RAAIT

Responsible Applied Artificial InTelligence

Onderzoeksagenda

maart 2023

Colofon

Aan de onderzoeksagenda werkten mee
Maaike Harbers, Sophie Horsman, Daan Kolkman, Stefan Leijnen, Marieke Peeters, Saskia Robben, Tjerk Timan, Pascal Wiggers

Namens het RAAIT-consortium

- Hogeschool van Amsterdam, Hogeschool Utrecht, Hogeschool Rotterdam
- Kernpartners: Gemeente Amsterdam, Gemeente Haarlem, Gemeente Rotterdam, Provincie Utrecht, Provincie Zuid-Holland, CGI, Media Perspectives, Cupola XS
- Diverse praktijkpartners

Vormgeving
Cerys Hancock, Irene Schipper

www.raait.nl

Een samenwerkingsverband tussen

Hogeschool van Amsterdam · HOGESCHOOL UTRECHT · HOGESCHOOL ROTTERDAM

In co-creatie met diverse praktijkpartners

Inhoudsopgave

1 Inleiding 4

2 Wat is Responsible Applied AI? 8

3 Onderzoeksagenda 13

4 Aanpak: Learning Communities 21

5 Conclusie 23

Referenties 24

Over RAAIT

In dit document presenteren we de onderzoeksagenda voor het Responsible Applied Artificial InTelligence (RAAIT) programma. RAAIT draait om het ontwerpen, ontwikkelen, en inbedden van AI-oplossingen waarbij getracht wordt om negatieve implicaties van AI-toepassingen voor individuen, samenleving en planeet te minimaliseren en positieve implicaties te stimuleren. Het RAAIT-programma is een samenwerking tussen drie hogescholen, namelijk de Hogescholen van Amsterdam, Rotterdam en Utrecht. Het betreft een achtjarig onderzoeksprogramma met als doel het opzetten van een infrastructuur ten behoeve van excellent onderzoek op een actueel, maatschappelijk vraagstuk: Responsible Applied AI.

Als we iets hebben geleerd van de afgelopen jaren, is het dat de ontwikkelingen op het gebied van AI elkaar in razendsnel tempo opvolgen. Deze onderzoeksagenda is dan ook niet in beton gegoten, en kan met voortschrijdend inzicht gedurende de periode van 8 jaar eens of meerdere malen worden geactualiseerd, al naar gelang de vraagstukken die leven in het consortium en breder in de maatschappij. Het doel van de onderzoeksagenda is het bieden van perspectief en samenhang voor het onderzoek dat binnen de RAAIT samenwerking plaatsvindt.

1 Inleiding

Stel je voor: een burger ontvangt een geautomatiseerd besluit over het recht op een toeslag. In de brief staat uitgelegd op basis van welke informatie het besluit genomen is. Ook staat er een verwijzing in naar een algoritmeregister waarin de burger de beschrijving van het algoritme kan vinden, kan zien wie er betrokken zijn geweest bij de ontwikkeling, hoe het algoritme gemonitord wordt, wat er uit de laatste audits kwam, etc. De burger is het niet eens met het besluit en gaat verhaal halen bij de gemeente. Daar kan een medewerker inzien welke factoren doorslaggevend waren voor het besluit, het besluit kan vergelijken met andere besluiten, en bekijken hoe het besluit zou zijn uitgevallen als bepaalde factoren anders waren geweest. De medewerker deelt de inzichten met de burger, maar deze is het nog altijd niet eens met de beslissing van de gemeente, en kiest ervoor het besluit aan te vechten. De advocaten en de rechter kunnen het besluit met het algoritme reproduceren en dezelfde analyses uitvoeren als de persoon van de gemeente. Ook hebben ze toegang tot de volledige audit-verslagen, en kunnen ze mensen van de gemeente interviewen over de wijze waarop het algoritme tot stand is gekomen, hoe het wordt onderhouden, welke feedback-loops er bestaan, hoe de AI governance is ingericht, en hoe incidentmanagement is ingericht.

Dit is een voorbeeld van hoe Responsible Applied AI er in de praktijk uit zou kunnen zien. Dit is één van vele mogelijkheden; Responsible Applied AI kan verschillende vormen aannemen. Wat dit voorbeeld laat zien, is de complexiteit van verantwoorde inzet van AI in de praktijk: er komt nogal wat bij kijken, en er zijn veel mensen en belangen mee gemoeid. In dit stuk beschrijven we het veld van Responsible Applied AI, zetten we de belangrijkste uitdagingen uiteen, en laten we zien hoe wij de komende jaren zullen gaan werken om een aantal van deze uitdagingen het hoofd te bieden.

De praktijk van Responsible Applied AI

Waar AI ooit een weinig bekend academisch onderzoeksveld betrof, waar alleen computerwetenschappers zich over bogen, zien we dat het tegenwoordig aan de orde van de dag is en breed ingezet wordt voor allerlei toepassingen. Zo wordt AI ingezet om het aanbod van content en advertenties te personaliseren, auto's relatief zelfstandig te laten inparkeren, robots de oppervlakten van Mars te laten onderzoeken, en in de logistiek allerlei processen te optimaliseren.

Een belangrijk potentieel van AI is dat het ons kan helpen bij het oplossen van grote maatschappelijke uitdagingen, zoals de klimaatcrisis, de energiecrisis, het tekort aan mensen op de arbeidsmarkt en vergrijzingsproblematiek. Daarbij is het belangrijk om oog te hebben voor de risico's van het toepassen van AI. Zo zien we bijvoorbeeld dat vooroordelen en uitsluiting doorsijpelen in onze algoritmen, resulterend in systemische

discriminatie (Barocas & Selbst, 2016). Ook zijn de uitkomsten van AI vaak moeilijk uit te leggen, waardoor menselijk toezicht ingewikkeld of zelfs onmogelijk wordt (Kolkman, 2022). Daardoor varen mensen soms blind op (verkeerde) voorspellingen, met alle gevolgen van dien.

Binnen het RAAIT-programma kiezen we voor een focus op *Responsible* Applied AI: hoe zorgen we ervoor dat de ontwikkeling en inzet van AI-algoritmen, -modellen en -toepassingen, en de producten die daaruit voortvloeien, in lijn zijn met ethische principes? En hoe zorgen we ervoor dat we met AI een positieve impact op de wereld realiseren, terwijl we de negatieve uitwerkingen zoveel mogelijk beheersen en inperken? Om een praktijk van Responsible Applied AI te verwezenlijken, moeten een hoop onderliggende vragen worden gesteld en beantwoord. In dit document identificeren we belangrijke vragen rondom Responsible Applied AI, en brengen we structuur aan in deze vraagstukken door deze onder te brengen in een onderzoeksagenda.

Bij het identificeren van de vraagstukken die wij binnen het RAAIT-programma willen oppakken is het praktische element leidend: waar de academische wetenschap zich heeft toegelegd op de theoretisch-wiskundige of juist ethisch-filosofische vraagstukken rondom thema's als eerlijkheid, uitlegbaarheid, of verantwoordelijkheid, willen we binnen het RAAIT-programma wetenschappelijk onderzoek doen naar praktische handvatten, technieken, methoden en instrumenten die onderzoekers, studenten, professionals, en organisaties kunnen inzetten om op verantwoorde wijze met AI te werken. Waar de toegepaste wiskunde bijvoorbeeld verschillende metrieken voor eerlijkheid ontwikkelt, willen wij nadenken over hoe je tot een keuze tussen die metrieken komt en deze in de praktijk hanteert gegeven een concrete use case of concreet vraagstuk. En waar bedrijfskunde verschillende interventies voor cultuurverandering voorstelt, willen wij onderzoeken welke interventie werkt voor een organisatie die haar mensen mee wil nemen in het identificeren van ethische vraagstukken rondom de inzet van AI, of het ontwikkelen van beleid voor de inzet van AI.

Wij vinden die praktische insteek belangrijk, omdat echte maatschappelijke verandering niet (alleen) plaatsvindt in kennisontwikkeling, of in de invoering van nieuwe wet- en regelgeving, maar uiteindelijk tot stand komt door gedragsverandering daar waar het ertoe doet: in de praktische realiteit zoals we die aantreffen bij organisaties die met AI werken. Om die reden hebben we in het RAAIT-programma de handen ineengeslagen met praktijkpartners die de maatschappelijke verantwoordelijkheid en betrokkenheid voelen om de benodigde verandering in te zetten, en met elkaar te zoeken naar hoe die verandering in de praktijk vorm kan krijgen.

Over onszelf en het RAAIT-programma

De onderzoeksgroep die aan de basis staat van het RAAIT-programma bestaat uit onderzoekers in dienst bij de drie hogescholen (Amsterdam, Utrecht en Rotterdam). De onderzoekers zijn allemaal opgeleid in het hoger onderwijs en hebben verschillende (studie)achtergronden, zoals informatica, kunstmatige intelligentie, psychologie, filosofie, sociologie, wiskunde en communicatiewetenschappen. De groep kent een min of meer gelijke verdeling tussen mannen en vrouwen, momenteel respectievelijk ongeveer 45% en 55%. De onderzoekers zijn voor het overgrote deel Nederlands, wit en zonder fysieke beperking. Wij vertegenwoordigen hiermee een perspectief dat niet representatief is voor de Nederlandse samenleving – een perspectief met bias. We erkennen dat we blinde vlekken hebben en dat dat een beperking vormt bij het onderzoeken van Responsible Applied AI, waarin bias een belangrijk ethisch aandachtspunt vormt. Deze interne bias is een blijvend aandachtspunt en wij zullen ons inzetten om andere perspectieven dan het onze mee te nemen en te betrekken in onze activiteiten en dialogen.

Als programma dragen we uit dat we staan voor een *ethische* toepassing van AI. Dit geeft een risico op 'ethics-washing', waarbij Responsible Applied AI slechts in naam wordt gesteund zonder dat zich dat vertaalt in concrete acties (Green, 2021). Wij willen dan ook expliciet benoemen dat activiteiten of producten niet vanzelfsprekend verantwoord zijn enkel omdat zij onder de paraplu van het RAAIT-programma zijn ontwikkeld. We zien het als taak van het programma om het geheel aan opvattingen en invullingen rondom Responsible Applied AI te overzien, kritisch te beschouwen (inclusief onze eigen rol) en de dialoog erover te bevorderen. We geloven dat juist door onderlinge verschillen er in deze dialoog nieuwe inzichten kunnen ontstaan.

Structuur van dit document

In de rest van dit document zullen we de grote diversiteit en complexiteit van de hedendaagse praktijk te schetsen, waarbij we in Hoofdstuk 2 beginnen met het toelichten van een aantal belangrijke concepten. In Hoofdstuk 3 introduceren we het RAAIT-model dat ons helpt de verschillende vraagstukken te ordenen, en daarbinnen de uitdagingen en knelpunten te identificeren en te adresseren in onderzoek, onderwijs, en beroepspraktijk. We brengen deze samen in een coherente onderzoeksagenda, die laat zien waar we de komende tijd aan willen werken, maar die ook laat zien welke onderwerpen er in de periode daarna mogelijk nog in het verschiet liggen. De wijze waarop we invulling willen geven aan de onderzoeksagenda, samen met de praktijkpartners, is toegelicht in Hoofdstuk 4, waarin we meer vertellen over de samenwerkingsvorm zoals we die voor ons zien in de hybride leeromgevingen. Hoofdstuk 5 biedt een samenvatting en een oproep tot actie richting de hybride leeromgevingen om deze onderzoeksagenda verder te concretiseren en in te vullen voor de verschillende toepassingsdomeinen.

2 Wat is Responsible Applied AI?

In dit Hoofdstuk geven we antwoord op de vraag wat wij onder Responsible Applied AI verstaan. Daarbij houden we een algemene definitie van AI aan (zie kader), en gaan we in op de begrippen 'Applied' en 'Responsible'.

Definitie van AI

In dit programma hanteren we de definitie voor AI opgesteld door de Europese High-Level Expert Group on AI: "systemen die intelligent gedrag vertonen door hun omgeving te analyseren en – met enige graad van autonomie – actie te ondernemen om specifieke doelen te bereiken." In overeenstemming met de WRR (Prins, 2021) zien we AI als een *systeemtechnologie*; een technologie die continu aan verandering - en verbetering onderhevig is, en indirect nieuwe producten en toepassingen op tal van gebieden mogelijk maakt, net als de stoommachine, elektriciteit, de computer en het internet dat eerder deden.

Applied AI

Huidig onderzoek in AI is vaak fundamenteel en gericht op het verbeteren van de technologie. Succesvolle implementatie van AI gaat echter verder dan de ontwikkeling van fundamentele AI-technologie. *Applied AI* gaat om het toepassen van technieken uit de kunstmatige intelligentie (bijv. agent-based modeling, computer vision en natuurlijke taalverwerking) in een specifieke context (zorg, retail, etc.) waarbij rekening wordt gehouden met de gebruiker en andere belanghebbenden.

Hoewel AI op verschillende plekken al succesvol wordt toegepast, is er nog weinig systematische, herbruikbare kennis beschikbaar over hoe dit te bereiken. AI-technieken zijn ontwikkeld op basis van een beperkt aantal goed gedefinieerde vraagstukken en datasets om zo resultaten onderling te kunnen vergelijken. Het toepassen van AI bestaat er juist uit om bestaande technieken toe te passen in nieuwe en steeds weer andere situaties. Dit vraagt niet alleen om het selecteren van geschikte data, het kiezen van geschikte technieken en het bepalen van relevante evaluatiecriteria, maar ook om bijvoorbeeld het ethisch evalueren van een toepassing, het ontwerpen van deinteractiemogelijkheden en de user interface van een AI-toepassing, het trainen van personeel en het inrichten van een governance-structuur om bestaande AI-toepassingen te monitoren.

Tot voor kort werd het publieke debat en wetenschappelijke discours rondom kunstmatige intelligentie gedomineerd door extremen. Hierbij was voornamelijk aandacht voor extreme “uitschieters” waarbij overdreven optimistisch dan wel extreem pessimistisch naar de potentie van AI-technologie werd gekeken (Ziewitz, 2016).

In de afgelopen jaren zien we dat er door de media op een meer genuanceerde wijze gesproken wordt over de potentie en gevolgen van AI en dat de maatschappelijke discussie hierdoor realistischer wordt. Ondanks dat er minder met de “hype” wordt meegegaan, ontbreekt het bij journalisten echter vaak nog aan praktische kennis en blijven analyses daardoor oppervlakkig (Ouchchy et al., 2020). In dit programma zijn we nadrukkelijk geïnteresseerd in een brede set aan AI-toepassingen, die niet zozeer spectaculair falen of slagen, maar die de praktijk en de impact op de maatschappij op een accurate wijze weerspiegelen.

Afbeelding 1 *DRAMA is een samenwerkingsverband met Hogeschool Rotterdam, Hogeschool Utrecht, Hogeschool van Amsterdam, en meerdere mediapartijen. Het project richt zich op hoe we mediaorganisaties kunnen ondersteunen en begeleiden bij het inbedden van verantwoorde AI in hun organisaties. Dit project is een groot onderzoeksproject tussen de betrokken Hogescholen die ook deelnemen aan het RAAIT-programma. Voor meer informatie over het project, bezoek de website: https://www.responsibleappliedai.nl*

Responsible Applied AI

Wij verstaan onder *Responsible Applied AI* (RAAIT) 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 (Vinuesa et al., 2020; Floridi et al., 2020).

Er is binnen de wetenschap inmiddels een breed gedragen overeenstemming over het belang van Trustworthy AI, namelijk dat de praktijk rondom de inzet van AI aan bepaalde randvoorwaarden moet voldoen, maar er is minder eenduidigheid over wat die randvoorwaarden precies zijn. In de afgelopen jaren zijn er tientallen ethische richtlijnen rondom randvoorwaarden voor kunstmatige intelligentie gepubliceerd in het bedrijfsleven (bijv. Google, IBM, Intel), de overheid (bijv. Europese Commissie, UNESCO) en de academische wereld (bijv. Future of Life Institute). Uit enkele systematische studies (Hagendorff, 2020; Floridi & Cowls, 2022; Jobin et al., 2019) is naar voren gekomen dat sommige ethische waarden of principes in bijna alle richtlijnen voorkomen, waaronder transparantie, privacy, rechtvaardigheid, uitlegbaarheid en het tegengaan van discriminatie. Echter, andere waarden, zoals diversiteit onder de makers van AI, regelingen voor klokkenluiders of de ecologische impact van AI-systemen, komen minder vaak voor in richtlijnen, terwijl ze niet minder belangrijk zijn (Hagendorff, 2020; Whittaker et al., 2019).

AI for Good en Trustworthy AI kunnen onafhankelijk van elkaar gerealiseerd worden. Een AI for Good applicatie, bijvoorbeeld het verzorgen van goed onderwijs voor jonge meisjes op het platteland in Pakistan, kan alsnog resulteren in privacy-schending, of alleen toegankelijk zijn voor meisjes van welgestelde ouders. Afhankelijk van welke randvoorwaarden je aan AI-toepassingen stelt, zou je kunnen concluderen dat een dergelijk systeem niet Trustworthy is. Andersom kan een AI-toepassing volledig voldoen aan alle vereisten van de ethische richtlijnen van de Europese Commissie, en toch als doel hebben om de positionering van een olieboorplatform op de Noordpool automatisch te corrigeren voor het weer en daarmee niet vallen onder AI for Good.

Als RAAIT-programma ondersteunen we de ambitie om AI in te zetten voor maatschappelijke uitdagingen. Tegelijk betekent dat niet dat we ons enkel richten op AI-toepassingen voor maatschappelijke uitdagingen. Veel AI-systemen worden niet per definitie met een 'AI for good' insteek ontwikkeld omdat bijvoorbeeld ook commerciële en concurrerende belangen meespelen. Veel toepassingen zijn bovendien misschien niet zuiver "AI for Good" of zuiver "AI for Bad", maar hebben zowel positieve als negatieve consequenties. Omdat AI in de praktijk in het RAAIT-programma centraal staat, omarmen wij het brede spectrum van AI-oplossingen, van oplossingen die zich onder AI for Good scharen, tot bovengenoemde oplossingen in een bedrijfscontext. Onze insteek is wel dat het voor iedere ontwikkeling en inzet van AI – ongeacht het doel van de AI-toepassing – van belang is om tegemoet te komen aan een verantwoorde inzet van AI (Trustworthy AI), zodat mogelijke negatieve (bij)gevolgen voor de mens, maatschappij en planeet zoveel mogelijk worden beperkt.

In het RAAIT programma richten wij ons niet op het implementeren van ethische principes of besluitvorming in de AI zelf, waarbij een AI-systeem zelf ethisch redeneervermogen ontwikkelt en onderscheid kan maken tussen "goed" en "kwaad" (*Machine Ethics*). Wij richten ons niet op het inbouwen van ethische redeneer- en besluitvormingscapaciteiten in een AI-systeem, omdat wij denken dat vanwege de pluriformiteit en veranderlijkheid van ethiek het automatiseren van ethisch redeneervermogen in de praktijk geen toereikende oplossing is.

Ook willen we tot slot benadrukken dat Responsible Applied AI wat ons betreft verder gaat dan 'compliance', het naleven van geldende wet- en regelgeving. Er wordt momenteel door Europese lidstaten gewerkt aan nieuwe wet- en regelgeving voor AI (zie kader: Wet- en regelgeving AI). Wij denken dat er zowel in de huidige als toekomstige wet- en regelgeving altijd ruimte zal blijven bestaan om binnen hetgeen wat wettelijk aanvaardbaar is, keuzes te maken die moreel onaanvaardbaar of onwenselijk zijn (Floridi, 2018). Daarom willen we met dit programma bedrijven en professionals aanmoedigen en helpen om verantwoorde keuzes te maken, die (zo veel mogelijk) in lijn liggen met eerdergenoemde ethische richtlijnen en waarden.

Wet- en regelgeving AI

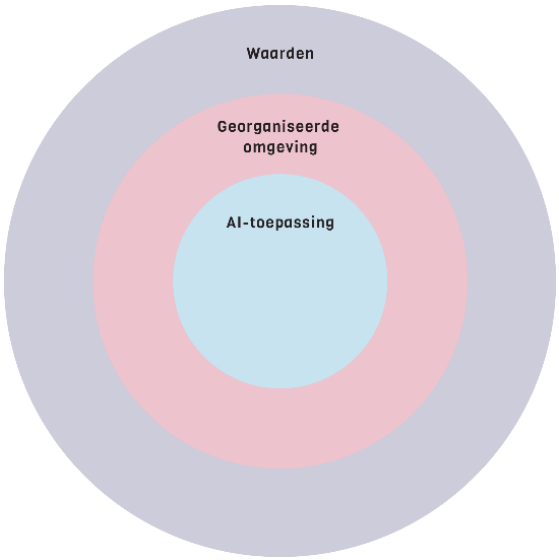
Er bestaat op dit moment in Nederland geen wetgeving specifiek gericht op AI-technologie. Echter, er bestaat wel wetgeving die AI raakt, aangezien AI in producten, diensten en processen zit die gereguleerd zijn. Voorbeelden zijn de AVG en de consumentenbeschermingsregelgeving (Peters, 2020). Momenteel wordt door verschillende Europese lidstaten gewerkt aan de AI Act. Dit is een Europees wetsvoorstel dat in april 2021 aan de Europese Commissie is geïntroduceerd. Dit voorstel neemt een sector-overschrijdende en op risico gebaseerde benadering die van toepassing is op alle aanbieders en gebruikers van AI-systemen op de Europese markt. Toepassingen van AI die als het meest schadelijk worden beschouwd, zullen worden verboden, terwijl een gedefinieerde lijst van "risicovolle" AI-systemen aan strikte eisen moet voldoen. Naar verwachting zal deze wetgeving ten vroegste vanaf 2024 van kracht gaan.

3 Onderzoeksagenda

Binnen het RAAIT-programma bieden we ruimte voor verschillende activiteiten en toepassingen. De overkoepelende onderzoeksvraag die daarbij centraal staat is: “Hoe begeleid je professionals, bedrijven en organisaties in de praktijk *die op dit moment weinig houvast hebben* naar een dagelijkse praktijk die in lijn is met *Responsible Applied AI*?” We zijn met name geïnteresseerd in wat ervoor nodig is om bij een “dagelijkse praktijk van Responsible Applied AI” uit te komen, en in welke mate dit verschilt per sector, organisatie en doelgroep. Om deze centrale vraag te beantwoorden hebben we vier ‘grand challenges’ geïdentificeerd. Voordat wij deze ‘grand challenges’ en bijbehorende onderzoeksvragen beschrijven, introduceren we het RAAIT-model. Het RAAIT-model biedt een kapstok om de diversiteit en complexiteit van Responsible Applied AI te beschrijven.

RAAIT-model

Het ontwikkelen en toepassen van AI is een sterk multidisciplinair en gelaagd proces. Daarom kiezen wij ervoor om uit te zoomen en een globaal model te schetsen, waarin verschillende belanghebbenden zich kunnen herkennen. Het RAAIT-model (Figuur 1) bestaat uit drie abstractieniveaus om naar verantwoorde AI te kijken: (1) het niveau van **waarden** die belangrijk zijn, (2) de **georganiseerde omgeving** (de context en omgeving) die van belang zijn voor een verantwoorde praktijk, en (3) de levenscyclus van de **AI-toepassing**. Dit RAAIT-model dient als overkoepelend raamwerk om met verschillende gesprekspartners (organisaties, professionals, onderzoekers) te communiceren over de transitie naar een verantwoorde AI-praktijk. De verschillende abstractieniveaus laten zien op welke manieren er invloed uitgeoefend kan worden om het geheel meer verantwoord te maken. Hieronder zullen we elk van de abstractieniveaus in meer detail beschrijven.



***Figuur 1**: RAAIT-model*

***Figuur 2**: RAAIT-model met drie abstractieniveaus: waarden, de georganiseerde omgeving en de AI-toepassing*

De abstractieniveaus van het RAAIT-model

Waarden: De buitenste cirkel beschrijft de waarden die belangrijk zijn voor een RAAIT-praktijk. We onderschrijven daarbij eerdergenoemde ethische richtlijnen, waarin waarden als privacy, autonomie, eerlijkheid, en inclusiviteit centraal staan. Deze lijst is voor ons niet uitputtend, aangezien we ook aandacht hebben voor stakeholders die (indirect) geraakt worden door AI-toepassingen en de ecologische impact van AI. Relevante thema's zijn onder andere: mensen die buitengesloten worden, misstanden rondom *crowdworkers* en vervuiling en duurzaamheid. Tenslotte vinden wij het belangrijk dat de prioritering van waarden ook afhangt van de visie van de organisatie en het specifieke project of product.

Georganiseerde omgeving: De middelste cirkel gaat over de omgeving en de context van de RAAIT-praktijk. Hier hebben we het over de invloedsfeer van de toepassingsgebieden (bedrijfssectoren) en de organisaties (of instituten) daarbinnen: wat kan er worden georganiseerd om een verantwoorde AI-praktijk te realiseren? Gedacht kan worden aan het beleid van een organisatie (o.a. financieel, governance-structuur, compliance, inkoopvoorwaarden), de mensen binnen een organisatie (o.a. training en opleiding, werving, cultuur, inclusiviteit en diversiteit), en ook technische/praktische zaken die een organisatie kan regelen, bijvoorbeeld rondom datamanagement, beschikbaarheid van tooling en infrastructuur.

AI-toepassing: De binnenste cirkel gaat over de levenscyclus van de AI-toepassing: van het ontwerp tot de evaluatie en het onderhoud van het model. We richten ons hierbij grotendeels op het team van ontwikkelaars dat aan een AI-oplossing werkt: wat doen zij of kunnen zij doen in de alledaagse praktijk om de AI-applicatie ten aanzien van potentiële gebruikers, de maatschappij en de planeet meer verantwoord te maken? Hierbij wordt onder meer gekeken naar de AI-technologie die in het product gebruikt wordt: aan welke

eisen moet een specifieke oplossing voldoen, en hoe zorgen we ervoor dat de oplossing vervolgens ook aan die vereisten blijft voldoen?

Grand challenges

We hebben in het RAAIT-programma vier “grand challenges” geformuleerd: rondom de ontwerppraktijk, de ontwikkelpraktijk, de organisatiepraktijk, en de integratie: het creëren van een geïntegreerde RAAIT-methodologie. Deze vier grand challenges zullen hieronder verder worden toegelicht aan de hand van het RAAIT-model.

Grand challenge 1: Ontwerppraktijk

Welke ontwerppraktijk past er bij het ontwerpen van Responsible Applied AI?

De complexiteit van AI, de onduidelijkheid over de mogelijkheden van AI en de nieuwe ethische vragen die AI-toepassingen oproepen maken dat bestaande ontwerpmethoden en -kennis tekortschieten (Yang et al., 2020). De mogelijkheid om AI in te zetten als ‘ontwerpmateriaal’ bij het creëren van technologische toepassingen, brengt dus nieuwe vragen met zich mee voor het ontwerpvakgebied (Holmquist, 2017).

Relaterend aan het RAAIT-model zijn er verschillende vragen te identificeren. Bij de eerste cirkel, ***waarden***, zijn relevante vragen voor de ontwerppraktijk bijvoorbeeld: Wiens waarden zijn van belang in een gegeven RAAIT-ontwerpvraagstuk? Welke waarden zijn dat? En wie bepaalt dat? Vragen bij de tweede cirkel, ***organisatie***, zijn: Welke processen en checks worden in een organisatie gevolgd om te zorgen voor een Responsible Applied AI-ontwerppraktijk? Wanneer in een ontwerpproces wordt bijvoorbeeld aandacht besteed aan ethische vraagstukken? Welke stakeholders worden daarbij betrokken? Wie heeft welke verantwoordelijkheid? Vragen bij de derde cirkel, ***AI-toepassing***, zijn bijvoorbeeld: Welke impact heeft een AI-toepassing op (waarden van) verschillende stakeholders (inclusief non-humans), hoe weeg je die impact, hoe kun je in het ontwerpproces keuzes te maken die de impact op een positieve manier beïnvloeden?

Vanuit het RAAIT-programma zien we in het bijzonder een rol voor ontwerpers in het op een zinvolle manier bij elkaar brengen van verschillende perspectieven die nodig zijn om tot een succesvolle toepassing van AI te komen (Harbers & Overdiek, 2022). Daarbij valt te denken aan het combineren van technische en juridische AI-expertise (waar al relatief veel aandacht voor is) met ethische en sociale invalshoeken, kennis over het toepassingsdomein, en kennis en ervaringen van gebruikers en andere stakeholders die geraakt worden door de AI-toepassing. Al deze verschillende stakeholders en perspectieven gaan uit van verschillende kennis, denkkaders en taalgebruik. Dat maakt communicatie en samenwerking soms lastig. Een bijkomende uitdaging bij het

samenbrengen van verschillende perspectieven is om ervoor te zorgen dat die perspectieven niet alleen aanwezig zijn, maar dat ze ook daadwerkelijk kunnen meepraten en -beslissen. Dit vereist bewustzijn van en aandacht voor machtsverhoudingen binnen het ontwerpproces, en nieuwsgierigheid naar en respect voor elkaars perspectief, expertise, en vakgebied. Zonder expliciete aandacht zullen dominante perspectieven al snel de overhand krijgen, bijvoorbeeld wanneer goed meetbare criteria zwaarder meegewogen worden dan lastig meetbare (maar niet minder belangrijke) criteria of wanneer de mening van experts of leidinggevenden zwaarder wegen dan die van gebruikers of burgers. We zien co-creatie, met de inzet van instrumenten zoals 'probes' (designmethoden om inspirerende ideeën op te halen) en 'boundary objects', als een veelbelovende manier om verschillende zienswijzen op een zinvolle manier bij elkaar te brengen (Harbers & Overdiek, 2022).

***Afbeelding 2**: Binnen RAAIT vallen niet enkel traditionele (onderzoeks)projecten, maar zoeken we actief de samenwerking op met kunstenaars en ontwerpers. Het boundary object van onderzoeker Tiwánee van der Horst en lector Anja Overdiek biedt een vernieuwende manier om het gesprek te voeren over de ethische kwesties rondom AI. Elke deelnemer wordt zich tijdens een co-creatie sessie bewust van welke machtsdynamieken van invloed zijn op AI. Welke partij trekt het hardste aan de touwtjes? Is het de overheid, Big Tech, onderwijsinstellingen of de gebruikers?*

Grand challenge 2: Ontwikkelpraktijk

Hoe zorg je ervoor dat ontwikkelaars geholpen en gestimuleerd worden om verantwoord te werken - en hoe evalueer je dat?

De ontwikkeling rondom AI zit in een stroomversnelling. Complexe AI-modellen en technieken worden steeds beter inzetbaar en het gebruik ervan is wijdverbreid. Het wordt echter ook steeds duidelijker dat er onbedoeld allerlei normatieve keuzes in de uiteindelijke AI-oplossing sluipen (bijvoorbeeld tijdens het selecteren van data, het trainen van het model, etc.).

Een andere ontwikkeling die bijdraagt aan een onverantwoorde AI-praktijk is dat het de afgelopen jaren steeds makkelijker is geworden om via het internet te experimenteren met 'off-the-shelf' AI-modellen. Deze laagdrempeligheid leidt ertoe dat ook programmeurs met beperkte achtergrondkennis op het gebied van wiskunde, statistiek en computerwetenschappen aan de slag kunnen met complexe technieken. Het gevolg is dat zij mogelijk niet beschikken over de noodzakelijke kennis om eventuele risico's goed te overzien en te mitigeren. Zo zouden dit soort 'copy-and-paste developers' bijvoorbeeld code kunnen ontwikkelen die niet aan bepaalde standaarden (validatie, wetenschappelijke methode, etc.) voldoet (Knight, 2022). Een ander risico is dat zij AI-modellen inzetten in een context waarvoor deze niet geschikt of gewenst zijn.

De alsmaar meer geavanceerde (technische) ontwikkelingen bieden ook kansen om de praktijk juist meer verantwoord te maken. Zo zijn er bijvoorbeeld de FAIR principes voor data ontwikkeld, het hanteren waarvan zorgt voor beter gebruik van data. Een ander voorbeeld is de opkomst van federated learning, welke kan zorgen voor betrouwbaardere modellen waarbij gevoelige data op een veilige plek kan blijven.

Relaterend aan het RAAIT-model kunnen er ten aanzien van de ontwikkelpraktijk verschillende vragen worden gesteld. Als het gaat om ***waarden***, is een cruciale vraag hoe bepaalde waarden zich vertalen naar de praktijk van de AI-ontwikkeling en het daaruit voortvloeiende AI-product (Leijnen et al., 2020; Morley et al., 2021). Bijvoorbeeld: wat betekent fairness voor de ontwikkeling van een specifieke AI-applicatie en welke onevenredige uitkomsten voor verschillende doelgroepen zijn acceptabel? Wat doe je als bepaalde waarden, zoals het recht op privacy en uitlegbaarheid met elkaar in conflict zijn? Als het gaat om de ***organisatie*** zijn er ook enkele vragen te stellen, zoals: Hoe richt je ontwikkelteams zo in dat er een gedeelde verantwoordelijkheid is voor het nemen van ethische beslissingen? Hoe monitor je dat? Hoe stuur je ontwikkelteams aan en wat voor handleidingen/richtlijnen reik je aan? Zijn er wellicht nieuwe rollen nodig binnen een organisatie? Ten slotte, als het gaat om de ***AI-toepassing*** zijn de volgende vragen relevant: Welke evaluatiemetrieken hanteer je? In het bijzonder: hoe evalueren we de prestatie van

AI-systemen niet enkel op *accuracy*, maar ook op bijvoorbeeld uitlegbaarheid en eerlijkheid? (Piersma, 2022). Hoe ga je om met state-of-the-art voorgetrainde modellen, die in hoog tempo gepubliceerd worden, maar niet altijd transparant zijn over de totstandkoming en onderliggende data? Tot slot: welke *design patterns* kunnen ervoor zorgen dat de (software-)kwaliteit van AI-oplossingen toeneemt, waarbij steeds terugkerende patronen in de AI-ontwikkeling op een zorgvuldige en transparante manier gebeuren (McAran, 2021)?

Grand challenge 3: Omgevingspraktijk

Hoe richt je de georganiseerde omgeving in op een manier die een verantwoorde AI-praktijk ondersteunt?

Het bedenken, ontwerpen, ontwikkelen, uitrollen, en opschalen van AI-gedreven innovaties vindt niet plaats in een vacuüm, maar binnen een -georganiseerde- omgeving, oftewel organisatie. Zo'n organisatie voorziet, afhankelijk van haar omvang en volwassenheid, in ondersteunende zaken zoals diensten, producten, processen, platformen, en - niet te vergeten - menselijke interacties en samenwerkingen. Wanneer we het hebben over een verantwoorde AI-praktijk, hebben we het al snel over de ontwerppraktijk en de ontwikkelpraktijk. En dat is terecht: een verantwoorde AI-praktijk kan en moet kiem vatten in die alledaagse praktijk van de AI-professional die op de werkvloer tot nieuwe oplossingen komt. Maar om echt wortel te kunnen schieten, helpt het wanneer de georganiseerde omgeving ondertussen ook de juiste voedingsbodem schept: training en opleiding, tijd en ruimte voor ethische discussies, een pool van experts die kunnen meedenken wanneer een team op complexe vraagstukken stuit, jong talent met geactualiseerde kennis over nieuwe methoden en technieken, cultuurverandering waarin nieuwe (ethische) waarden verankerd liggen, tooling voor specifieke operationele vereisten, zoals reproduceerbaarheid, privacy, data-kwaliteit of transparantie, en technische infrastructuur die het mogelijk maakt om monitoring en signalering op ethische randvoorwaarden, zoals fairness, bias, model drift of data drift, deels te automatiseren (Krijger et al., 2022; Smit & Eybers, 2022).

Ook de wijze waarop de AI-toepassing wordt ingebed in de gebruikscontext, en hoe menselijke terugkoppeling op een laagdrempelige manier kan worden gestimuleerd, zijn belangrijke vraagstukken binnen de grand challenge van de omgevingspraktijk (Friedman & Hendry, 2019): hoe stellen we belanghebbenden (denk aan eindgebruikers, consumenten, data subjecten, IT/beheer medewerkers, en/of toezichthouders) in staat om gedurende het gebruik van het systeem betrokken te blijven bij de iteratieve verbetering van dat systeem? Al deze groepen belanghebbenden kunnen waardevolle informatie terugkoppelen over potentiële mogelijkheden en functie-uitbreidingen, maar ook over geïdentificeerde risico's, tekortkomingen en beperkingen. Hoe zorgen we ervoor

dat juist zij waakzaam blijven, en over hun bevindingen in gesprek blijven met de ontwerpers en ontwikkelaars van het product? Welke kennis hebben zij nodig om als een goed geïnformeerde gesprekspartner deel te nemen aan het gesprek over de zin en onzin van een AI-oplossing (Fortuna & Gorbarniuk, 2022)?

En ten slotte bestaat er nog het vraagstuk over de inrichting van AI governance (Mannes, 2020; Chhillar & Aguilera, 2022): welke rollen, processen, afspraken, informatie-uitwisseling, en documentatie is er nodig om vanuit een organisatorisch perspectief een verantwoorde praktijk rondom AI te faciliteren? Wie neemt de leiding in het opleiden en trainen van de medewerkers, het vastleggen van, en controleren op, het nakomen van afspraken rondom technische en functionele vereisten, of het delen van richtlijnen, best practices en lessons learned?

Als we de vraagstukken rondom de organisatiepraktijk relateren aan het RAAIT-model zien we verschillende vragen ontstaan op de drie abstractieniveaus. Bij de eerste cirkel, ***waarden***, kunnen we onszelf de vraag stellen: Welke bedrijfs- of organisatiewaarden zijn van belang om in ogenschouw te nemen bij het ontwikkelen (of inkopen) van AI-oplossingen? Hoe kan een organisatie omgaan met eventuele conflicterende belangen en/of waarden, zoals commerciële belangen, maatschappelijke/ecologische belangen, maar ook belangen van verschillende doelgroepen (zoals toezichthouders, medewerkers, consumenten, en data subjecten)? Vragen bij de tweede cirkel, ***organisatie***, gaan bijvoorbeeld over de wijze waarop de organisatie stapje voor stapje kan bewegen richting een meer verantwoorde AI-praktijk, maar ook hoe de bedrijfs- en organisatiewaarden kunnen worden uitgedragen in de cultuur van de organisatie, en hoe daarmee rekening kan worden gehouden in het wervings-, selectie- en trainingsproces. Vragen bij de derde cirkel, ***AI-toepassing***, zijn gericht op het organiseren van structurele en gestandaardiseerde organisatie-brede ontwerp- en ontwikkelwerkwijzen, processen, tools en infrastructuur die het zo eenvoudig en eenduidig mogelijk maken voor alle medewerkers om te werken in lijn met de waarden en de cultuur van de organisatie, en tot AI-producten te komen die daarmee in overeenstemming zijn.

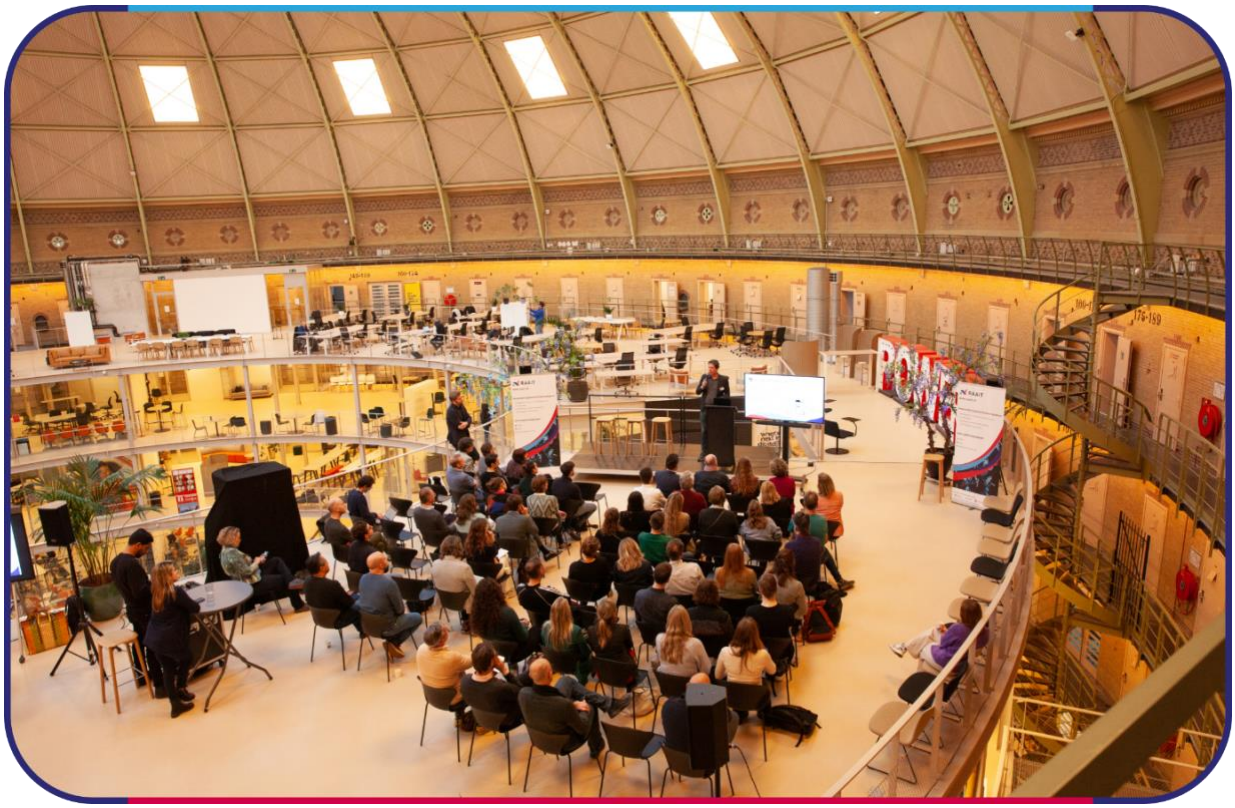
Grand Challenge 4:

De ontwikkeling van een overkoepelend raamwerk in de vorm van een RAAIT-methodologie waarin de drie onderzoekslijnen geïntegreerd worden.

De drie tot nu toe beschreven grand challenges staat niet los van elkaar. De vierde grand challenge gaat dan ook over het ontwikkelen van een overkoepelend raamwerk dat de inzichten uit de drie onderzoekslijnen integreert. Het doel binnen deze grand challenge is om een RAAIT-methodologie te ontwikkelen. Deze methodologie bevat een overkoepelende visie, die wordt geladen met praktische tools, methodes, prototypes en

werkwijzen die elkaar versterken en breed inzetbaar zijn. Aanvullend op deze praktische handvatten bevat de methodologie ook aanbevelingen en richtlijnen voor wanneer en onder welke omstandigheden de verschillende tools en methodes inzetbaar zijn.

Het RAAIT-model vormt een eerste startpunt voor het ontwikkelen van een overkoepelende visie. Om de RAAIT-methodologie verder te ontwikkelen is kennisopbouw en -deling nodig die wordt ontwikkeld in co-creatie met onderzoekers, de beroepspraktijk en andere belanghebbenden. Door het bieden van een gemeenschappelijk vertrekpunt en praktische handvatten voor een RAAIT-praktijk, ondersteunt de RAAIT-methodologie professionals om AI structureel op verantwoorde wijze toe te passen.



***Afbeelding 3**: Op 1 februari jl. heeft er een partnerdag van RAAIT plaatsgevonden in de Cupola XS – de oude koepelgevangenis - te Haarlem. Deze dag stond in het teken van uitwisseling tussen de partners van RAAIT en andere geïnteresseerden. Er zal jaarlijks een groot RAAIT evenement worden georganiseerd, waardoor partners elkaar op de hoogte kunnen houden van de vorderingen in het programma en kennisuitwisseling binnen- en tussen de learning communities kan worden gefaciliteerd.*

4 Aanpak: Learning Communities

Het onderzoek naar bovenstaande uitdagingen vindt plaats in Learning Communities. Een Learning Community (ook wel “hybride leeromgeving”) biedt een omgeving waarin bedrijven, overheden, onderzoekers, studenten en eindgebruikers met elkaar samenwerken aan een bepaald vraagstuk. Deze aanpak is gekozen omdat de uitdagingen over de ontwerp-, ontwikkel- en organisatiepraktijk van de verantwoorde AI-toepassingen contextafhankelijk zijn. Dat wil zeggen dat er verschillen zijn tussen AI-toepassingen in wat de belangrijkste risico’s, uitdagingen en oplossingen voor die specifieke toepassing zijn. Om recht te doen aan de contextspecifieke kenmerken van een AI-toepassing, is het van belang dat bovenstaande onderzoeksvragen onderzocht worden in - en met - de praktijk. In een Learning Community trekken de verschillende partijen, waaronder praktijkpartners, actief samen op om in co-creatie innovatieve oplossingen te ontwerpen en te toetsen. Hiermee vormen Learning Communities een omgeving die leidt tot nieuwe inzichten en kennis, die gegrond zijn in de praktijk en daarmee dus recht doen aan de contextafhankelijkheid van verantwoorde AI-toepassingen.

In het RAAIT-programma zal elk van de drie hogescholen in de eerste twee à drie jaar zich richten op het opzetten van een Learning Community voor een specifiek toepassingsgebied. Er is gekozen voor de toepassingsgebieden Retail (HvA), Zakelijke dienstverlening (HR) en Media (HU). Daarna zullen Learning Communities voor andere toepassingsgebieden volgen. Aan elke Learning Community zijn praktijkpartners verbonden uit het betreffende toepassingsgebied. Samen met deze praktijkpartners zullen in co-creatie met overheden, onderzoekers, studenten en eindgebruikers praktische tools, instrumenten, prototypes, leer- en professionaliseringstrajecten worden ontwikkeld. Vragen over de verantwoorde toepassing van AI die leven bij de verbonden praktijkpartners zijn leidend voor de activiteiten in een Learning Community. De Learning Communities stellen de (praktijk)deelnemers in staat om zich te (blijven) ontwikkelen tot de koplopers Responsible Applied AI in een specifiek toepassingsgebied. De Learning Communities moeten hierbij gezien worden als “coalitions of the willing”, waarbij een groep vooruitlopende bedrijven kan helpen om ook de achterblijvers op sleeptouw te nemen.

Hoewel de exacte invulling per Learning Community zal verschillen, geldt dat het uitgangspunt van een Learning Community altijd samenwerking tussen onderzoek, onderwijs en praktijk is; die drie elementen streven we na in elk partnerschap. Daarbij wordt voortdurend gestreefd naar multidisciplinariteit en gelijkwaardigheid in de samenwerking. Ook zal in elk van de Learning Communities specifieke aandacht uitgaan naar de maatschappelijke impact van de AI-toepassingen, rekening houdend met ethische vraagstukken (Schipper et al., 2022). Voorbeelden van activiteiten zijn onder

andere het inzetten van stagiairs, het inzetten van werkstudenten, het organiseren van workshops, debatten, inhoudelijke dineravonden of andere evenementen, het ontwikkelen van e-learning modules (zoals de Teach de Teacher leergang ontwikkeld door de NLAIC), het doen van gezamenlijk onderzoek (met afstudeerders, professional doctorates en promovendi).

Parallel aan de domein-specifieke activiteiten binnen de Learning Communities, zal er afstemming plaatsvinden tussen de verschillende Learning Communities met als doel om van elkaars ervaringen te leren en om domein-overstijgende kennis op te doen. Een workshop of prototype ontwikkeld in de ene Learning Community zou bijvoorbeeld kunnen worden ingezet en getoetst in een andere Learning Community. De ervaringen die worden opgedaan met dergelijke experimenten geven inzicht in welke aspecten van een oplossing of aanpak domein-specifiek zijn, en welke juist domein-overstijgend. De onderzoekers verbonden aan de verschillende Learning Communities zullen het voortouw nemen in de afstemming tussen de Learning Communities. Deze domein-overstijgende activiteiten dragen bij aan de vierde Grand Challenge, het ontwikkelen van een overkoepelende RAAIT-methodologie.

5 Conclusie

Met de beschreven uitdagingen en aanpak ambiëren we over acht jaar een RAAIT methodologie te hebben ontwikkeld die (toekomstige) professionals, bedrijven en organisaties op een werkbare en zinvolle wijze ondersteunt in hun RAAIT praktijk. Deze methodologie zal bestaan uit enerzijds een overkoepelende RAAIT visie en anderzijds een verzameling aan methodes, instrumenten, prototypes en werkwijzen waaruit al naar gelang wat in een bepaalde situatie wenselijk is geput kan worden. Deze methodologie wordt ontwikkeld met onze praktijkpartners aan de hand van experimenten en pilot-onderzoek naar Responsible AI binnen de Learning Communities. Gezamenlijk beoogd deze RAAIT methodologie (toekomstige) professionals, bedrijven en organisaties in staat te stellen om op verantwoorde wijze AI-toepassingen te ontwerpen, ontwikkelen en in te zetten.

Dit document beschrijft onze huidige kijk op Responsible Applied AI en het RAAIT-programma, de vraagstukken waarmee we aan de slag willen gaan in het RAAIT-programma en de wijze waarop we van plan zijn dat te gaan doen. Hiermee vormt deze onderzoeksagenda het uitgangspunt van de activiteiten die we de komende jaren zullen ontplooien. De onderzoeksagenda zoals gepresenteerd in dit document vormt een momentopname van onze huidige kijk op de zaken. De vragen en uitdagingen waar we voor staan, zullen handen en voeten krijgen binnen de Learning Communities waarin we met bedrijven, overheden, onderzoekers en studenten aan de slag gaan om de praktijk van Responsible Applied AI concreet te maken en verder te ontwikkelen. We zijn ervan overtuigd dat deze ervaringen, in combinatie met ontwikkelingen in het vakgebied, onze kijk de komende jaren zal doen aanpassen, verdiepen en verrijken.

Referenties

Mannes, A. (2020). Governance, risk, and artificial intelligence. Ai Magazine, 41(1), 61-69.
Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California law review, 671-732.

Chhillar, D., & Aguilera, R. V. (2022). An Eye for Artificial Intelligence: Insights Into the Governance of Artificial Intelligence and Vision for Future Research. Business & Society, 61(5), 1197-1241.

Europese Commissie (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L. (2018). Soft ethics and the governance of the digital. Philosophy & Technology, 31, 1-8.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, 535-545.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. Science and Engineering Ethics, 26(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Fortuna, P., & Gorbaniuk, O. (2022). What is behind the buzzword for experts and laymen: Representation of “artificial intelligence” in the IT-professionals’ and non-professionals’ minds. Europe’s Journal of Psychology, 18(2), 207-218.

Friedman, B., & Hendry, D. G. (2019). Value sensitive design: Shaping technology with moral imagination. Mit Press.

Green, B. (2021). The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice. Journal of Social Computing, 2(3), 209–225. https://doi.org/10.23919/JSC.2021.0018

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Harbers, M., & Overdiek, A. (2022). Towards a Living Lab for Responsible Applied AI. Proceedings of Design Research Society Conference, pp. 28-33.

Holmquist, L. E. (2017). Intelligence on tap: artificial intelligence as a new design material. interactions, 24(4).

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: The global landscape of ethics guidelines.

Knight, W. (2022). Sloppy Use of Machine Learning Is Causing a 'Reproducibility Crisis' in Science. Wired Business. https://www.wired.com/story/machine-learning-reproducibility-crisis/

Kolkman, D. (2022). The (in) credibility of algorithmic models to non-experts. Information, Communication & Society, 25(1), 93-109.

Krijger, J., Thuis, T., de Ruiter, M., Ligthart, E., & Broekman, I. (2022). The AI ethics maturity model: a holistic approach to advancing ethical data science in organizations. AI and Ethics, 1-13.

Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., & Ossewaarde, R. (2020). An agile framework for trustworthy AI. In NeHuAI@ ECAI (pp. 75-78).

McAran, D. (2021). Privacy, Ethics, Trust, and UX Challenges as Reflected in Google's People and AI Guidebook. In HCI in Business, Government and Organizations: 8th International Conference, HCIBGO 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings (pp. 588-599). Cham: Springer International Publishing.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Ethics, governance, and policies in artificial intelligence, 153-183.

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. AI & SOCIETY, 35, 927-936.

Piersma, N. (2022). System error, please restart: Hoe we verantwoorde IT-systemen kunnen bouwen.

Peters, M. (2020). Zeven acties voor verantwoord innoveren met AI. Bericht aan het parlement.

Prins, C. (2021). Opgave AI: De nieuwe systeemtechnologie. Wetenschappelijke Raad voor het Regeringsbeleid.

Schipper, T. , M. Vos, C. Wallner (2022). Position Paper Learning Comminities. Online beschikbaar: https://www.health-holland.com/sites/default/files/downloads/Landelijk%20Position%20Paper%20Learning%20 Communities_DEF_31-3-2022_0.pdf

Smit, D., & Eybers, S. (2022). Towards a socio-specific artificial intelligence adoption framework. In Proceedings of 43rd Conference of the South African Insti (Vol. 85, pp. 270-282).

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 233. https://doi.org/10.1038/s41467-019-14108-y

Whittaker, M., Alper, M., College, O., Kaziunas, L., & Morris, M. R. (2019). Disability, Bias, and AI. AI Now Institute, 8.

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-13).

Ziewitz, M. (2016). Governing algorithms: Myth, mess, and methods. Science, Technology, & Human Values, 41(1), 3-16.