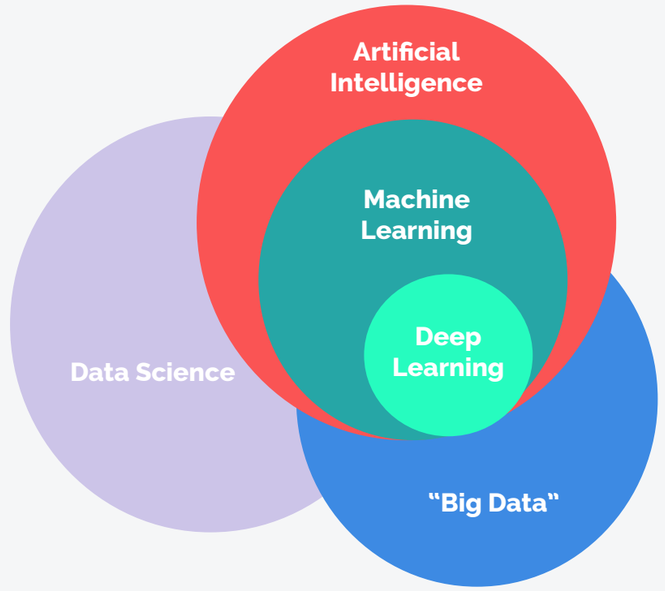
Het kernvraagstuk van deze continueringsnota is: Hoe moet AI worden ontwikkeld en toegepast als krachtige en mensgerichte technologie? Dit vraagstuk wordt uitgewerkt in drie onderzoekslijnen:

Explainable AI richt zich op onderzoek naar en ontwikkeling van hulpmiddelen om AI transparant en uitlegbaar in te zetten

Future Machine Learning richt zich op het verkennen van nieuwe machine learning technologieën en het onderzoeken van de bruikbaarheid daarvan in een praktijkgerichte context

Cooperative AI richt zich op verantwoorde AI-systemen en de plek die zij innemen in socio-technische structuren door tools, prototypen en methoden te onderzoeken die ontwikkelaars helpen om mensgerichte samenwerking met AI te realiseren

Het lectoraat AI zal zich via deze onderzoekslijnen in de komende periode bezighouden met praktijkgerichte onderzoeksprojecten zoals Designing Responsible AI for Media Applications (DRAMA) en Explainable AI in de Financiële Sector; de ontwikkeling van de Internationale Master Human Centered AI; AI-innovaties in het HU onderwijs zoals Ethics Inc. en The Skills Consultant; en het ontwikkelen van een mix van kleine en grote projectvoorstellen voor SIA, NWO, SURF, EU en AiNed.



Figuur 1: Het gebruik van de term 'Artificial Intelligence' in een technologische context (Bron: Thakur, N. (2020). The differences between Data Science, Artificial Intelligence, Machine Learning, and Deep Learning. Medium. Retrieved from https://ai.plainenglish.io/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5)

1

TERUGBLIK

2008 - 2018

Architectuur van Digitale Informatie System

De oorsprong van het lectoraat Artificial Intelligence bevindt zich in 2008, met de kenniskring die destijds is opgericht onder de naam Architectuur van Digitale Informatie Systemen (ADIS) onder leiding van bijzonder lector Wiebe Wiersema. In 2010 is Raymond Slot als dragend lector aangesloten waarna in 2014 het lectoraat voor een nieuwe periode van vier jaar werd gecontinueerd. Doel van dit lectoraat was verbetering, afbakening en professionalisering van het architectuurdomein binnen de ICT. De onderzoekslijnen in de vroege jaren ’10 richten zich op het nut en de waarde van software architectuur en de kwaliteit en de toepassing van die architectuur. Naast deze langlopende onderzoekslijnen werden compliance en waarden, duurzaamheid, open en big data, architectural technical depth, en zorg en technologie als thema’s gedefinieerd.

De onderzoekslijnen werden in belangrijke mate gevoed in samenwerking met andere lectoraten in gesubsidieerde SIA RAAK projecten. Hierbij werd een brug geslagen met het onderwijs door bachelorstudenten van de opleiding ICT in te zetten bij onderzoeks- en ontwikkelprojecten. De software die voor deze projecten werd geschreven werd in het daaropvolgende jaar overgedragen aan een nieuwe lichting studenten, waarmee de waarde van een goede architectuur voor studenten werd geïllustreerd.

Doordat deze projecten doorgaans in het teken stonden van een andere beroepspraktijk dan ICT, was het lectoraat zelf niet altijd even zichtbaar als penvoerder. Een doel van de continueringsnota ADIS 2018 was dan ook om het lectoraat en het penvoerderschap meer zichtbaarheid te geven. Op basis hiervan zijn nieuwe netwerkpartijen betrokken, waaronder Aegon, DUO, Data Kitchen, VU, RUG, TU Eindhoven en Saxion.

Sinds de invoering van de opleiding HBO-ICT in 2015 is er een actieve bijdrage geleverd aan het onderwijs. Zo zijn duizenden ICT studenten betrokken via lessen of projecten bij onderwerpen die zijn aangedragen door drie ICT-gerelateerde lectoraten (ADIS, Betekenisvol Digitaal Innoveren en Procesinnovatie & innovatiesystemen) in afstemming met het Instituut ICT en het beroepenveld.

In de strategische visie op de doorontwikkeling van het lectoraat ADIS in 2018 wordt gekenschetst dat ICT en de rol van ICT in de maatschappij in een razend tempo aan het veranderen is. Cross-sectorale uitdagingen en onderzoeksthema’s ontstaan vanuit de ontwikkeling van digitale sleuteltechnologieën zoals robotica, machine learning, artificial intelligence en blockchain in combinatie met nieuwe manieren van werken zoals scrum, data-driven decision making en agile development.

De onderzoekslijnen van de derde periode van het lectoraat veranderden mee met deze technologische veranderingen. Drie nieuwe onderzoekslijnen (System Design, Lean IT, en Data Science & AI) dienden als perspectief op digitale transformatie binnen organisaties. Hierin legde System Design zich toe op de kwaliteit en toepassen van architectuur, Lean IT op nieuwe manieren van lerend werken en het iteratief ontwikkelen van software, en Data Science & AI op het gebruik en waarde van toenemende hoeveelheden data.

Met de introductie van de huidige kenniscentra in 2017 heeft het toenmalige lectoraat ADIS gekozen zich aan te sluiten bij het kenniscentrum Economisch Sterke & Creatieve Stad, met het oog op de inhoudelijke verbinding met de financieel-economische thema’s en de daarbij behorende digitalisering van diensten.

2018-2020

Intelligent Data Systems

Met de vernieuwing van het lectoraat in 2018 is gekozen voor een nieuwe naam: Intelligent Data Systems (IDS), en daarmee ook met een nieuwe focus die zich richt op het gebruik van data voor de ontwikkeling van intelligente systemen. Kort na aanvang van deze lectoraatsperiode vertrok Raymond Slot als lector, waardoor er enige tijd sprake is geweest van een lectoraat zonder (functioneel) lector met eerst Ander de Keijzer en later Marlies van Steenbergen, terwijl er werd gezocht naar een nieuwe lector die paste bij het profiel van het pas gecontinueerde lectoraat. In september 2019 is Stefan Leijnen begonnen als lector bij het lectoraat IDS, en is de kenniskring gegroeid door collega's van opleidingen ICT, CMD, IBS en IPB aan te trekken als onderzoeker en externe onderzoekers met onderwijsambities te werven.

2020-heden

Artificial Intelligence

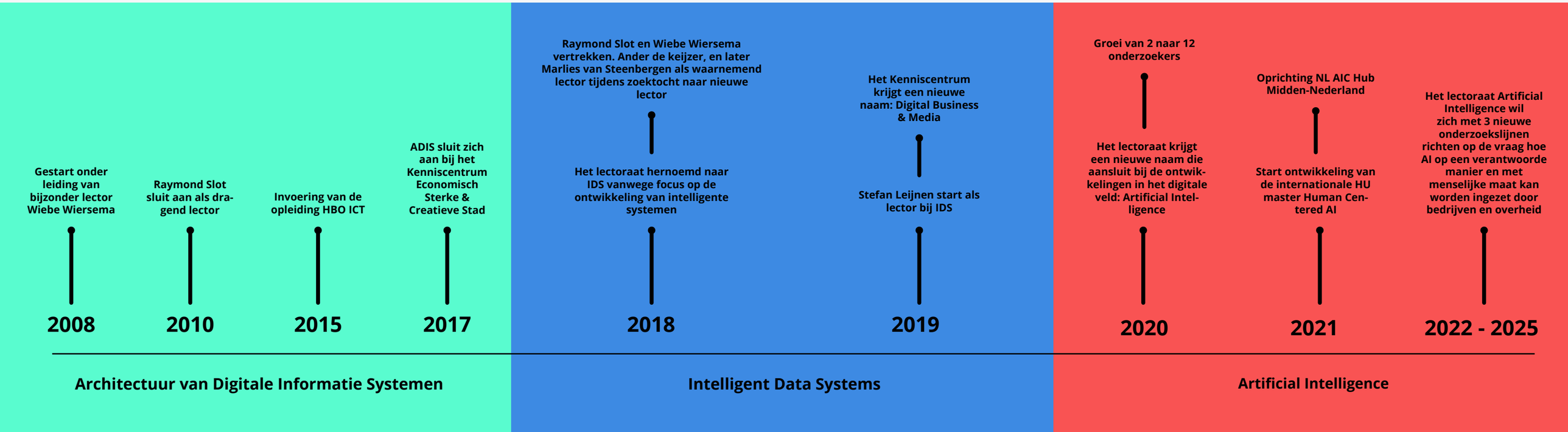
Data Science & AI nemen in toenemende mate een centrale plaats in in de digitale transformatie van de bedrijven, overheden en samenleving en daarmee ook in de ICT beroepspraktijk. In 2020 kreeg het lectoraat haar huidige naam: Artificial Intelligence.

Het lectoraat AI is betrokken bij de ontwikkeling van een internationale master Human Centered AI[1], en een aantal projecten die de kern van de thematiek van dit lectoraat raken, zoals het SPRONG consortium Responsible Applied AI, onderzoeksproject Designing Responsible AI for Media Applications[2] (DRAMA) en onderwijs-innovatieproject The Skills Consultant. Ook is het lectoraat betrokken bij regionale en nationale AI-ecosystemen, zoals de AI-hub Midden-Nederland van de Nederlandse AI coalitie[3], het TTT AI programma[4], de UU/ HU AI labs[5,6], het focusgebied Human Centered Artificial Intelligence van de Universiteit Utrecht[7], ontwikkelingen op AI-gebied van bedrijfsleven, instellingen en overheid in de regio Utrecht[8] en de ambities van de HU op het gebied van digitalisering.

In het licht van bovenstaande ontwikkelingen wordt met deze continueringsnota een pad geschetst voor de doorontwikkeling van het lectoraat AI in de periode 2022-2025, waarbij het lectoraat zich de komende jaren toe zal leggen op de vraag hoe Artificial Intelligence op een verantwoorde manier en met oog voor de menselijke maat kan worden ontwikkeld en gebruikt.

1 hcaim - human centred artificial intelligence masters. (2021). Hcaim. Retrieved from http://www.humancentered-ai.org/
2 talkshow - responsible al voor media. (2021, 7 oktober). YouTube. Retrieved from https://www.youtube.com/watch?v=9y8ae9_31YU
3 NL AI Coalitie. (2021). Algoritmen die werken voor iedereen. Retrieved from https://nlaic.com/
4 About TTT.AI. (2021). ICAI. Retrieved from https://icai.ai/about-ttt-ai/
5 Onze Labs. (2021). Universiteit Utrecht. Retrieved from https://www.uu.nl/onderzoek/ai-labs/onze-labs
6 https://www.aimedialab.nl/
7 Human-centered Artificial Intelligence. (2021, 20 september). Universiteit Utrecht. Retrieved from https://www.uu.nl/onderzoek/human-centered-artificial-intelligence
8 https://www.romutrechtregion.nl/nieuws/ai-hub-midden-nederland-van-start/

Figuur 2: Lectoraten ADIS, IDS en AI door de tijd



2

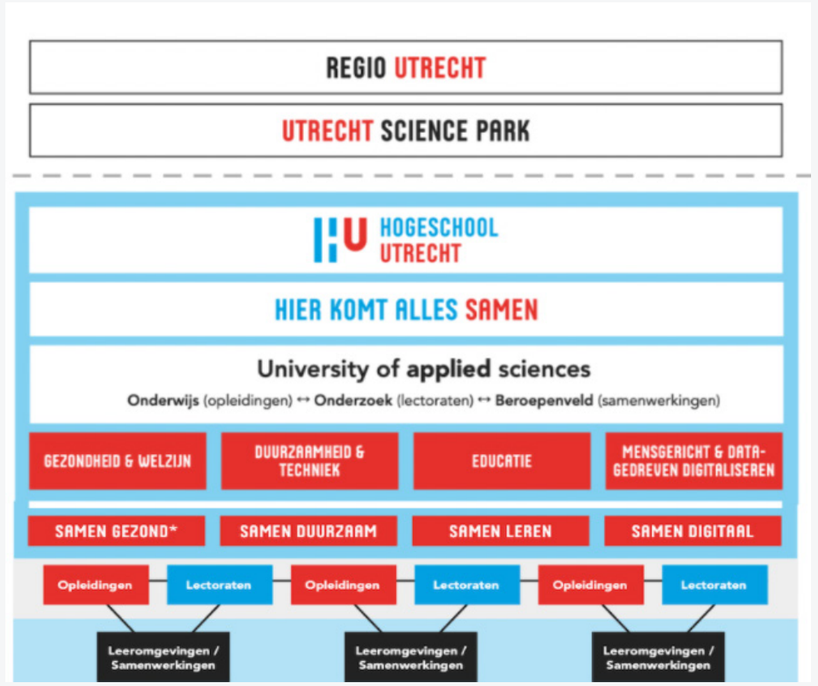
SAMENWERKEN IN ECOSYSTEMEN

Het lectoraat AI is onderdeel van het HU Kenniscentrum Digital Business & Media. In samenwerking met de 7 andere lectoraten van het Kenniscentrum willen we mensen, organisaties en samenleving toerusten om succesvol te worden in een digitale wereld.

De Hogeschool Utrecht heeft gekozen voor vier zwaartepunten waar onderwijs, onderzoek en beroepspraktijk samen komen: Samen Digitaal, Samen Gezond, Samen Duurzaam en Samen Lerend (zie figuur 3). De verbinding tussen deze zwaartepunten stelt ons in staat om bij te dragen aan maatschappelijke vraagstukken in de regio die veelal multidisciplinair van aard zijn.

Het Kenniscentrum is een van de trekkers van het zwaartepunt Samen Digitaal. Utrecht kent relatief veel zakelijke dienstverlening, waaronder veel ICT-gerelateerde bedrijven actief in uiteenlopende markten: media, gaming, digital design, fintech, legal tech en edutech. In werkgelegenheid niet de grootste, maar wel in economische zin. Zij vormt de groeimotor van de Utrechtse economie. Daarnaast wordt digitalisering gezien als sleuteltechnologie voor oplossingen in andere contexten, zoals bijvoorbeeld de zorg.

Het lectoraat AI draagt vanuit een technologische kennis- en expertise-basis met haar drie onderzoekslijnen (zie hoofdstuk 4) een fundament aan van ieder van deze vier vraagstukken, en legt daarbij in het bijzonder een focus op het vraagstuk Responsible Data met een cluster van onderzoeksactiviteiten rond Responsible Applied AI.



Figuur 3: Context en zwaartepunten van de Hogeschool Utrecht

Het thema digitalisering is uitgewerkt naar een onderzoeksprogrammering met 4 actuele digitaliseringsvraagstukken rond het thema mensgericht & datagedreven digitaliseren (zie figuur 4). Het lectoraat AI neemt daarbij een leidende rol voor het vraagstuk Responsible Data: Hoe kunnen we data inzetten om ons leven te verbeteren en systemen te bouwen die by design rekening houden met menselijke waarden als privacy, veiligheid en autonomie? Het lectoraat AI heeft een steunende rol in de samenwerking met andere lectoraten voor de overige drie vraagstukken in het zwaartepunt Samen Digitaal.

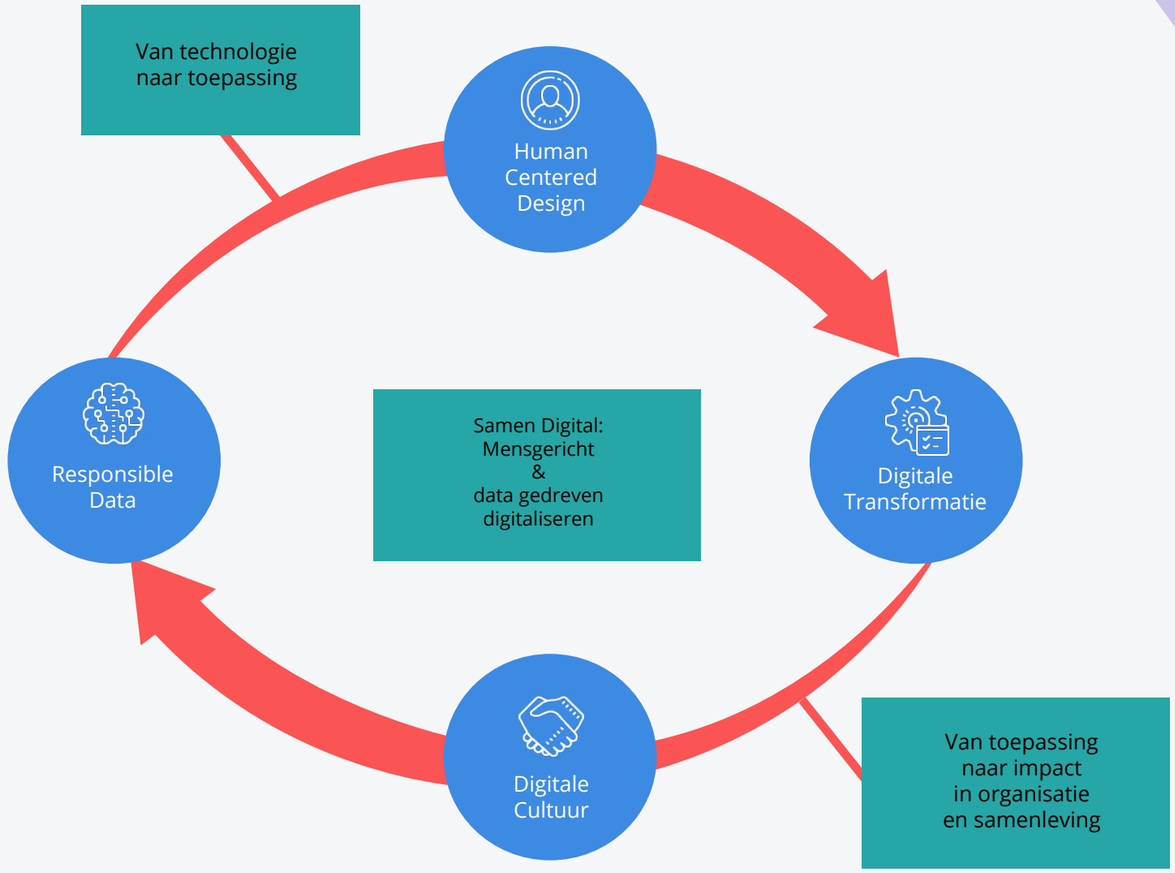
Centraal in de onderzoeksprogrammering voor het vraagstuk Responsible Data staat het recent opgerichte SPRONG consortium Responsible Applied AI, waarbinnen het lectoraat AI in nauwe samenwerking met de lectoraten Betekenisvol Digitaal Innoveren en Kwaliteitsjournalistiek in Digitale Transitie, zich zal richten op methodieken en instrumenten die (toekomstige) professionals helpen om op verantwoorde wijze AI-toepassingen te ontwikkelen en in te zetten in hun organisatie. De toepassing van AI roept nog veel vragen op. Hoe zorg je ervoor dat iedereen gelijkwaardig wordt behandeld? Wie heeft zeggenschap over welke beslissing, hoe staat het met transparantie? En hoe kunnen de behoeften van mens en samenleving centraal staan in de toepassingen? Binnen dit beoogde excellente onderzoekscluster, waar ook de Hogeschool van Amsterdam en de Hogeschool Rotterdam deel van uitmaken, willen wij zorgen dat organisaties AI op een realistische en positieve manier benutten.

Naast samenwerking met andere lectoraten op onderzoeksprogrammering en zwaartepuntontwikkeling zet het lectoraat AI ook nadrukkelijk in op samenwerking met de opleiding via hybride leeromgevingen, waarin competentiegericht onderwijs, labs en lerende netwerken elkaar vinden. De kansen en opgaven die AI biedt dienen zich aan vanuit de beroepspraktijk en zullen via de professionals van de toekomst ook weer hun weg terug naar de beroepspraktijk moeten vinden. Te denken valt bijvoorbeeld aan het gebruik van algoritmen voor kandidaatselectie (Institute for People & Business), digitaal ondernemerschap met AI (International Business Studies) en pluriforme aanbevelingssystemen voor mediacontent (Instituut voor Media).

Vanuit het SPRONG consortium zal in Utrecht worden gestart met een hybride leeromgeving die zich richt op AI & Media. Naar verwachting zullen er hybride leeromgevingen voor andere sectoren volgen naarmate er middelen hiervoor beschikbaar komen, zoals de Kickstart gelden van de Nederlandse AI Coalitie.

Met de internationale master Human Centered AI wordt in samenwerking met het Instituut ICT gebouwd aan future-proof infrastructuur voor onderzoek en onderwijs waarmee de HU zich nationaal en internationaal kan onderscheiden op het gebied van mensgericht digitaliseren. Deze master, gefinancierd door het CEF-instrument van de Europese Commissie, wordt ontwikkeld in samenwerking met de Technologische Universiteit van Dublin, de Universiteit van Napels Federico II, de Technologisch & Economische Universiteit van Boedapest binnen een internationaal consortium met bedrijfspartners en onderzoeksinstituten.

Buiten de muren van de hogeschool zoekt het lectoraat AI de samenwerking op met de Universiteit Utrecht, via aansluiting bij de focus-area Human Centered AI, het UMC, HKU en de AI Hub Midden-Nederland. Als partner van het TTT.AI consortium is het lectoraat betrokken bij de nationale incubator die AI startups en scale-ups door studenten, docenten en onderzoekers wil stimuleren. Lectoraat en kenniscentrum zijn aangesloten op nationale en internationale gremia via lector Stefan Leijnen, die als adviseur bij de Nederlandse AI Coalitie betrokken is bij het opstellen van de onderzoeks- en innovatieagenda, het Nationaal Groeifonds en de internationale relaties van de Nederlandse AI Coalitie.



Figuur 4: Expertisegebieden van het Kenniscentrum Digital Business & Media voor het zwaartepunt Samen Digitaal

LANDSCHAP VAN SAMENWERKINGEN

Onderzoeksinstellingen

Universiteit Utrecht
Freudenthal Instituut
Hanze Hogeschool
Hogeschool Rotterdam,
Hogeschool van Amsterdam
Universiteit van Amsterdam,
Universiteit Twente
Budapest University of Technology and Economics
Rijksuniversiteit Groningen VU Universiteit
Università degli Studi di Napoli Federico II
Stichting Toekomstbeeld der Techniek
European Software Institute
Technological University
Dublin
TU Delft
CeADAR

Universiteiten
HBO's
TNO

Media

RTL
Media Perspectives
Beeld & Geluid
NPO
VPRO

Mediabedrijven
Contentcreators

Publieke diensten

DNB
National Research Council Italy
AFM
NEN

Belastingdienst
Uitvoeringsinstanties
Toezichthouders
Departementen
Provincies
Gemeenten

Utrecht Inc
ICAI
ELSA

Kennis netwerken

TTT.AI
Nederlandse AI Coalitie
AI Hub Midden-Nederland
Utrecht Tech Community
CLAIRE

ICT leveranciers
Digitale consultancybureaus
(studenten) AI startups en scale-ups

ICT

Real AI
Deepkapha
Sogeti
Info Support
Human & Tech Institute
Nathean Technologies
Citel Group
Researchable
Numworx
ICT Institute
Ascom
Audoir

Banken
Verzekeraars
Kleine en middelgrote financiële dienstverleners
Fintech bedrijven

Financiële sector

Floryn, Nordea
Nationale Nederlanden
Verbond van Verzekeraars
Volksbank
Achmea
NVB

Gezondheid & Zorg

UMC UtrechtGrootbanken
Patiëntenfederatie Nederland
VU-MC
Alzheimer Centrum
Winterlight Labs
Immuneering
BioSensics
Erasmus UMC

(Academische) ziekenhuizen
Patiëntenorganisaties
Healthtech bedrijven

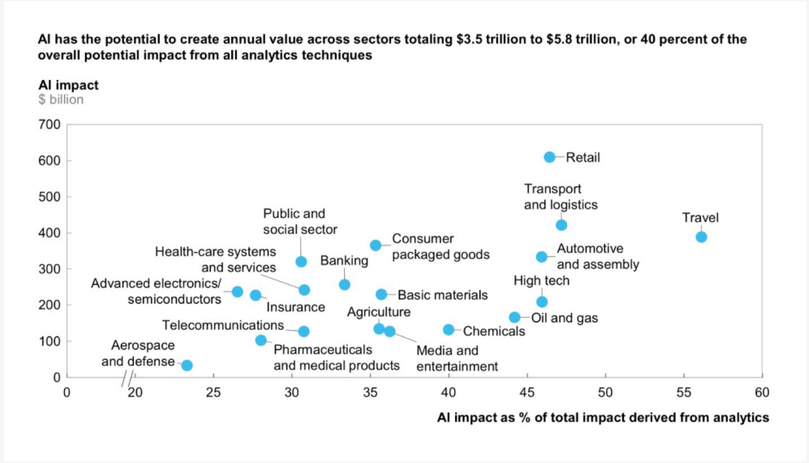
3

STRATEGISCHE VISIE EN DOELEN

Huidige staat van AI

Waar Artificial Intelligence zo'n tien jaar geleden nog een relatief onbekend begrip was, is de impact van AI tegenwoordig dagelijks terug te zien in het nieuws, van platformbedrijven zoals Amazon[1] en Uber[2] die intelligente algoritmes gebruiken om hun werknemers te sturen, deepfakes[3] en filter bubbles[4] die politiek en media in hun greep houden, zelfrijdende auto's[5], drones[6], smartwatches tot robots die ervoor moeten zorgen dat de zorg in de toekomst betaalbaar blijft[7]. De groei die AI in het afgelopen decennium heeft doorgemaakt is het gevolg van de alomtegenwoordige beschikbaarheid van data door het internet, toenemende computerkracht - in het bijzonder met de opkomst van GPU-chips voor het doen van razendsnelle matrix berekeningen[8] – de wildgroei aan nieuwe intelligente algoritmen en computermodellen, en de naadloze integratie van digitale interfaces in het dagelijks leven.

De digitale transformatie wordt mogelijk gemaakt door AI-technologie heeft impact in alle sectoren, en die impact wordt inmiddels ook breed maatschappelijk onderkend[9] – en bevraagd[10]. Advertentiebedrijven zoals Google en Facebook hebben forse bedragen geïnvesteerd in AI om beter te kunnen voorspellen wanneer een gebruiker van hun diensten op een advertentie klikt[11]. De invloed van dit soort AI op de persoonlijke levenssfeer van de gebruikers van hun producten is behoorlijk beperkt, zeker als we die vergelijken met die van producten waar AI inmiddels voor wordt gebruikt.



Figuur 5: AI heeft impact in alle sectoren (Bron: https://www.aei.org/technology-and-innovation/innovation/beyond-the-hype-on-artificial-intelligence-the-reality-of-intelligent-infrastructure-and-human-augmentation/)

1 Regulating Amazon's Warehouse Algorithms Is About More Than Injuries, https://newrepublic.com/article/163588/amazon-warehouse-algorithms-injuries-california-bill

2 Uber faces legal challenge for algorithmic dismissals. Sifted, van https://sifted.eu/articles/uber-algorithm-firing-drivers/

3 "Deepfakes" - a political problem already hitting the EU. EUobserver, van https://euobserver.com/opinion/151935

4 The Social Media Filter Bubble's Corrosive Impact On Democracy And The Press. Forbes, van https://www.forbes.com/sites/kalevleetaru/2019/07/20/the-social-media-filter-bubbles-corrosive-impact-on-democracy-and-the-press/?sh=73e45f2fad42

5 EC-OECD STIP Compass, van https://stip.oecd.org/stip/policy-initiatives/2019%2Fdata%2FpolicyInitiatives%2F26808

6 Europe is now in the fast lane to implementing UAS traffic management systems, van https://www.eurocontrol.int/article/europe-now-fast-lane-implementing-uas-traffic-management-systems

7 Four ways AI can slash healthcare costs around the world. World Economic Forum, van https://www.weforum.org/agenda/2018/05/four-ways-ai-is-bringing-down-the-cost-of-healthcare

8 Understanding the Efficiency of GPU Algorithms for Matrix-Matrix Multiplication, van https://graphics.stanford.edu/papers/gpumatrixmult/gpumatrixmult.pdf

9 Nederlandse AI Coalitie. (2019, October). Position paper "Algoritmen die werken voor iedereen". https://nlaic.com/wp-content/uploads/2019/10/position_paper_algoritmen_die_werken_voor_iedereen-1.pdf

10 Nederlandse AI Coalitie. (2020, November). Mensgerichte Artificiële Intelligentie. https://nlaic.com/wp-content/uploads/2020/11/Manifest-Mensgerichte-Artificiele-Intelligentie_November-2020.pdf

11 Mass personalization: Predictive marketing algorithms and the reshaping of consumer knowledge. Big Data & Society , 7 (2), 205395172095158. https://doi.org/10.1177/2053951720951581

We willen graag dat er goed wordt nagedacht over zelfrijdende auto's, smartwatches die de gezondheid van het lichaam kunnen monitoren[1] en medische specialisten kunnen waarschuwen als er iets dreigt mis te gaan, drones die vanuit de lucht kunnen waarnemen of gewassen er goed bij liggen[2], of in een andere context mensen kunnen doden in een vijandig land, en de algoritmen die helpen bepalen of iemand in aanmerking komt voor een hypotheek of een toeslag van de Belastingdienst[3]. Het is van groot maatschappelijk belang dat de overheid hiervoor strenge en heldere kaders opstelt, en dat die worden nageleefd.

De maximale complexiteit van traditionele ICT-systemen en -architecturen is beperkt tot de complexiteit die wij als mensen kunnen ontwerpen. In zekere zin is digitalisering daarmee beperkt tot het denkvermogen van de ontwerper. Maar AI-systemen waarvoor zelflerende algoritmen (machine learning) worden ingezet,kunnen voorbij die grens reiken en tot zulke complexe systemen en modellen leiden dat wij mensen de onderliggende logica niet langer kunnen volgen. De werking van machine learning systemen is doorgaans moeilijker te begrijpen en daarmee minder voorspelbaar en beheersbaar[4]. De kracht van AI en machine learning (complexe modellen) is daarmee meteen ook haar zwakte .Dit black box probleem vormt een belangrijke en urgente uitdaging voor toekomstige AI-systemen[5].

1 Smartwatch data help detect COVID-19, van Smartwatch data help detect COVID-19

2 Agriculture Drone Market Worth USD 3,697.4 Million by 2027, van https://finance.yahoo.com/news/agriculture-drone-market-worth-usd-154500711.html?guccounter=1

3 Fiscus gebruikte in toeslagenaffaire algoritmes die mensenrechten schenden. NU - Het laatste nieuws het eerst op NU.nl. Geraadpleegd van https://www.nu.nl/tech/6164117/fiscus-gebruikte-in-toeslagenaffaire-algoritmes-die-mensenrechten-schenden.html

4 https://www.bol.com/nl/nl/p/creativity-and-constraint-in-artificial-systems/9200000036483972/

5 https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf

6 Kamphuis, Y., & Leijnen, S. (2021, February 22). Here's what you need to know about the new AI 'arms race'. Retrieved from World Economic Forum: https://www.weforum.org/agenda/2021/02/heres-what-you-need-to-know-about-the-new-ai-arms-race/

7 The High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy A", EU Document, 2019, Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

8 European Union. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence. Retrieved from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

9 European Union. (2021). A European approach to artificial intelligence. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Mensgerichte AI

Het sociaal-maatschappelijke aspect van de technologische vooruitgang, met name de ontwikkeling van AI, is alomvattend en wint recentelijk aan aandacht en invloed. Zo heeft de Europese Unie, in tegenstelling tot China, duidelijk stelling genomen op het ontwikkelen van verantwoordelijke kunstmatige intelligentie[6]. Met de introductie van de vereisten van betrouwbare AI[7] en het recent publiceren van de draft rondom AI-regulering[8], heeft de EU laten zien dat zij zichzelf graag wil positioneren als pionier op mensgerichte kunstmatige intelligentie[9].

"Door datagebruik en algoritmes hebben we de turbo gezet op de onnauwkeurigheden in onze waarneming"

Maxim Februari
VPRO Zomergasten
August 18,2019

Uiteindelijk vallen veel vragen in het digitale domein terug te brengen tot één kernvraag: hoe kan ik dit systeem vertrouwen[1]? Daarmee zijn een aantal lessen geleerd in het verleden. Zo weten we dat het begrijpen van een systeem, het doorzien van de werking, de doelen, en de opbrengsten van het systeem kunnen bijdragen aan vertrouwen. Daarnaast is de mate waarin wij als mens kunnen ingrijpen in een systeem ook een positieve factor voor vertrouwen. Voorbeelden zijn een menselijke piloot die de automatische piloot van een vliegtuig kan overriden, de rechter die uiteindelijk de strafmaat bepaalt of de huisarts die door het huisartsinformatiesysteem wordt geadviseerd. Het is goed denkbaar dat op termijn AI systemen ook leren van dit soort ingrepen en zich hieraan aanpassen.

De afstemming op de menselijke maat heeft in de abstracte zin weinig tot geen betekenis. Pas wanneer een AI-toepassing in een context wordt geplaatst met reële belangen, die goed of slecht af kunnen lopen, worden de afwegingen concreet. Daarbij bieden conflicterende waarden, zoals privacy en veiligheid, autonomie en efficiëntie, een uitnodiging naar creatieve oplossingen waarin die waarden niet langer in conflict zijn[2].

CAUSE: 5 aspecten van mensgerichte AI

Mens en computer verschillen fundamenteel van elkaar in hoe ze doelen formuleren, functioneren, en communiceren.

Waar de mens de computer heeft bedacht en programmeert met een bepaald doel voor ogen, zo kan de mens ook zichzelf programmeren waarbij het doel niet vooraf vaststaat, maar het doel aan zichzelf oplegt. Vanuit een existentieel filosofische perspectief zou men kunnen zeggen dat de mens de computer schept en zichzelf; het eerste als middel, het tweede als doel.

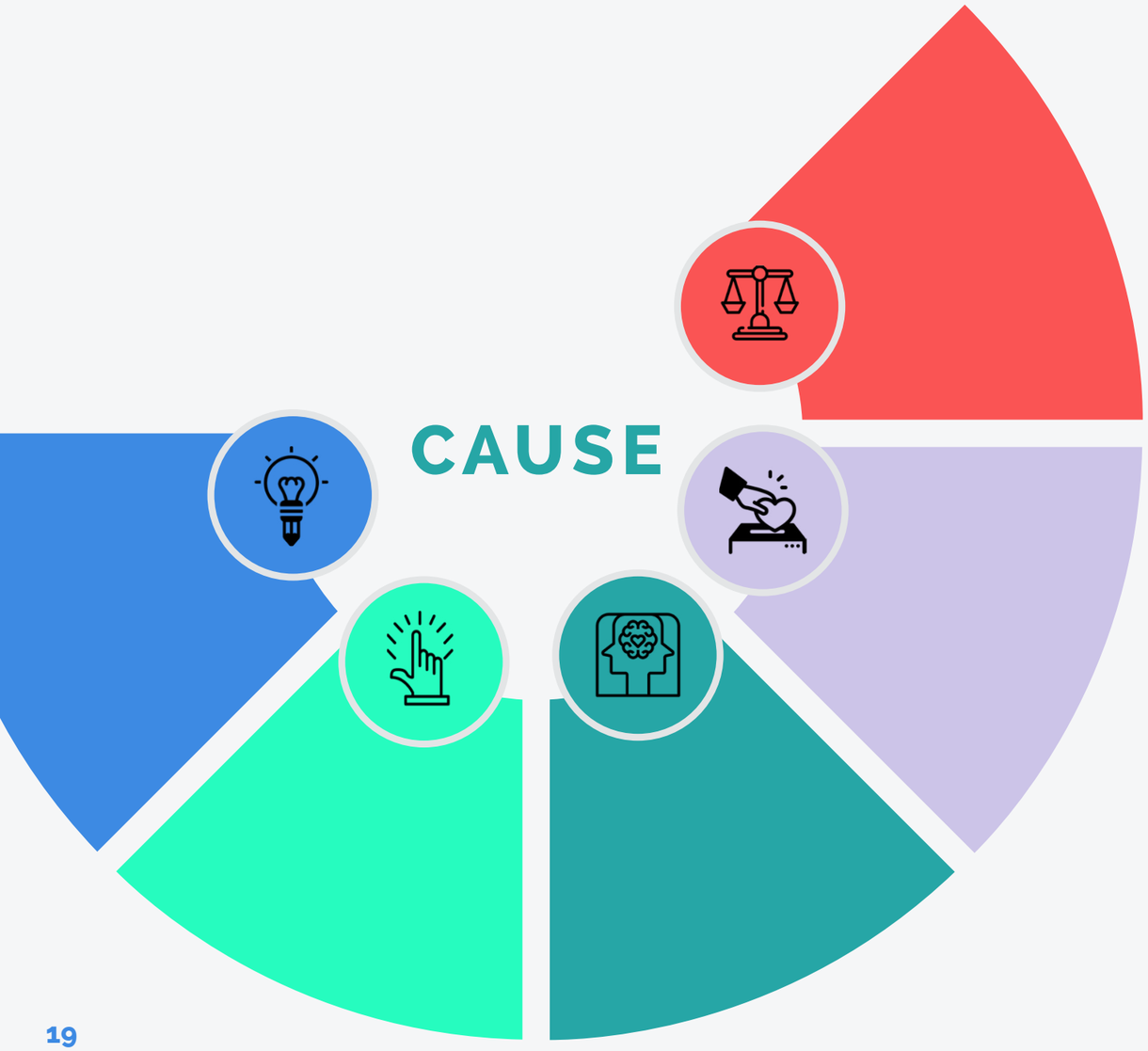
De sociaal-maatschappelijke uitdagingen die bij de doorontwikkeling en toepassing van AI ontstaan gaan over menselijke ervaringen en eigenschappen die niet of nauwelijks te vatten zijn in formules, features en simulaties. Juist daarom stellen we deze eigenschappen centraal in het onderzoek naar de praktische implementatie van AI. We kiezen daarbij vijf niet-mechanische aspecten van menselijk gedrag, die moeilijk in algoritmen te vangen zijn en daarom nader onderzoek verlangen: creativiteit, autonomie, begrip (understanding), gevoel (sentience) en ethiek.

Deze vijf aspecten vormen samen het CAUSE onderzoeksprogramma, dat in de komende periode verder zal worden uitgewerkt met het schrijven van een position paper en geladen door de onderzoeksprojecten. Doel van dit programma is om zowel kadering als inspiratie te bieden aan toegepast onderzoek naar de interactie en harmonisering tussen mensen en intelligente systemen. Hieronder volgt een korte toelichting van de relevantie van de vijf CAUSE aspecten voor mensgericht en datagedreven digitaliseren.

1 Ribes, D., Henchoz, N., Portier, H., Defayes, L., Phan, T. T., Gatica-Perez, D., & Sonderegger, A. (2021). Trust Indicators and Explainable AI: A Study on User Perceptions. Human-Computer Interaction – INTERACT 2021, 662–671. https://doi.org/10.1007/978-3-030-85616-8_39

2 Van Belkom, R., Leijnen, S., Aldewereld, H., Bijvank, R., & Ossewaarde, R. (2020, 16 juni). An Agile Framework for Trustworthy AI. ResearchGate, van https://www.researchgate.net/publication/343106635_An_Agile_Framework_for_Trustworthy_AI

Het woord CAUSE (als acroniem) vat deze vijf aspecten. Daarnaast duidt de term (met gevolg als vertaling) op causaliteit, het begrijpen van de oorzakelijkheid van processen en uitkomsten van complexe systemen dat een actuele uitdaging vormt in onderzoek naar AI-systemen. Als laatste verwijst het (met reden als vertaling) naar de publieke functie van een kennisinstelling en de maatschappelijke taak om, met praktijkgericht onderzoek en het opleiden van de AI professional van de toekomst, de ontwikkeling van mensgerichte AI in goede banen te leiden.



CREATIVITY

Creativiteit stelt ons in staat nieuwe oplossingen te vinden buiten de oorspronkelijke probleemdefinitie, of nieuwe waarde te creëren met behulp van beschikbare bouwstenen. Sommige AI-systemen kunnen creatief gedrag vertonen, en muzikale tophits, kunstwerken en speelfilms genereren op basis van data. Het gebruik van AI-systemen (bijvoorbeeld geautomatiseerde beoordelingssystemen in het onderwijs) kan de creativiteit van mensen die met deze systemen omgaan verrijken of beperken.

AUTONOMY

Autonomie verwijst naar wie beslissingen neemt: zijn machines (of mensen in een toezichthoudende rol) in staat om hun eigen oordeel te gebruiken en hun eigen beslissing te nemen, of zijn de beslissingen gebaseerd op de instructies die ze hebben gekregen? Mensen hebben autonomie, maar kunnen worden beperkt in het uitoefenen van hun autonomie. Machines vertonen doorgaans geen autonoom gedrag, maar bepaalde AI-systemen kunnen autonoom lijken. AI-systemen kunnen, wanneer ze op bepaalde manieren worden toegepast, de menselijke autonomie wegnemen maar ook vergroten.

UNDERSTANDING

Menselijke experts kunnen niet alleen beslissingen nemen, maar zijn in staat om hun beslissingen uit te leggen. Ze zijn zich bewust van hun eigen begrip van het probleemdomein en kunnen redeneren over hun eigen mogelijkheden en beperkingen. Voor AI-systemen is het soms onduidelijk wat voor begrip er is van de onderliggende concepten.

SENTIENCE

Gevoel stuurt de besluitvorming van levende wezens, maar speelt geen rol in de besluitvorming van computers. AI-systemen zijn veelal niet gemaakt om gevoelens mee te nemen in een meting en rekening te houden met emotionele input en output, die soms moeilijk blijkt te kwantificeren.

ETHICS

Als samenleving hebben we bepaalde gemeenschappelijke normen en waarden (bijv. eerlijkheid, geen discriminatie) die ethisch gedrag definiëren, maar zijn we het vaak oneens over de uitwerking van deze ideeën. Bij het implementeren van AI-systemen moet worden overwogen welke ontworpen ethische principes in de AI-systemen zijn ingebed en welke ethische principes mogelijk verdere aandacht behoeven.

AI Wetgeving

De roep om een dergelijke mensgerichte benadering van AI komt mede vanuit Brussel. De Europese Commissie (EC) heeft zichzelf recent tot taak gesteld om voor haar burgers digitale technologie te laten ontwikkelen die aansluit op de menselijke maat, door Europese waarden door te laten klinken in beperkingen die via wetgeving worden opgelegd aan AI-technologie. De AI Verordening zal naar verwachting over 2 tot 4 jaar bepalend zijn voor alle producenten, leveranciers en afnemers van AI. Deze wetgeving kan op termijn leiden tot een Europese industrie van AI-technologiebedrijven die zich expliciet richten op Europese waarden en beperkingen (compliance-by-design).

In het in 2019 gepubliceerde document 'Ethics Guidelines for Trustworthy AI'[1] beschrijft de high-level expert group on AI van de EC een aantal waarden en uitgangspunten waar AI-toepassingen in de toekomst aan dienen te voldoen. In april 2021 is door de EC een voorstel gepresenteerd voor de eerste wet- en regelgeving op AI wereldwijd[2]. Deze wetgeving, die in vorm en functie overeenkomsten vertoont met de GDPR[3], zal binnen enkele jaren definitief worden en in werking treden. Ieder bedrijf dat in Europa AI-diensten produceert, levert of afneemt, zal aan deze strenge eisen moeten voldoen.

Risico-gebaseerde aanpak

Onderdeel van de nieuwe EU AI Verordening vormt een risico assessment waarbij AI-systemen worden geclassificeerd in een van vier risicocategorieën (zie figuur 6). Toepassingen die niet in lijn zijn met Europese waarden zijn onwenselijk en worden verboden. Aan de onderkant van de piramide bestaat een grote groep AI-algoritmes die minimaal risico opleveren en niet met deze wetgeving worden gereguleerd.

Vanuit een onderzoeksperspectief zijn de middelste twee categorieën het meest interessant. Aan toepassingen die een beperkt risico vormen, zoals chatbots of deepfakes, worden transparantie eisen gesteld. Dat wil zeggen dat de menselijke gebruiker waarmee het systeem interacteert weet dat het om AI of gemanipuleerde media gaat, en dus tot op zekere hoogte begrijpt vanuit welk perspectief de interactie plaatsvindt. Voor hoog-risico systemen zal een compliance (self-)assessment moeten worden uitgevoerd, zoals dat nu ook gebruikelijk is voor bijvoorbeeld het CE-keurmerk. Omdat bij veel AI-toepassingen gebruik wordt gemaakt van complexe modellen en processen die voor mensen niet altijd goed te begrijpen en controleren zijn, en die door de tijd heen ook nog eens veranderen, zal de voorgenomen verordening de komende jaren een belangrijke en urgente uitdaging vormen voor de beroepspraktijk.

1 The High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy A", EU Document, 2019, Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

2 European Commission (2021). "Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts". Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

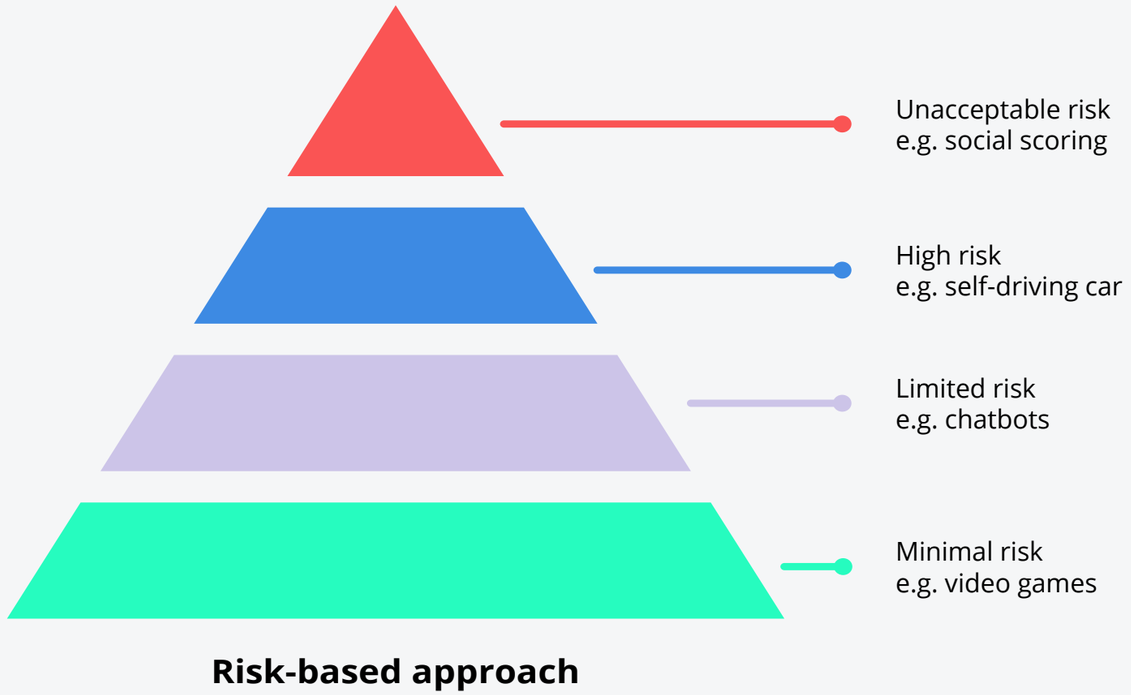
3 Gegevensbescherming in de EU. (2021). European Commission. Retrieved from https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_nl

Praktijkgerichte uitdagingen

Krachtige AI die op een juiste manier wordt ingezet leidt tot transformatie, innovatie, verhoogde efficiëntie en nieuwe oplossingen. Hoe die inzet moet worden versterkt, begeleid of juist ingeperkt door sociale, ethische en juridische (ELS) kaders is een actueel vraagstuk, gegeven het maatschappelijke debat over de wenselijkheid van ongebreidelde technologie en de aangekondigde wetgeving. Het lectoraat AI onderzoekt innovatieve oplossingen die werken in de praktijk, met expertise van zowel technologie als de noodzakelijke randvoorwaarden.

We beschouwen daarmee het conflict tussen waarden en wetgeving enerzijds, en de mogelijkheden die AI-technologie biedt anderzijds, niet noodzakelijkerwijs als dilemma of geforceerde keuze maar als startpunt van een creatieve zoektocht naar betere technologie, innovaties die zorg dragen voor de toekomst, die op verstandige manieren samenwerken met mensen, en verantwoord handelen met oog voor de menselijke maat. Centraal staat daarbij de vraag welke ontwikkelproces- en technische keuzes tot betere technologie leiden. Maar in belangrijke mate is dit ook een menselijke en sociale uitdaging, waar de verantwoorde inzet van technologie geborgd dient te zijn in regelgeving en processen, en er bij het opleiden van de professionals van de toekomst ook duidelijk aandacht nodig is voor de menselijke en maatschappelijke kanten van AI.

Figuur 6: De voorgenomen EU AI Verordening onderscheidt vier risiconiveaus voor AI-toepassingen



4

ONDERZOEKSLIJNEN

De context van deze ontwikkelingen op het gebied van technologische innovatie, de wens om producten en diensten mensgericht te maken, en de aankomende wet- en regelgeving die compliance-by-design vereist geven een kader voor een AI die in dienst staat van mensen en aansluit bij Europese waarden. Daarbij geven de mensgerichte thema's van het CAUSE-programma focus bij het onderzoeken van de kernvraag van deze continueringsnota:

Hoe moet AI worden ontwikkeld en toegepast als krachtige en mensgerichte technologie?

De risico-gebaseerde categorisering van AI-systemen zoals die wordt voorgesteld in de EU AI Verordening biedt een leidraad voor de drie onderzoekslijnen die het lectoraat AI de komende jaren wil doorontwikkelen.

Voor zowel laag- als hoog-risico systemen zal transparantie worden vereist. Bij laag-risicosystemen, zoals chatbots, impliceert transparantie dat gebruikers weten dat ze niet met een menselijke actor maar met een AI te maken hebben, en tot op zekere hoogte ook begrijpen hoe de AI opereert en besluiten neemt. Bij hoog-risico systemen, zoals de beoordeling van kredietwaardigheid, gaat de eis aan transparantie veel verder. Bij zo'n AI-toepassing moet de gebruiker niet alleen weten dat die met een AI te maken heeft, maar ook een specifieke en begrijpelijke uitleg krijgen van zijn/haar beoordeling. Dit roept vragen op in hoeverre het al dan niet aanwezig zijn van een uitleg en het begrip van zo'n uitleg (understanding) invloed heeft op de autonomie van de gebruiker. In welke mate transparante AI technisch te verwezenlijken is, welke behoeftes aan uitleg er bestaan bij gebruikers en werknemers, en welke processen en architecturen transparantie en uitlegbaarheid borgen in een organisatie, zijn vragen die centraal staan in de onderzoekslijn **Explainable AI**.

In de EU AI Verordening worden hoog-risico systemen aan een regime van compliance assessments onderworpen om te toetsen of er in voldoende mate rekening is gehouden met ethische en juridische kaders bij de ontwikkeling en toepassing van een AI product of dienst. Daarom is het van belang die kaders integraal onderdeel te laten zijn van het ontwikkelproces om compliance-by-design af te dwingen (ethiek), technieken te onderzoeken die verantwoorde AI technisch garanderen of tenminste tot op zekere hoogte stimuleren (sentience), te bepalen waar en hoe menselijke actoren een rol zouden moeten spelen bij AI-systemen (autonomie), en hoe dit alles systematisch te borgen zodat de resultaten niet afhankelijk zijn van de individuele keuzes van programmeurs of gebruikers. Dit komt samen in de onderzoekslijn **Cooperative AI**.

Waar de eerste twee onderzoekslijnen voortkomen uit de noodzaak en urgentie AI-systemen te controleren en passend te maken binnen kaders, biedt de derde onderzoekslijn, **Future Machine Learning**, een handvat voor toekomstige AI-innovaties en de nieuwe mogelijkheden die daarmee worden geboden. We onderzoeken welke nieuwe (wetenschappelijke) vindingen en ontwikkelingen hun weg kunnen vinden naar het onderwijs en de beroepspraktijk, zoals de integratie van zelflerende technieken met klassieke logische methoden (sentience), en welke opkomende AI-technologie bruikbaar is in nieuwe contexten (creativiteit).

EXPLAINABLE AI

I. Hoe is explainable AI technisch te verwezenlijken?

II. Welke behoeften aan uitleg bestaan er bij gebruikers, organisaties en toezichthouders?

III. Welke processen, architecturen, tools en best practices helpen explainable AI te borgen in een organisatie?

IV. Hoe kan explainable AI bijdragen aan het herkennen en voorkomen van ongewenste bias en discriminatie?

Praktijkvraag

De onderzoekslijn Explainable AI richt zich op onderzoek naar en ontwikkeling van hulpmiddelen om AI transparant en uitlegbaar in te zetten, met een focus op de sectoren financiële dienstverlening, publieke dienstverlening en media.

De toenemende inzet van AI gaat samen met het gebruik van steeds complexere machine learning modellen. Daarmee worden AI-systemen in toenemende mate ondoorzichtig en wordt het voor gebruikers en AI-ontwikkelaars onduidelijk hoe onderliggende modellen precies tot uitkomsten en beslissingen komen. Explainable AI (afgekort xAI) is ontstaan uit de noodzaak om licht te schijnen in de black box van een model of algoritme, en zodoende een zekere mate van transparantie te bewerkstellingen. Vaak wordt met de term xAI verwezen naar technieken om nieuwe informatie over de werking van een model te verkrijgen, maar dat betekent niet automatisch dat daarmee de black box voor iedere gebruiker wordt geopend. Mensgerichte AI zou daarom moeten streven naar een uitleg die voor alle relevante stakeholders te begrijpen is (inclusiviteit).

Explainable AI wordt gezien als een belangrijke randvoorwaarde voor het mensgericht toepassen van AI. In de huidige Algemene Verordening Gegevensbescherming (AVG) is transparantie van besluitvorming als harde eis vastgelegd; daarnaast eist de AVG nu al menselijk toezicht op de inzet van automatische besluitvorming.

Naar verwachting worden deze wettelijke eisen van transparantie verder uitgebreid in de aankomende European AI Verordening.

In de financiële sector gebruiken zowel grootbedrijven als MKB-organisaties AI bij onder andere leningverstrekking, claimafhandeling en bestrijding van financiële criminaliteit zoals anti witwassen. Het World Economic Forum[1] merkt op dat de ondoorzichtigheid van AI-toepassingen een serieus risico vormt voor het gebruik van AI in de financiële sector: gebrek aan transparantie kan leiden tot verlies van controle door financiële instellingen en schaadt daarmee het vertrouwen van consumenten en maatschappij. Gelet op de cruciale rol van vertrouwen in de financiële sector wordt uitlegbaarheid van de uitkomsten en werking van AI-toepassingen als noodzakelijk beschouwd. Voor alle sectoren en use cases waar vertrouwen een belangrijke rol speelt is xAI van groot belang. Naast de financiële sector bijvoorbeeld ook voor sectoren als publieke dienstverlening (overheid), media en gezondheidszorg.

1 McWaters, J. R., & Galaski, R. (2018). The New Physics of Financial Services: Understanding how artificial intelligence is transforming the financial ecosystem. In World Economic Forum.

PROJECT: XAI IN DE FINANCIËLE SECTOR

Het belang van explainable AI in de financiële sector was reden voor het lectoraat AI om een raamwerk te ontwikkelen om soorten uitleg te relateren aan verschillende groepen stakeholders. Het raamwerk is toegepast in een project van het iForum, een samenwerking tussen De Nederlandsche Bank (DNB), de Autoriteit Financiële Markten (AFM), de Nederlandse Vereniging van Banken (NVB), de Hogeschool Utrecht (HU) en drie grootbanken. Het doel van het project is om in samenwerking met organisaties in de financiële sector praktijkgericht onderzoek te doen naar uitlegbaarheid en daarbij de randvoorwaarden van uitlegbaarheid in beeld te brengen. Dit bestaat enerzijds uit het helder krijgen van de stakeholders en welke uitleg zij verwachten en anderzijds hoe die uitleg het beste tot stand kan worden gebracht.

Een van de bevindingen was dat financiële instellingen explainability weliswaar als een van hun ethische principes hebben omarmd maar nog zoekende zijn naar hoe dit principe te implementeren over de hele levenscyclus van een AI-toepassing. Er blijkt behoefte aan een meer gedetailleerde aanpak voor het implementeren van explainable AI waarbij de mogelijkheden van de techniek en de behoeften van stakeholders op elkaar worden afgestemd.

Het aantal beschikbare technieken om een AI-black box te ontsluiten stijgt snel, maar het is vooralsnog onduidelijk welke techniek in welke situatie het beste resultaat geeft. Daarnaast komen burgers en andere stakeholders ook steeds vaker in het geweer en eisen van gemeente of kredietverstrekker uitleg over beslissingen waar (AI) algoritmes en modellen aan ten grondslag liggen. De toeslagenaffaire[1] is hier een actueel voorbeeld van.

Voor meer achtergrond en informatie, zie het white paper Explainable AI in the Financial Sector[2] en het rapport van iForum iForum: Perspective on xAI[3].

1 De Vrede, T. (2020). Explainable AI: het recht op een goede uitleg. AG Connect. Retrieved from https://www.agconnect.nl/artikel/explainable-ai-het-recht-op-een-goede-uitleg

2 Van den Berg, M., & Kuiper, O. (2020). XAI in the Financial Sector. Hogeschool Utrecht, Lectoraat Artificial Intelligence. Retrieved from https://www.hu.nl/-/media/hu/document-en/onderzoek/projecten/whitepaper-xai.ashx

3 De Nederlandsche Bank & iForum (2020). Perspectives on Explainable AI in The Financial Sector. Retrieved from https://www.dnb.nl/media/jifjijeq/perspectives-on-explainable-ai-in-the-financial-sector.pdf

Onderzoeksvragen

Vanuit de ICT-beroepspraktijk bestaat een behoefte aan mensen en kennis met betrekking tot xAI technieken en hoe die technieken toe te passen in een organisatie. Daarnaast is er brede behoefte aan best practices en processen hoe om te gaan met transparantie en uitlegbaarheid binnen de organisatie, bijvoorbeeld hoe via toezicht verantwoording afgelegd kan worden over explainable AI.

Het is daarbij van belang om xAI niet te zien als doel, maar als middel om te komen tot mensgerichte AI. AI moet vrij zijn van vooringenomenheid of bias, mag niet discrimineren en moet mensen ondersteunen bij het maken van goede beslissingen (zie ook de Europese Ethics guidelines for trustworthy AI). Dergelijke mensgerichte AI vereist uitlegbaarheid van AI. In dat kader komen we tot de volgende onderzoeksvragen:

I. Hoe is explainable AI technisch te verwezenlijken?

II. Welke behoeften aan uitleg bestaan er bij gebruikers, organisaties en toezichthouders?

III. Welke processen, architecturen, tools en best practices helpen explainable AI te borgen in een organisatie?

IV. Hoe kan explainable AI bijdragen aan het herkennen en voorkomen van ongewenste bias en discriminatie?

Plannen

We richten ons met de xAI onderzoekslijn in eerste aanzet op financiële en publieke dienstverlening en de mediasector. Wat betreft de financiële sector wordt een KIEM-project FinTech uitgevoerd waarin we aspecten in kaart gaan brengen die een rol spelen bij het implementeren van xAI en waarbij we deze aspecten relateren aan stadia in de AI-lifecycle (CRISP-DM). Gedurende het KIEM-project wordt het consortium (bestaande uit de Volksbank, Floryn en Researchable) verder uitgebreid om een RAAK-MKB aanvraag in te dienen met de focus op het ontwikkelen van tools en best practices voor het implementeren van xAI. Hierbij geven we antwoord op de eerste drie onderzoeksvragen.

Om meer zicht te krijgen op vragen die leven bij het inzetten van de techniek achter xAI (zoals SHAP[1] en LIME[2]) is het streven in 2022 een aanvraag in te dienen voor een KIEM xAI technologie. Op dit moment loopt er een “Digital Innovation Lab”-project waarbij studenten van het ICT-instituut van de HU een eerste verkenning doen naar de state-of-the-art xAI technieken. Deze kennis bouwen we verder uit in de vorm van een RAAK-MKB aanvraag. Deze projecten gaan met name in op de eerste en derde onderzoeksvraag.

Ernst-Jan Hamel heeft een promotievoorstel ingediend dat zich richt op xAI in de mediasector. Met zijn onderzoek wil hij zich richten op hoe nieuwsmedia fairness, accountability en transparantie kunnen en moeten inbedden in het gebruik van algoritmen voor het modereren van reacties onder nieuwsartikelen.

Explainable AI speelt hierin ook een belangrijke rol, met name voor de tweede en vierde onderzoeksvraag. Voor het onderzoeken van vragen rond xAI die leven in de publieke sector zijn we van plan een RAAK-Publiek aanvraag in te dienen in 2023.

1 Explainable AI (XAI) with SHAP - regression problem. Medium, van https://towardsdatascience.com/explainable-ai-xai-with-shap-regression-problem-b2d63fdca670

2 Understanding model predictions with LIME - Towards Data Science. Medium , van https://towards-datascience.com/understanding-model-predictions-with-lime-a582fdff3a3b

I. Welke technieken stellen ontwikkelaars in staat coöperatieve AI te verwezenlijken?

II. Hoe bereiken we optimale en wenselijke coöperatie tussen mens en AI met behoud van betekenisvolle menselijke controle?

III. Hoe zorgen we dat de wettelijke en ethische verantwoordelijkheid bij de inzet van AI gegarandeerd is vanuit de ontwerp- en organisatieprincipes?

COOPERATIVE AI

Praktijkvraag

De onderzoekslijn Cooperative AI richt zich op het verantwoord realiseren van AI-systemen en de plek die zij innemen in socio-technische structuren (zie figuur 7). Daarbij worden tools, prototypen en methoden onderzocht die ontwikkelaars helpen om mensgerichte samenwerking met AI te realiseren, met een focus op de sectoren health, media, publieke diensten.

AI-systemen worden in toenemende mate onderdeel van een steeds complexere socio-technische realiteit. AI raakt geïntegreerd in de samenleving; iedere burger, gebruiker, consument en werknemer komt in aanraking met intelligente systemen. Deze systemen zullen in meerdere of mindere mate autonoom zijn, beslissingen nemen en interacteren met de omgeving en met mensen[1]. AI toepassingen staan echter niet zomaar op zichzelf, maar hebben interactie met andere systemen of gebruikers en veranderen voortdurend door die interactie. Daardoor vervagen grenzen tussen wat door mensen gedaan wordt en wat door AI-systemen; zo is AI bijvoorbeeld inmiddels beter in het herkennen van huidkanker dan klinische experts.

Een essentieel aspect in de samenwerking tussen gebruiker en AI-systemen is het hebben van begrip en inzicht in de vaardigheden en doelen van de ander, en het hebben van gepast vertrouwen hierin. Maar hoe kunnen we ervoor zorgen dat de mens gepast vertrouwen heeft in de AI-technologie en in diens besluitvorming en autonome acties? Wat vraagt gepast vertrouwen en welke eisen stelt dit aan AI?

Om gepast vertrouwen te verkrijgen in AI-systemen, dient er aandacht besteed te worden aan de plek die AI inneemt: enerzijds met de gebruiker en ontwikkelaar (met belangrijke aspecten als autonomie, betekenisvolle menselijke controle en menselijk toezicht), anderzijds met de maatschappij (met ELS aspecten als ethische naleving, wettelijkheid en verantwoord ontwerp) en tenslotte met de organisatie (d.w.z. met betrekking tot de samenwerking met andere componenten van het systeem, dus mensen en/of andere systemen). Alleen als alle relaties met aandacht worden ontworpen kan er sprake zijn van het borgen van een coöperatie tussen mens en AI.

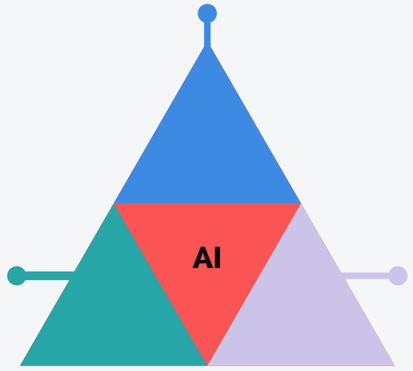
1 Ribes, D., Henchoz, N., Portier, H., Defayes, L., Phan, T. T., Gatica-Perez, D., & Sonderegger, A. (2021). Trust Indicators and Explainable AI: A Study on User Perceptions. Human-Computer Interaction – INTERACT 2021, 662–671. https://doi.org/10.1007/978-3-030-85616-8_39

TEAM

Het team duidt de samenwerking van het systeem met andere componenten (systemen of mensen).

MAATSCHAPPIJ

De maatschappij (organisatie) legt op welke beperkingen er bestaan op de mogelijkheden van het AI systeem, cf. normen en waarden



GEBRUIKER

De gebruiker verwacht een gewenste interactie met het systeem, bijv. level of control, mate van beslissingsvrijheid

Fig 7: Positie van AI in relatie tot de gebruiker, de maatschappij en organisaties.

PROJECT: DESIGNING RESPONSIBLE AI FOR MEDIA APPLICATIONS

AI speelt een belangrijkere rol in mediaorganisaties bij de automatische creatie, personalisatie, distributie en archivering van mediacontent. Dit gaat gepaard met vragen en bezorgdheid in de maatschappij en de mediasector zelf[1] over verantwoord gebruik van AI. Het doel van het DRAMA project is om mediaorganisaties te ondersteunen en begeleiden bij het ontwerpen, ontwikkelen en inzetten van verantwoorde AI-toepassingen, door domeinspecifieke ethische instrumenten te ontwikkelen.

Dit gebeurt aan de hand van drie praktijkcasussen die zijn aangedragen door mediaorganisaties: pluriforme aanbevelingssystemen, inclusieve spraakherkenningssystemen voor de Nederlandse taal, en collaboratieve productie-ondersteuningssystemen.

Voor meer achtergrond en informatie, zie het artikel op de HU website.

1 https://mediaperspectives.nl/intentieverklaring/

Onderzoeksvragen

De verregaande integratie tussen mens en machine maakt het lastiger om te bepalen waar aspecten als verantwoording en verantwoordelijkheid behoren te liggen. Is het wenselijk dat systemen (autonoom) kunnen beslissen over ingrijpende zaken, en hoe wordt de verantwoordelijkheid gedeeld tussen mens en AI? Hoe kan de professional zich verhouden tot ML modellen die zichzelf doorlopend en autonoom aanpassen?

Een notie van controle over intelligente en autonome systemen zou niet alleen de rol van de gebruiker verduidelijken, maar ook die van ontwerpers, ontwikkelaars en beleidsmakers. De achterliggende hypothese is dat de inzet van meer AI-technologie niet noodzakelijkerwijs minder menselijke controle met zich meebrengt, maar wel een nieuw begrip van de menselijk rol en verantwoordelijkheid in het systeem vereist. In dat kader komen we tot de volgende onderzoeksvragen:

I. Welke technieken stellen ontwikkelaars in staat coöperatieve AI te verwezenlijken?

II. Hoe bereiken we optimale en wenselijke coöperatie tussen mens en AI met behoud van betekenisvolle menselijke controle?

III. Hoe zorgen we dat de wettelijke en ethische verantwoordelijkheid bij de inzet van AI gegarandeerd is vanuit de ontwerp- en organisatieprincipes?

Plannen

De hierboven beschreven coöperatie tussen mens en AI wordt gekenmerkt door het beantwoorden van vragen over autonomie en betekenisvolle menselijke controle (gerelateerd aan onderzoeksvraag II) en aan het bepalen van noodzakelijke kaders voor AI-ontwerp (gerelateerd aan onderzoeksvraag III). Het op dit moment lopende project RAAK-Publiek DRAMA bekijkt beide aspecten van Cooperative AI. Het lopende DLO-project Skills Consultant heeft een sterke focus op verantwoord automatiseren van onderwijsondersteuning.

Ook zijn er een aantal RAAK aanvragen in ontwikkeling die inhoud geven aan vragen over autonomie en zich richten op de mogelijkheden om AI in te zetten om de expert te helpen zonder de menselijk maat te verliezen. Daarnaast wordt er in een in voorbereiding zijnde KIEM-aanvraag gekeken naar het samenspel van AI-technologie en de noodzakelijke kaders voor verantwoord AI-ontwerp.

FUTURE MACHINE LEARNING

I. Welke opkomende ML-technieken zijn bruikbaar in nieuwe praktische contexten?

II. Hoe kunnen state-of-the-art ML-technieken worden toegepast door studenten en MKB-bedrijven?

II. Hoe kan ML bijdragen aan het ontwikkelen van systemen waar we mee samen willen werken en leven?

IV. Welke aanpassingen aan ML kunnen bijdragen aan een duurzame en inclusieve samenleving?

Praktijkvraag

De toegenomen bruikbaarheid van machine learning (ML) technieken heeft AI gemaakt tot sleuteltechnologie voor impactvolle innovaties in tal van sectoren. Deze zelflerende algoritmen zijn niet nieuw, het wetenschappelijk fundament van technieken als neurale netwerken, backpropagation en genetische algoritmen is gelegd in de jaren tachtig en negentig van de vorige eeuw en reikt soms nog verder terug[1]. Door convergentie van deze technieken, en met de opkomende rekenkracht van nieuwe hardware (zoals GPU-chips) en de toegankelijkheid van data in steeds grotere hoeveelheden (Big Data) is het gebruik van machine learning in de toepassing van vrijwel alle intelligente systemen actueel en relevant.

Door de explosieve groei van interesse en toepasbaarheid in ML staat de ontwikkeling van het veld zelf allesbehalve stil. Recente trends in ML richten zich op:

- het optimale samenspel tussen lerende machines en de mens (hybride systemen[2], NLP[3]);
- menselijker vormen van intelligentie[4], foundation models[5];
- grotere diversiteit in predictive modelling (personalisatie[6], generatieve AI[7]);
- bredere inzetbaarheid van ML voor het MKB (open data[8], commodificatie);
- en de duurzamere inzet van data, computerkracht en energie (federated learning[9], sustainable computation[10])

Het is dan ook belang voor zowel de rol van het lectoraat in het onderwijs als de doorwerking naar de beroepspraktijk – zowel de ICT-beroepspraktijk als professionals uit andere disciplines - dat nieuwe machine learning technologieën worden verkend en getoetst op bruikbaarheid in een praktijkgerichte context. Wetenschappelijke doorbraken en trends in de implementatie van machine learning vinden op die manier een korte weg naar ontwikkeling en toepassing bij bedrijven en door studenten.

1 The Neural Network Zoo (Van Veen & Leijnen, 2019), van https://www.mdpi.com/2504-3900/47/1/9

2 The Hybrid Intelligence Centre. (2021). Hybrid Intelligence , van https://www.hybrid-intelligence-centre.nl/

3 Education, I. C. (2021, 17 augustus). Natural Language Processing (NLP). IBM, van https://www.ibm.com/cloud/learn/natural-language-processing

4 Open-endedness: The last grand challenge you've never heard of. (2017, 19 december). O'REILLY. https://www.oreilly.com/radar/open-endedness-the-last-grand-challenge-youve-never-heard-of/

5 The Future Brain AI's Paradigm Shift to Foundation Models (Cami Rosso, 2021), van https://www.psychologytoday.com/intl/blog/the-future-brain/202108/ais-paradigm-shift-foundation-models

6 Walch, K. (2020, 16 november). 8 examples of AI personalization across industries. SearchEnterpriseAI, van https://searchenterpriseai.techtarget.com/feature/8-examples-of-AI-personalization-across-industries

7 Brownlee, J. (2019, 19 juli). A Gentle Introduction to Generative Adversarial Networks (GANs). Machine Learning Mastery, van https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

8 Fair principles, van https://www.go-fair.org/fair-principles/

9 Federated learning, van https://federated.withgoogle.com/

10 AI ALS VERSNELLER VAN DE ENERGIETRANSITIE, van https://nlaic.com/wp-content/uploads/2021/09/Position_paper_AI_als_versneller_van_de_energietransitie.pdf

PROJECT: BEELDHERKENNING VAN INVARIANTEN

In deze zomer van 2019 heeft promovendus Roelant Ossewaarde meegedaan aan een prijsvraag van farmaceut en chemiebedrijf Merck. Het ging over een actueel thema in beeldherkenning: als we dingen waarnemen, dan is dat per definitie altijd onder andere omstandigheden. Neem bijvoorbeeld een koe, die staat nooit in precies dezelfde houding, in precies hetzelfde licht of in precies dezelfde hoek ten opzichte van de kijker. Toch kunnen mensen heel nauwkeurig een koe waarnemen, ook al ziet diezelfde koe er altijd anders uit - er is blijkbaar iets aan een koe dat onafhankelijk van de omstandigheden waargenomen wordt. Dat worden "invarianten" genoemd. Merck vroeg om computermodellen die deze vorm van perceptie kunnen nabootsen, met extra voorwaarde dat de voorgestelde oplossing ook lijkt op het proces dat in menselijke hersenen plaatsvindt. Binnen de AI is dit een notoir probleem - alle goed werkende kunstmatige intelligentie-systemen lijken juist niet op de manier waarop mensen het met echte intelligentie lijken op te lossen.

Er deden oorspronkelijk 78 teams mee. Na een eerste schifting bleven er 14 kansrijke voorstellen over. In juni 2019 ging Roelant voor 2 dagen naar het hoofdkwartier van Merck in Darmstadt, om zijn voorstel toe te lichten. De zomervakantie was er om ideeën samen met hun onderzoeksafdeling verder te volmaken. En dat betaalde zich uit: op de laatste vrijdag van augustus vond in Darmstadt de finale plaats en werd zijn voorstel bekroond met de tweede plaats; de eerste prijs ging naar een onderzoeksteam van Huawei/Audi dat met studenten een systeem heeft gebouwd voor invariant-detectie in bewegende beelden. Roelant's model beschrijft hoe menselijke klanken (klinkers) herkend kunnen worden met minimale training - zo'n model is belangrijk om te begrijpen hoe kleine kinderen hun eigen taal leren, ook als ze daar nog helemaal niet zo veel van hebben gehoord.

Onderzoeksvragen

De huidige technology-push van machine learning technieken wordt gedreven door universiteiten, onderzoeksgroepen bij ‘Big Tech’ bedrijven en deep-tech startups. Hogescholen hebben in de verdere ontwikkeling, toepassing en commodificatie van ML een belangrijke rol te nemen: als opleider van AI professionals met actuele kennis en skills, als verantwoord toepasser van AI technologie met een brede maatschappelijke blik, en als motor van AI innovatie voor het MKB.

I. Welke opkomende ML-technieken zijn bruikbaar in nieuwe praktische contexten?

II. Hoe kunnen state-of-the-art ML-technieken worden toegepast door studenten en MKB-bedrijven?

II. Hoe kan ML bijdragen aan het ontwikkelen van systemen waar we mee samen willen werken en leven?

IV. Welke aanpassingen aan ML kunnen bijdragen aan een duurzame en inclusieve samenleving?

Plannen

De onderzoekslijn future machine learning zal zich in eerste aanzet ontwikkelen over twee assen: het aansluiten bij wetenschappelijk onderzoek via promotietrajecten, Europese projecten en NWA-projecten, en het opbouwen van een inhoudelijk sterk ML team. In een later stadium is ook het aanstellen van een L.INT lector voorzien. Via wetenschappelijke projecten in de medische sector, die zich met name zullen richten op onderzoeksvragen 1, 3 en 4, wordt voeling met state-of-the-art ML ontwikkelingen aan universiteiten ontwikkeld en het stelt het lectoraat in staat om ideeën over praktische toepassing te toetsen bij inhoudelijk experts. Daarnaast bieden projecten ruimte voor verdere capaciteitsopbouw in deze onderzoekslijn. Vanaf 2023 wordt, met dit uitgebreide team, gewerkt aan toegepaste onderzoeksprojecten met MKB-partners waarvan de resultaten zullen landen in het onderwijs, eerst middels KIEM en later via RAAK-MKB.

5

DOORWERKING

VOORBEELDPROJECT: ETHICS INC

Ethische aspecten rondom AI, zoals transparantie, non-discriminatie, en privacy krijgen steeds meer publieke belangstelling. Maar hoe krijgt men deze principes ook vooraf geïntegreerd in het ontwerpproces van AI, in plaats van achteraf als nazorg?

Het project Ethics Inc. ontwikkelt een online serious game om bedrijven, overheid, en opleidingen te ondersteunen in de bewustwording van AI en ethiek.

De basis voor het spel zijn de principes uit het EU-rapport “Ethics guidelines for trustworthy AI”, praktisch en sprekend gemaakt door middel van agile spelelementen.

VOORBEELDPROJECT: MASTER HUMAN CENTERED AI

Om de digitale vaardigheden rond AI te versterken heeft de Europese Commissie subsidies verstrekt voor de ontwikkeling van vier internationale masters op het gebied van AI, waaronder de master Human Centered AI.

Het lectoraat AI is in samenwerking met HU Instituut ICT, de Technologische Universiteit van Dublin, de Universiteit van Napels Federico II en de Technologisch & Economische Universiteit van Boedapest gestart met de ontwikkeling van de internationale master Human-Centered Artificial Intelligence die in het studiejaar 2022/2023 van start zal gaan.

De master gaat zich richten op de mensgerichte en ethische aspecten van Artificial Intelligence (AI). De ethische richtlijnen voor betrouwbare AI vanuit de EU worden hierbij als uitgangspunt genomen. Tijdens de ontwikkeling van deze Europese master werken experts, onderzoekers, onderwijsinstellingen. en grote ICT organisaties vanuit deze verschillende Europese landen samen.

Focus	Modeling	Implementation	Evaluation	
Theme	Classic ML	Deep Learning	Future AI	
Period	A	B	C	D
Technical BoKS	Fundamentals of AI, Machine Learning	Advanced AI, Deep Learning	Future AI, Learning	Master Thesis Project
Practical Work	AI Modeling: Data prep, Clean pre pro-cess	AI in Action Organizational AI	Socially Responsible AI	
Ethical BoKS	Ethics by De-sign, Bias Detection & Mitigation	Risk & Explainable AI	Compliance, Legality & Humanity	

DOORWERKINGSTABEL

	Beroepspraktijk	Onderwijs en professionalisering	Wetenschap
Kennis - ontwikkeling (Researching)	• Kennis en tools die ontwikkeld worden in het kader van transparante en semi-autonome AI vinden via gezamenlijke onderzoeksprojecten de weg naar de beroepspraktijk. • Er vindt kennisdeling plaats over de impact van AI via openbare lezingen, interviews en maatschappelijke publicaties.	• In gastcolleges, colloquia en summerschools wordt kennis gedeeld over nieuwe mogelijkheden met AI technologie. • Met opleidingen wordt nagedacht over kansen en uitdagingen met AI die specifiek zijn voor het vakgebied, zoals contentcreatie (IvM) en de vraag hoe mensen en machines in de toekomst gaan samenwerken (IPB). • Kennis en tools die ontwikkeld worden in het kader van transparante en semi-autonome AI vinden via onderwijs en curricula hun weg naar de leerroutes van studenten.	• Door publicatie van de resultaten van het onderzoek in peer-reviewed journals en op wetenschappelijke conferenties te publiceren, wordt de wetenschap gevoed met praktijkgerichte kennis, instrumenten, best practices en voorbeelden uit de beroepspraktijk • Het CAUSE onderzoeksprogramma wordt in een position paper uitgewerkt en verder geladen met wetenschappelijke publicaties. Hiermee wordt een raamwerk opgetuigd voor praktijkonderzoek naar mensgerichte AI.
Persoons - ontwikkeling (Learning)	• In het kader van onderzoeksprojecten worden professionals uit de verschillende beroepspraktijken in samenwerking betrokken bij actuele wetenschappelijke ontwikkelingen, innovaties en nieuwe contexten voor implementatie van AI producten en diensten. • Dit biedt hen nieuwe handelingsperspectieven, zoals de intentieverklaring voor verantwoord gebruik van AI in de mediasector.	• Onderzoekers gaan vanuit hun aanstelling bij het lectoraat ook doceren in het onderwijs, zoals de masters Data-Driven Business en Data-Driven Design, en zijn actief bij de Summerschool Machine Learning en de Master Human Centered AI.	• Docenten worden uitgedaagd in de rol van AI onderzoeker, waarmee de verbinding met de instituten aangegaan wordt, uit zich dit onder andere op de korte termijn in een promotietraject voor Ernst-Jan Hamel (IvM) en post-doc trajecten voor Ouren Kuiper, Roelant Ossewaarde (ICT) en Sietske Tacoma (IPB). • Studenten die met onderzoeksstages en -projecten bij het lectoraat betrokken zijn worden uitgenodigd hun werk te publiceren in journals en conferenties.
Product - ontwikkeling (Designing)	• Binnen onderzoeksprojecten worden tools, prototypes, instrumenten, trainingen en best practices ontwikkeld die in onderlinge samenwerking de weg vinden naar de beroepspraktijk. • De serious game Ethics Inc. helpt AI ontwikkelaars met het bespreekbaar maken van sociale, maatschappelijke en ethische problemen die kunnen ontstaan door AI producten en diensten.	• De inzet van mensgerichte AI in het onderwijs wordt door het lectoraat onderzocht in het kader van het Skills Consultant project, om een tool te ontwikkelen die studenten en docenten faciliteert het kiezen van leerroutes voor soft skills.	• In studentenprojecten, afstudeerstages en binnen onderzoeksconsortia worden AI prototypes ontwikkeld die ten dienste staan aan wetenschappelijk onderzoek naar de technische (on)mogelijkheden van AI. • Het lectoraat streeft hierbij naar de standaarden van open data en open science.
Systeem - ontwikkeling (Changing)	• Het lectoraat wil een actieve bijdragen leveren aan de ontwikkeling van learning communities op de HU, waarin In het kader van de SPRONG Responsible Applied AI zal met deze hybride leeromgevingen in 2022 een start worden gemaakt. • De mediasector geldt hierbij voor de HU als proeftuin voor andere sectoren.	• Het lectoraat werkt, in samenwerking met het Instituut voor ICT, met een Europees consortium aan de ontwikkeling van een master Human Centered AI die AI professional van de toekomst wil opleiden om technische kennis toe te passen met een brede maatschappelijke blik. • Dit vergt het nader tot elkaar brengen van ICT en Design, en studenten op te leiden die zich thuis voelen in beide disciplines.	• Binnen de AI Hub Midden-Nederland draagt het lectoraat bij aan het regionale ecosysteem van AI onderzoek, met de Universiteit Utrecht en het UMC als go-to kennispartners. • Voorbeeld hiervan is het recent opgestarte initiatief van de UU/HU AI Labs.

6

TEAM

Stefan Leijnen

Lector
Interesses: AI governance, human-centered AI, machine creativity & sentience
stefan.leijnen@hu.nl

Marlies Sandee

HHD
Interesses: AI in onderwijs, student & ondernemerschap, change management
marlies.sandee@hu.nl

Huib Aldewereld

HHD
Interesses: verantwoord AI ontwerp, sociale coördinatie in normatieve systemen, cooperative AI
huib.aldewereld@hu.nl

Martin van den Berg

Docent-onderzoeker
Interesses: Explainable AI, enterprise architecturen, digitale transformatie
martin.m.vandenberg@hu.nl

Ouren Kuiper

Docent-onderzoeker
Interesses: explainable AI, human-centered AI, AI ethiek & wetgeving
ouren.kuiper@hu.nl

Sietske Tacoma

Docent-onderzoeker
Interesses: transfer learning, machine learning, AI in onderwijs
sietske.tacoma@hu.nl

Roelant Ossewaarde

Promovendus
Interesses : computationele linguïstiek, neurale netwerk architecturen, Bayesiaanse uitlegbare modellen
roelant.ossewaarde@hu.nl

Tina Mioch

Onderzoeker
Interesses: cooperative AI, human-AI interaction, designing responsible AI
tina.mioch@hu.nl

Ida Buikema

Management ondersteuning
ida.buikema@hu.nl

Rianne van Os

Docent-onderzoeker
Interesses: data science, machine learning
rianne.vanos@hu.nl

Roland Bijvank

Docent-onderzoeker
Interesses: process mining, deep learning, generative adversarial networks
roland.bijvank@hu.nl

Ernst-Jan Habel

Promovendus
Interesses: journalistiek, AI & ethiek, publieke waarden & democratie
ernst-jan.hamel@hu.nl

Jonas Moons

Docent-onderzoeker
Interesses: Generatieve modellen, AI & menselijk gedrag, predictieve algoritmen
jonas.moons@hu.nl

Sieuwert van Otterloo

Onderzoeker
Interesses: toegepaste AI, privacyrecht en auteursrecht, kwaliteit van software
sieuwert.vanotterloo@hu.nl

Sophie Rust

Docent-onderzoeker
Interesses: design ethics & AI, grafisch ontwerp
sophie.rust@hu.nl

Marc van der Peet

Projectleider
Master Human Centered AI
marc.vanderpeet@hu.nl

Meer informatie of vragen over het HU lectoraat Artificial Intelligence?

Neem contact op met lector Stefan Leijnen
(stefan.leijnen@hu.nl)