MASTER'S THESIS

Een risico framework voor Responsible AI

Grimbergen, S.

Award date:
2022

Een risico framework voor Responsible AI

A risk framework for Responsible AI

Opleiding:	Open Universiteit, faculteit Bètawetenschappen
	Masteropleiding Business Process Management & IT
Programma:	Open University of The Netherlands, faculty of Science
	Master of Science Business Process Management & IT
Cursus:	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student:	Stefan Grimbergen
Identiteitsnummer:	
Datum:	27 februari 2022
Afstudeerbegeleider	Laury Bollen
Meelezer	Tim Huygh
Derde beoordelaar	
Versie nummer:	1
Status:	definitief

Abstract

The field of AI has proven that organization can benefit from using AI, but that there are risks involved as well. It's necessary to address these risks in order to implement AI in a responsible way. In this research, a framework is proposed that can be used to deal with these risks. The main research question of this paper is 'How should a framework be developed that can be used to deal with risks in the context of Responsible AI?'

This research paper describes the way in which the RAI Risk Framework is developed and provides a visual presentation of this framework. Aligned with the principles of RAI, the focus of the framework is 'a responsible implementation'. The five principles which are used in the framework are: ethics, responsibility, accountability, privacy & security and explainability.

Sleutelbegrippen

Responsible AI, Risk framework, AI Governance, RAI principles

Samenvatting

Dit onderzoek is gedaan binnen het vakgebied van Responsible AI (RAI). Omdat een algemeen geldende definitie van RAI ontbreekt in de literatuur, is in dit onderzoek uitgegaan van een eigen definitie die luidt: ‘Een verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een ‘governance framework’.

AI heeft bewezen grote voordelen te kunnen bieden aan organisaties, maar ook risico’s met zich mee te brengen. Om AI op een verantwoordelijke manier te kunnen implementeren dient omgegaan te worden met de risico’s. Dit onderzoek richt zich op het omgaan met de risico’s van RAI door het ontwikkelen van een framework. De hoofdonderzoeksvraag die is opgesteld is ‘Hoe kan een framework worden vormgegeven dat gebruikt kan worden om met risico’s om te gaan in de context van Responsible AI?’.

Dit onderzoeksrapport beschrijft de wijze waarop het RAI risico framework is ontwikkeld en toont vervolgens een visuele presentatie van het ontwikkelde framework. In lijn met de grondbeginselen van RAI, ligt de focus van dit framework op ‘een verantwoorde implementatie’. In het framework staan vijf beginselen centraal. Deze vijf beginselen zijn: ethiek, verantwoordelijkheid, aansprakelijkheid, privacy & veiligheid en uitlegbaarheid.

Voor ieder van de vijf beginselen zijn, op basis van de literatuur, risico’s en beheersmaatregelen vastgesteld. Deze risico’s en beheersmaatregel zijn opgenomen in een visuele presentatie van het RAI Risico Framework. Door middel van een multiple case study is het RAI risico framework in de praktijk gevalideerd.

Het RAI risico framework kan worden gebruikt om te beoordelen of de implementatie van AI, op het gebied van risicobeheersing, voldoet aan de definitie van responsible AI. Hiervoor dient elk van de risico’s in het framework te worden beoordeeld, en dient te worden gekeken of de beheersmaatregel op een acceptabele wijze is geïmplementeerd.

Summary

This research paper focusses on the field of Responsible AI (RAI). The existing literature lacks a generally applicable definition of RAI, thus a definition had to be created for this research. The definition that was created is: 'A responsible way of implementing AI within organizations in the form of a governance framework'

The field of AI has proven that organization can benefit from using AI, but that there are risks involved as well. It's necessary to address these risks in order to implement AI in a responsible way. In this research, a framework is proposed that can be used to deal with these risks. The main research question of this paper is 'How should a framework be developed that can be used to deal with risks in the context of Responsible AI?'

This research paper describes the way in which the RAI Risk Framework is developed and provides a visual presentation of this framework. Aligned with the principles of RAI, the focus of the framework is 'a responsible implementation'. The five principles which are used in the framework are: ethics, responsibility, accountability, privacy & security and explainability.

Based on existing literature, multiple risks as well as risk mitigating measurements were determined for each of the five mentioned principles. Both the risks and the risk mitigating measurements are included in the visual presentation of the RAI Risk Framework. The framework was validated in practice through a multiple case study.

The RAI Risk Framework can be used to determine whether or not an implementation of AI is, in the context of risk management, in accordance with the definition of RAI. This determination can be done by evaluating each of the risks in the framework and then assess if the related risk mitigating measurement was implemented in an acceptable way.

Inhoudsopgave

1. Introductie

1.1. Achtergrond

De inzet van Artificial Intelligence algoritmen biedt veel kansen ten opzichte van menselijk handelen. Uit onderzoek gepubliceerd in het Harvard Business Review blijkt dat het inzetten van AI het meest effectief is wanneer dit ter ondersteuning is van mensen. Dit blijkt zelfs effectiever dan volledige automatisering door AI omdat het laatste vaak enkel op korte termijn grote voordelen behaalt. Voorbeelden van elementen waar de inzet van AI ter ondersteuning van de mens verbetering teweeg heeft gebracht zijn flexibiliteit, snelheid, schaal, besluitvorming en personalisatie (H. J. Wilson & Daugherty, 2018).

De inzet van AI brengt ook uitdagingen met zich mee. De basis waarop een AI algoritme tot een uitkomst komt op basis van de input, is niet altijd makkelijk verklaarbaar. Deze onverklaarbare werking wordt ook wel de 'black box' van AI genoemd. Het gebied van Explainable AI (XAI) gaat hier verder op in. Een uitkomst kan ook, feitelijk of ethisch, onjuist zijn, met alle consequenties van dien. Een voorbeeld van een ethisch vraagstuk is de vraag hoe een algoritme voor een zelfrijdende auto moet handelen in het geval van een crash-scenario waarin de beslissing fataal kan zijn (Nyholm, 2018). Moet de eigenaar van de auto beschermd worden, moet de weg van de minste schade gekozen worden, of is er een andere oplossing? Ook kunnen praktische vragen worden gesteld zoals 'Wie is verantwoordelijk wanneer fouten worden gemaakt wordt door een algoritme?'

Hierboven genoemde scenario's tonen aan dat het simpelweg inzetten van AI in de bedrijfsvoering niet altijd risicoloos is. Vanuit een risicomanagement perspectief is het in het belang van organisaties om mogelijke risico's te adresseren. Het vakgebied van AI is een breed vakgebied. Het in de vorige alinea benoemde voorbeeld van hoe een zelfrijdende auto de beslissingen zou kunnen nemen toont ook aan dat AI gevaarlijk kan zijn. Maar het gevaar van een mensenleven hoeft uiteraard niet bij elke AI toepassing het geval te zijn. De omgeving waarin een AI wordt ingezet, en welke AI toepassing wordt ingezet, zijn ook van invloed kan zijn op risico's.

De inzet van Artificial intelligence kan dus risico's met zich meebrengen. Governance op het gebied van Artificial Intelligence is daarom nodig om hier richting aan geven. Er bestaan meerdere frameworks met het doel om AI om een verantwoordelijke manier in te zetten. In 2018 heeft een groep van experts in opdracht van de Europese Commissie een framework opgesteld om op een betrouwbare manier met AI om te gaan (European Commission, 2016). De overkoepelende term 'Trustworthy AI' wordt gehanteerd, ofwel TAI. De vier grondslagen van TAI zijn volgens het gepubliceerde rapport van de Europese Commissie 'respect voor menselijke autonomie', 'preventie van schade', 'rechtvaardigheid', 'verantwoording'.

Naast TAI bestaat ook het 'Responsible AI' framework, ook wel RAI genoemd. Volgens een rapport gepubliceerd door Accenture en opgesteld door de CTO, Dominic Delmonino, focust RAI op 'ensuring the ethical, transparent and accountable use of AI technologies in a manner consistent with user expectations, organizational values and societal laws and norms' (Accenture, 2018). Omdat TAI een relatief nieuw framework is met een bredere scope dan enkel de risico's van AI, is dit onderzoek niet gericht op TAI, maar op het beter aansluitende RAI. Dit onderzoek beoogd een aanvulling te zijn op RAI in de vorm van een risico framework. Het risico framework beoogd richting te kunnen geven aan het probleem hoe er omgegaan kan worden met risico's van AI door organisaties op een verantwoorde manier.

1.2. Gebiedsverkenning

Responsible AI is (RAI) is een framework binnen het vakgebied van AI met een sterke focus op verantwoordelijkheid. Binnen RAI staan gebruikers, organisaties en de maatschappij centraal. Er bestaan ook verwante frameworks met een net andere scope zoals Explainable AI (XAI) en Trustworty AI (TAI).

Bij XAI heeft voornamelijk transparantie en uitlegbaarheid een sterke focus. Bij RAI is dit ook een belangrijke factor, maar dat betekent niet dat hier dezelfde waarde aan gegeven hoeft te worden binnen RAI, als bij XAI het geval is.

Er is een duidelijke overloop te herkennen tussen RAI en TAI. TAI poogt een allesomvattend framework te zijn dat breder is dan RAI. In beide frameworks zijn veel dezelfde aspecten te herkennen. De focus van TAI ligt sterk op betrouwbaarheid. Even simplistisch gezegd, ligt de scope van RAI ergens tussen XAI en TAI in.

1.3. Probleemstelling

Het vakgebied van AI heeft bewezen grote voordelen te kunnen bieden aan organisaties. In lijn met de beginselen van Corporate Governance is het voor organisaties belangrijk om schade bij belanghebbende te voorkomen. Het inzetten van AI kan gepaard gaan met risico's. Om tot een beheersbare situatie te komen dient omgegaan te worden met deze risico's.

1.4. Opdrachtformulering

Hoe kan een framework worden vormgegeven dat gebruikt kan worden om met risico's om te gaan in de context van Responsible AI?

1.5. Motivatie / relevantie

Responsible AI is een framework met een sterke focus op het begrip 'verantwoordelijkheid' in de context van het omgaan met AI toepassingen binnen organisaties. Risico's zijn een inherent onderdeel van deze focus. Het kan een waardevolle toevoeging aan RAI zijn om een governance framework te ontwikkelen dat specifiek gericht is op het beheersen van de risico's die relevant (kunnen) zijn voor organisaties in het kader van een AI implementatie. Aanvullend kan een risico framework een instrument zijn dat verder onderzoek stimuleert binnen responsible AI.

1.6. Aanpak in hoofdlijnen

In hoofdlijnen zal allereerst een theoretisch kader worden opgesteld dat het RAI framework duidelijk beschrijft. De definitie van RAI zal duidelijk gedefinieerd worden en vervolgens zal worden vastgesteld welke grenzen RAI kent, in de vorm van beginselen van RAI. Met behulp van een theoretisch kader omtrent de beginselen van RAI zal een risico framework worden opgesteld tezamen met beheersmaatregelen gekoppeld aan deze risico's. Dit risico framework zal in de praktijk worden gevalideerd door middel van een kwalitatief onderzoek in de vorm van een multiple case studie teneinde om tot een gevalideerd RAI risico framework te komen.

2. Theoretisch kader

Het doel van het theoretisch kader is om een basis neer te leggen op basis van beschikbare literatuur. Deze basis wordt vervolgens gebruikt voor het onderzoek naar de risico's van RAI.

2.1. Onderzoeksaanpak

Om op een gestructureerde wijze te zoeken naar de theorie die aansluit bij het onderzoek is de informatiebehoefte uitgesplitst in een viertal deelvragen. De deelvragen zijn de volgende:

(1) Wat is responsible AI?
(2) Welke beginselen van responsible AI kunnen worden vastgesteld?
(3) Wat houdt het concreet in om verantwoordelijk om te gaan met AI?
(4) Hoe ziet het risico framework voor responsible AI eruit?

Grote technologie conglomeraten en consultancy bedrijven zijn op de voorgrond aanwezig in de wereld van AI. Uit het vooronderzoek is gebleken dat bedrijven die zich over het onderwerp uitlaten voornamelijk technologiebedrijven als Microsoft, Google, Meta en IBM zijn, alsmede bedrijven actief binnen consultancy dienstverlening zoals Accenture en Deloitte. Ook geeft de Europese Commissie richtlijnen voor het gebruik van AI in het TAI framework. In de zoekstrategie wordt daarom meegenomen hoe toonaangevende bedrijven en instanties omgaan met responsible AI en wat zij hierover publiceren.

Tevens wordt een literatuurstudie uitgevoerd op basis van wetenschappelijke bronnen. Voor iedere sub onderwerp wordt op basis van kernwoorden naar literatuur gezocht in de EBSCO database. De hits van de zoekquery worden gesorteerd op relevantie. De eerste twee pagina's (20 artikelen) worden beoordeeld op basis van de titel, de onderwerpen(subjects) en de abstract. Indien meer dan 20 hits worden getoond, dan zullen verdere filters worden toegepast waaronder een filter op datum om de meest recente artikelen eerst te tonen. De reden hiervoor is dat het gebied van AI snel ontwikkeld en recentere literatuur daarom de voorkeur heeft boven oudere literatuur. Daarnaast wordt gefilterd op 'Academic journals'.

Op basis van de snowballing techniek (forward en backwards) wordt naar meer relevante artikelen gezocht. Een overzicht van de kernwoorden die per deelvraag zijn gesteld, welke filters zijn toegepast en welke resultaten dit heeft opgeleverd, is te vinden in bijlage 1.

2.2. Resultaten en conclusies

De resultaten en de conclusies van de onderzoeksaanpak zijn hieronder gespecificeerd met een uitsplitsing per deelvraag.

2.2.1. Wat is responsible AI?

Volgens (Clarke, 2019) kan de positie worden ingenomen dat het in het belang is van organisaties om ervoor te zorgen dat schade aan belanghebbende wordt voorkomen. Deze gedachtegang ligt in lijn met de beginselen van Corporate Governance. In het kader van risicobeheersing legt (Clarke, 2019) de nadruk op risico's die komen kijken bij AI door de mate van complexiteit en de 'mysterieuze werking' die AI kan hebben. Bij het volgen bij de redenering kan worden geconcludeerd dat het in het belang van organisaties is om in het kader van risico's terughoudend te zijn bij het implementeren van AI vanwege potentiële schade bij belanghebbenden.

Aan de andere kant heeft AI bewezen een grote positieve impact te kunnen hebben voor organisaties op ten minste flexibiliteit, snelheid, schaal, besluitvorming en personalisatie (H. J. D. Wilson, Paul R., 2018). In tegenstelling tot de terughoudendheid die organisaties zouden moeten hebben in het kader van risico's voor belanghebbende, biedt de technologie dus ook kansen voor de organisaties en de belanghebbende. Deze tegenstrijdigheid, waarbij enerzijds risico's meewegen en anderzijds kansen, geeft noodzaak voor een verantwoorde manier om AI te implementeren zodat risico's kunnen worden gemanaged. Deze verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een governance framework wordt in dit onderzoeksrapport gezien als 'responsible AI (RAI)'.

De klemtoon bij RAI ligt op de 'verantwoorde' implementatie. De definitie van RAI die wordt gehanteerd voor dit onderzoeksrapport is daarom 'Een verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een governance framework'.

De term RAI is ook al in de literatuur gedefinieerd door andere onderzoekers. De gehanteerde definities verschillen, maar zeggen in de kern hetzelfde. (Barredo Arrieta et al., 2020) omschrijft RAI als '*de methodologie om op grote schaal AI methoden te implementeren in organisaties met in achtneming van eerlijkheid, explainability en verantwoordelijkheid in de kern*'.

Wat RAI wezenlijk anders maakt van andere vormen van AI is het in acht nemen van de beginselen die RAI kenmerken. In de bovenstaande definitie van (Barredo Arrieta et al., 2020) zijn deze beginselen 'eerlijkheid', 'explainability' en 'verantwoordelijkheid'. Het vaststellen van welke beginselen wordt uitgegaan in dit onderzoek is noodzakelijk om richting te geven aan de definitie 'Een verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een governance framework'.

2.2.2. Welke beginselen van responsible AI kunnen worden vastgesteld?

Op basis van welke beginselen RAI kan worden geïmplementeerd is niet eenduidig vastgesteld binnen de literatuur, echter worden verwante beginselen wel gehanteerd in verschillende artikelen of door organisaties. Binnen de context van dit onderzoeksrapport wordt uitgegaan van principes die breed gedekt zijn binnen organisaties en de literatuur.

Zoals in paragraaf 2.2.1 is aangegeven wordt de term RAI wordt door (Barredo Arrieta et al., 2020) omschreven als '*de methodologie om op grote schaal AI methoden te implementeren in organisaties met in achtneming van eerlijkheid, explainability en verantwoordelijkheid in de kern*'. De genoemde beginselen 'eerlijkheid', 'explainability' en 'verantwoordelijkheid' zijn in literatuur dus genoemde kenmerken van RAI. Literatuuronderzoek van Doorn wijst uit dat 'transparantie', 'justice & fairness', 'responsibility', 'accountability', 'privacy' en 'non-maleficence' het meest prominent zijn (Doorn, 2021). Clarke belicht de ethiek en risicobeheersing als aspecten van RAI (Clarke, 2019).

Toonaangevende instanties bieden ook een kijk op wat Responsible AI inhoudt. Accenture noemt 'Ethiek', 'transparantie' en 'verantwoordelijk' gebruik als de aspecten die bij RAI komen kijken (Accenture, 2018). Microsoft hanteert in de kijk die zij hebben op RAI, die zij de 'Microsoft AI principes' noemen, de aspecten 'Fairness', 'Reliability & Safety', 'Privacy & Security', 'Inclusiveness', 'Transparancy' en 'Accountability' (Microsoft, 2021).

Binnen de context van dit onderzoeksrapport wordt uitgegaan van een set van vijf beginselen welke tot stand zijn gekomen uit de bovengenoemde literatuur. In bijlage 2 is de link gelegd tussen de beginselen en de benaming of invulling die hieraan wordt gegeven op basis van de literatuur. De vijf gehanteerde beginselen zijn:

1. Ethiek
2. Verantwoordelijkheid
3. Aansprakelijkheid
4. Privacy & veiligheid
5. Uitlegbaarheid

Deze vijf beginselen zijn een aanvulling op de definitie 'Een verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een governance framework'. De beginselen geven richting aan het deel '*De verantwoorde manier*' van de AI implementatie.

De vijf beginselen zijn hetgeen dat RAI onderscheid van gewone AI. Wanneer op een juiste manier om wordt gegaan met deze 'kapstok' van vijf beginselen kan worden gesteld dat de implementatie van AI ook daadwerkelijk Responsible AI is.

2.2.3. Wat houdt het concreet in om verantwoordelijk om te gaan met AI?

Uit de literatuurstudie is gebleken dat ethiek, verantwoordelijkheid, aansprakelijkheid, privacy & veiligheid en uitlegbaarheid van belang zijn voor een verantwoordelijke implementatie van AI. In deze paragraaf wordt gekeken naar de manier hoe invulling gegeven kan worden aan deze vijf beginselen van responsible AI op basis van de beschikbare literatuur.

Ethiek

Savliev claimt dat AI geen puur technologisch fenomeen meer is, maar juist meer een sociaal fenomeen. Reden voor deze claim zijn de zorgen die op grote schaal zijn uitgesproken over de mogelijke schade die AI kan hebben op de maatschappij (Anton Saveliev, 2020). Volgens Unesco is AI voorbestemd om de toekomst te veranderen, maar is nog onduidelijk hoe die verandering er precies uit zal zien en dat leidt volgens Unesco tot een staat van fascinatie en angst tegelijkertijd. Mogelijke zorgen die door Unesco worden genoemd zijn onder andere autonoom opererende wapens, datacollectie die privacy kan schenden en algoritmes die een voorkeur hebben voor een bepaald ras (Unesco, 2018). Een dekkend 'legal framework' bestaat volgens Unesco nog niet. Corporate Social responsibility wordt breed genoemd in relatie tot de ethische kant van AI.

In aanvulling op het onderzoek van Savliev, dat stelt dat AI een sociaal fenomeen is, stelt Sullins dat robots onder bepaalde omstandigheden kunnen worden gezien als 'moral agents' (Sullins, 2006). Dit is het geval wanneer het soort interacties met robots erg lijken op de interactie met andere mensen. De robots zouden dan een besef moeten hebben van wat goed of fout is volgens Sullins. Volgens het onderzoek wordt het ethische vraagstuk bij de inzet van AI met name relevant wanneer een AI een sociale rol inneemt. Een voorbeeld hiervan die Sullins noemt is een situatie waarin een arts een 'morele verplichting' heeft naar patiënten toe. De stelling die Sullins inneemt is dat in het geval dat de arts wordt vervangen door een robot, waarbij de robot op significant niveau autonoom kan handelen, dat de robot dan kan worden gezien als een 'moral agent' en daarmee moreel besef zou moeten hebben.

Volgens het onderzoek van Savliev zijn de meningen verdeeld in het vraagstuk of morele kwesties überhaupt zouden moeten worden uitbesteed aan een AI. Sommige experts claimen dat een AI nooit morele kwesties zouden moeten beoordelen, anderen denken dat het juist de morele autonomie van mensen zou kunnen versterken.

Om de potentie van AI te kunnen benutten op een wijze dat gericht is op ethische aanvaardbaarheid is tijdens de Asilomar conferentie in 2017 een set van AI guidelines opgesteld door de Future of Life Institute (Future Of Life Institute, 2017). Deze richtlijnen omvatten onder andere richtlijnen wanneer AI wel en niet onderzocht zou moeten worden en welke ethische- of morele waarden in acht genomen zouden moeten worden zoals veiligheid, vrijheid en waarden van mensen.

Verantwoordelijkheid

Het aspect van verantwoordelijkheid kan breed worden opgevat waardoor dit een verdere afbakening vereist voor dit onderzoek. In bijlage 2 is de link is gelegd tussen wat literatuur en organisaties aangeven met betrekking tot het RAI beginsel 'verantwoordelijkheid'. Er is vastgesteld dat onder andere 'verantwoordelijk gebruik', 'betrouwbaarheid & veiligheid', 'risicobeheersing' en 'niet-kwaadaardigheid' vallen onder het beginsel verantwoordelijkheid.

In een onderzoek door (Saleema Amershi et al., 2019) wordt gesteld dat wanneer er interactie is tussen mens en AI, dat het mogelijk is dat AI systemen onvoorspelbaar gedrag vertonen. Dit kan volgens het artikel leiden tot disruptie, verwarring, aanvallend gedrag of tot gevaar. Om deze gevolgen te kunnen beperken wordt in het artikel voorgesteld om algemeen toepasbare design-richtlijnen te incorporeren in het design van de AI.

In het kader van verantwoord gebruik van AI maakt Hatfield onderscheid tussen enerzijds AI als aanvulling op de expertise van mensen en anderzijds AI ter vervanging van de expertise van mensen (Hatfield, 2019). Hatfield suggereert door middel van een onderzoek binnen belastingrecht dat het meer verantwoord is om AI in te zetten als een aanvulling op expertise en niet als vervanging daarvan. Het vervangen van de mensen met expertise door een AI wordt door Hatfield zelfs afgeraden. De reden die Hatfield geeft is dat het vervangen van expertise een negatieve invloed heeft op professionaliteit en daarnaast ook een risico kan zijn in het kader van betrouwbaarheid voor de klant. In het kader van risicobeheersing, betrouwbaarheid en veiligheid alsmede verantwoordelijk gebruik kan hiermee de link gelegd worden met het beginsel verantwoordelijkheid. Hierbij dient wel de kanttekening gemaakt te worden dat dit een onderzoek betrof binnen belastingrecht waarbij dus kritisch gekeken dient te worden in hoeverre de resultaten representatief zijn voor organisatie in andere omgevingen.

De kijk waarin AI een aanvulling kan zijn op de mens wordt sterk gedeeld door (H. J. Wilson & Daugherty, 2018) in een onderzoek over Collaborative Intelligence waarin wordt geconcludeerd dat AI het beste resultaat geeft op een verantwoordde manier wanneer deze opereert in samenwerking met de mens. Het doel van de AI moet volgens het artikel zijn om de maximale potentie van de mens te benutten, die resulteert in het beste resultaat.

Robbins heeft onderzoek gedaan naar wat envelopment, een term uit de roboticawereld, betekent in relatie tot AI. De claim van Robbins is dat zowel robots als AI algoritmen succesvol en verantwoord zijn wanneer deze opereren in een gecreëerde mini-omgeving waarin wel de gewenste output behaald kan worden, maar waarin deze mini-omgeving is afgeschermd van onverwachte factoren (Robbins, 2020). Volgens Robbins brengt het extra risico's met zich mee wanneer een AI algoritme

opereert zonder de restrictie van een afgeschermde omgeving. Deze risicobeheersingsmaatregel past daarom goed bij het beginsel van verantwoordelijkheid.

Aansprakelijkheid

Een machine of software kan niet aansprakelijk worden gehouden voor de consequenties van het autonoom handelen. Hierdoor kan het onduidelijk zijn wie er wel verantwoordelijk is voor. (Will Orr, 2019) maakt onderscheid tussen onder andere aansprakelijkheid op het gebied van compliance en op het gebied van ethics. In het onderzoek wordt gesteld dat over compliance en ethics moet worden nagedacht in het design. De aansprakelijkheid op het gebied van ethics heeft overlap met het beginsel 'ethiek'. (Will Orr, 2019) geeft aan dat de wetgeving op het gebied van AI te los, slecht gedefinieerd en ongrijpbaar is omdat het vakgebied AI zeer snel groeit. De wetgeving groeit hierin dus niet snel genoeg mee volgens het artikel, waardoor veel organisaties eigen standaarden ontwikkelen.

Clarke claimt dat publieke regulatie noodzakelijk is voor het verantwoord kunnen benutten van AI technologie vanwege de publieke risico's die bij de technologie komen kijken. Clarke noemt meerdere methoden om regulatie van AI te realiseren en neemt het standpunt in dat 'co-regulatie' de meeste passende is. Co-regulatie houdt in dat een regelgevende instantie een regulatief framework ontwikkeld. Dit regulatieve framework vormt de basis van de wetgeving, maar gaat niet in op de details voor specifieke ondernemingen. De details, welke een aanvulling zijn op het regulatieve framework, worden voor de specifieke ondernemingen vastgesteld in een consultief processen met de regelgevende entiteit.

Bedrijven die responsible AI implementeren dienen te voldoen aan geldende wetgeving, ongeacht welke methode er wordt toegepast. Een voorbeeld van geldende wetgeving die in de Europese Unie geldt is de GDPR wetgeving (European Commission, 2016).

In een onderzoek naar de aansprakelijkheid van AI bij het gebruik van chatbots geeft (Sara Suárez-Gonzalo, 2019) twee mogelijkheden om aansprakelijkheid toe te wijzen, de structuralist en de context-dependent approach. De situatie van de Twitterbot 'Tay', een AI experiment van Microsoft dat in 2016 onethische berichten twitterde, is gebruikt als casus in het onderzoek.

De structuralist approach gaat uit van aansprakelijkheid voor het design en management van de AI. Het artikel van (Sara Suárez-Gonzalo, 2019) noemt een argument tegen de structuralist approach, namelijk dat aansprakelijkheid voor het design en management niet altijd terecht is omdat er niet altijd een causaal verband is tussen de actie van een AI en de intentie die het design en management voor ogen heeft gehad.

De context-dependent approach gaat uit van aansprakelijkheid gebaseerd op de omgeving waarin de AI opereert. De AI leert van input uit de omgeving en daardoor kunnen volgens (Sara Suárez-Gonzalo, 2019) volgens de context-dependent approach de actoren in die omgeving verantwoordelijk worden gehouden. Echter kunnen vraagtekens gesteld worden of de context-dependent approach representatief is voor alle responsible AI, of dat dit voornamelijk geldt voor chatbots.

De conclusie van het onderzoek van (Will Orr, 2019) ligt in lijn met de structuralist approach. Een voorbeeld wordt genoemd waarin 3.000 werknemers van Google een open brief aan de CEO sturen om geen zaken te doen met het Pentagon voor de ontwikkeling van een 'AI Predator Drone'. Dit is

een voorbeeld waarin vanuit het design verantwoordelijkheid genomen wordt op ethisch gebied. Vanuit het design kunnen bepaalde situaties dus worden voorkomen.

Uit de resultaten van de interviews van (Will Orr, 2019) wordt tevens de casus gemaakt dat de structuralist approach ook nadelen heeft. Soms zijn er onbedoelde gevolgen bij het handelen door een AI, en dat kan fatale afloop hebben. De argumenten die worden gemaakt zijn dat zelf in het geval dat er iets wordt gemist in het design door een fout, dat dit niet hoeft te komen door nalatigheid, maar dat het kan komen door de mate van complexiteit. Van belang in zulke situaties is volgens het interview dat aantoonbaar de juiste veiligheidsanalyse wordt gemaakt en het juiste ethical framework wordt toegepast.

Er kan worden geconcludeerd dat het belangrijk is om aansprakelijkheid toe te wijzen, en dat er meerdere mogelijkheden zijn om dit te doen. Een duidelijke best-practice is niet eenduidig aan te wijzen, maar het belang van een juist design op het gebied van ethiek en compliance wordt onderstreept.

Privacy & veiligheid

Er is een verband te leggen tussen privacy en ethiek met betrekking tot AI. Unesco noemt de mogelijkheid van AI om manieren van data collectie toe te passen die privacy schendt (Unesco, 2018).

Microsoft claimt dat de implementatie van AI complexiteit toevoegt aan het bewaken van privacy & veiligheid. Voor machine learning modellen wordt doorgaans veel data gebruikt. Er dient onder andere nagedacht te worden waar de data vandaan komt, waar het heen gaat, waar modellen gedraaid worden en hoe kan worden nagegaan of de data niet uitlekt of moedwillig wordt aangepast (Microsoft, 2021).

Accenture geeft aan dat het noodzakelijk is om privacy, net als transparantie en veiligheid, in het design te incorporeren (Accenture, 2018). Dit is in lijn met de vereiste van ‘Data protection by Design’ dat vanuit de GDPR wetgeving verplicht is binnen Europa. In deze wetgeving zijn onder andere opgenomen welke persoonsdata mag worden verwerkt, onder welke omstandigheid, hoe lang deze mag worden bewaard en hoe dit gecommuniceerd moet worden aan de persoon van wie deze data is (European Commission, 2016).

In de AI guidelines die tijdens de Asilomar conferentie zijn opgesteld, wordt gesteld dat de toegang en de controle tot privacygevoelige data beheerd zou moeten worden door individuen zelf, en mensen daarnaast niet onredelijk in privacy zouden mogen beperken (Future Of Life Institute, 2017). Accenture wijst ernaar dat het geven van toegang en controle tot privacygevoelige data aan de individuen zelfs een potentiële ‘best-practice’ is (Accenture, 2018) en verwijst hierin naar een studie die is uitgevoerd door MIT, waarbij patiënten die zelf complete controle kregen over hun eigen genetische informatie, 83% vaker meededen met een genetische test ten opzichte van een basislijn.

Uitlegbaarheid

Het aspect uitlegbaarheid komt tot stand vanuit de aspecten ‘transparantie’ en ‘explainability’ volgens de literatuurbronnen uit bijlage 2.

Robbins claimt dat voor het maken van keuzes over de regulatie van AI algoritme, dat we meer zouden moeten weten over hoe deze algoritmen werken (Robbins, 2020). Volgens Robbins rust dit met name op de lastig te verklaren manier van de werking van de algoritmen. Specifiek de training data, inputs, outputs, functies en grenzen moeten bekend zijn. Hier kan de link gelegd worden met het vakgebied van XAI, welke ingaat op het verklaren van de werking van AI.

De gedachte dat uitlegbaarheid van belang zijn voor responsible AI wordt gedeeld door Accenture. Accenture stelt dat het belangrijk is, en in sommige situaties zelfs volgens de wet verplicht, om in het design van een AI systeem na te denken over hoe verklaard kan worden wat de grondgedachte is van een AI systeem bij de beslissingsvorming (Accenture, 2018). Accenture noemt specifiek het voorbeeld van de bancaire sector waarbij het, in het geval van het afwijzen van een creditcard applicatie, verplicht is om aan klanten te kunnen verantwoorden waarom die keuze is gemaakt.

Volgens diverse toonaangevende bedrijven, waaronder Google, is XAI een set van tools en frameworks om de uitkomst van voorspellingen van een machine learning model te kunnen verklaren en begrijpen (Google, 2021). Er bestaande diverse aanbieders van tools of frameworks voor Explainable AI, waaronder open source varianten zoals AI Explainability 360 (IBM Research, 2021).

De gedachte dat explainability, of XAI, van belang is voor responsible AI wordt tevens gedeeld door onderzoeker (Barredo Arrieta et al., 2020), welke onderstreept dat wanneer beslissingen van een AI impact kunnen hebben op mensenleven, het steeds belangrijker wordt om te snappen hoe beslissingen van de AI tot stand komen. Specifiek worden de vakgebieden ‘medicijnen’, ‘recht’ en ‘defensie’ genoemd waar explainability van extra belang is.

Als aanvulling op de vakgebieden ‘medicijnen’, ‘recht’ en defensie, kan volgens de AI guidelines die tijdens de Asilomar conferentie zijn opgesteld, worden gesteld dat het onethisch is niet te kunnen verklaren waarom bepaalde keuzes worden gemaakt in het geval dat en AI systeem schade aanricht. Daarnaast zou het te alle tijde mogelijk moeten zijn voor een competent mens om te kunnen achterhalen waarom een beslissing is gemaakt bij autonome AI systemen (Future Of Life Institute, 2017).

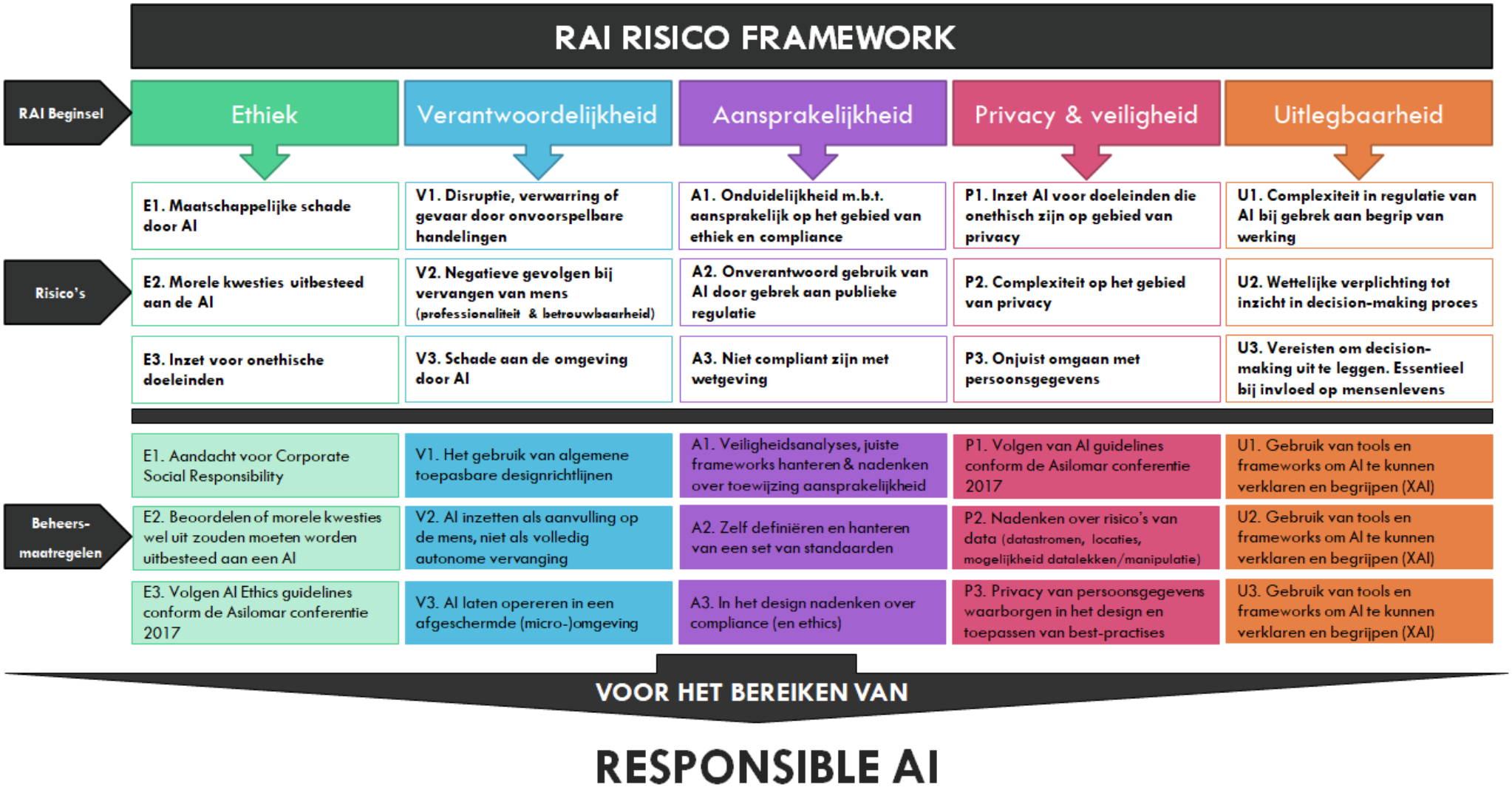
2.2.4. Hoe ziet het risico framework voor responsible AI eruit?

In paragraaf 2.2.3 is de huidige beschikbare kennis van responsible AI verzameld vanuit de literatuur en toonaangevende bedrijven. Op basis van deze theorie is een risico framework ontwikkeld dat kan worden gebruikt door organisaties om op een verantwoorde manier om te gaan met de risico's van het implementeren van AI. Het risicoframe is opgenomen in afbeelding 1 op de volgende pagina.

In het risico framework is onderscheid gemaakt tussen de vijf beginselen van Responsible AI die zijn gedefinieerd in paragraaf 2.2.2. Voor ieder van de vijf beginselen van responsible AI is een analyse uitgevoerd welke risico's en welke beheersmaatregelen kunnen worden geïdentificeerd op basis van de theorie uit hoofdstuk 2.2.3. De identificatie van risico's en beheersmaatregelen is opgenomen in bijlage 3. De bijlage dient tevens als de documentatie van het risico framework. Toeval heeft bepaald dat er exact drie risico's zijn geïdentificeerd voor ieder van de vijf beginselen. Ieder risico's is gekoppeld aan een uniek nummer dat correspondeert met het risico framework.

Met behulp van het risico framework kan worden beoordeeld of een implementatie van AI, op het gebied van risico's, voldoet aan de definitie van responsible AI: 'Een verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een governance framework'.

De beoordeling wordt gedaan in een proces van twee stappen. Als eerste stap wordt beoordeeld welk van de risico's van AI mogelijk van toepassing is op de specifieke implementatie. Als tweede stap wordt voor ieder van de risico's die van toepassing is, beoordeeld of de corresponderende beheersmaatregel op een acceptabele wijze is geadresseerd. Op deze wijze kan per beginsel worden bepaald of wordt voldaan aan de definitie. Om volgens het risico framework te kunnen spreken van responsible AI dient een implementatie van AI volledig conform alle vijf beginselen uit het model te zijn



Afbeelding 1: RAI risico framework. Documentatie is opgenomen in bijlage 3.

2.2.5. Conclusie

Op basis van het onderzoek is de definitie van responsible AI vastgesteld als ‘een verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een governance framework’. Om aan responsible AI te kunnen voldoen zijn vijf beginselen van responsible AI vastgesteld. Deze beginselen zijn: ethiek, verantwoordelijkheid, aansprakelijkheid, privacy en uitlegbaarheid.

Voor elk van de vijf beginselen zijn risico’s en beheersmaatregelen geïdentificeerd. Op het gebied van ethiek is geconcludeerd dat het onzeker is hoe de toekomst van AI eruit ziet. Het belang van Corporate Social Responsibility wordt daarom onderstreept. Wanneer een AI een functie vervult waarin morele kwesties worden uitbesteed aan een AI, en de AI daardoor een functie als ‘moral agent’ kan krijgen, dan ontstaan er onder onderzoekers ethische bedenkingen. Meningen verschillen of morele kwesties überhaupt uitbesteed zouden moeten worden aan een AI. De inzet van AI voor niet-ethische doeleinden is potentieel een risico. Tijdens de Asilomar conferentie in 2017 zijn AI Ethics Guidelines opgesteld om hier richting aan te geven.

Op het gebied van verantwoordelijkheid kan worden geconcludeerd dat de werking van AI onvoorspelbaar kan zijn. Wanneer een AI wordt ingezet in een omgeving waarbij interactie is met mensen, kan dat potentieel leiden tot disruptie, verwarring of gevaar. Dit kan worden gematigd door het volgen van algemeen toepasbare designrichtlijnen. Daarnaast bestaat het risico op schade aan de omgeving. Dit risico kan beperkt worden door een afgesloten omgeving waarin de AI kan opereren en alsnog de gewenste output kan behalen. Ten slot toont onderzoek aan dat het in bepaalde situaties een risico kan zijn voor de professionaliteit en betrouwbaarheid wanneer menselijke experts worden vervangen door AI. Dit risico kan worden beperkt door de inzet van AI ter ondersteuning van de mens, niet als vervanging. Deze aanpak is volgens onderzoek tevens als effectiever gebleken.

Op het gebied van aansprakelijkheid kan worden geconcludeerd dat het niet altijd duidelijk is wie aansprakelijk is in het geval dat een actie door een AI is gedaan. Aantoonbare veiligheidsanalyse, de juiste (ethische) frameworks en nadenken over het vraagstuk van het toewijzen van aansprakelijkheid kunnen hierin ondersteunen. De theorie noemt onder andere de structuralist approach en de context-dependent approach als mogelijke methoden van toewijzen van aansprakelijkheid. Onderzoek toont aan dat publieke regulatie achterblijft op de groei van AI en dat kan een risico zijn op het gebied van aansprakelijkheid. Dit kan worden beperkt door organisaties door middel van het toepassen van een eigen set van standaarden. Daarnaast dient voldaan te worden aan wetgeving die er wel is. Dit kan het beste gemanaged worden door dit in het design te incorporeren.

Op het gebied van privacy brengt AI een potentieel nieuwe mogelijkheden met zich mee voor het schenden met privacy. Om te zorgen dat een AI verantwoord geïmplementeerd wordt kunnen de AI Ethics Guidelines worden gevolgd die zijn opgesteld tijdens de Asilomar conferentie in 2017. Ook brengt AI extra complexiteit met zich mee mede door de hoeveelheid data die benodigd is voor machine learning modellen. Er dient hiervoor goed nagedacht te worden waar modellen gedraaid worden, hoe datastromen lopen en hoe ervoor kan worden gezorgd dat data niet uitlekt of wordt gemanipuleerd. Op het gebied van privacy dient tevens correct opgegaan te worden met persoonsgegevens. In Europa geldt hiervoor de GDPR wetgeving te worden gevolgd. Daarnaast is het volgens onderzoek potentieel een ‘best-practise’ om de toegang en het beheer van persoonsgegevens bij individuen te leggen.

Op het gebied van uitlegbaarheid kan het reguleren van een AI complex worden indien het niet duidelijk is hoe de AI precies werkt. Volgens onderzoek is het met name van belang dat de training data, input, outputs, functies en grenzen bekend zijn. Aanvullend is het in sommige situaties zelfs wettelijk verplicht om uitleg te geven bij een beslissing. In het geval dat een beslissing gemaakt wordt door een AI betekent dat ook dat deze beslissing uitgelegd moet kunnen worden. In situaties waar een AI invloed kan hebben op mensenlevens, specifiek de gebieden 'medicijnen', 'recht' en 'defensie' komen uit onderzoek naar voren, dan kan het als ethisch onaanvaardbaar worden gezien om geen uitleg te kunnen geven. Om deze risico's met betrekking tot uitlegbaarheid te kunnen managen kan Explainable AI worden gebruikt, wat inhoudt dat tools en/of framework worden toegepast om de uitkomst van AI te kunnen verklaren.

De bovengenoemde risico's en beheersmaatregelen zijn gebruikt om een risico framework te ontwikkelen. Dit risico framework kan worden gebruikt om te beoordelen of een implementatie van AI, op het gebied van risicobeheersing, voldoet aan de definitie van responsible AI. Dit kan worden gedaan door de risico's die mogelijk van toepassing kunnen zijn voor de specifieke implementatie van AI te identificeren in het risico framework, en te beoordelen of de beheersmaatregelen op acceptabele wijze zijn geadresseerd. Om te kunnen spreken van responsible AI volgens het risico framework dient aan alle vijf de beginselen te worden voldaan.

2.3. Doel van het vervolgonderzoek

Het doel van het vervolgonderzoek is om het risico framework dat is opgenomen in paragraaf 2.2.4 in afbeelding 1 te valideren in de praktijk. De intentie van het risico framework is om organisaties te helpen met het beheersen van de risico's van AI, zodat AI op een verantwoorde wijze kan worden geïmplementeerd. Hiermee is het doel van het onderzoek om bij te dragen aan het gebied van responsible AI.

Het framework is gebaseerd op theorie uit het literatuuronderzoek en zal worden gevalideerd in de praktijk. Het vervolg onderzoek wordt dus deductief uitgevoerd. Dit zal op kwalitatieve wijze worden gedaan door middel van één of meerdere case studies. Hiervoor zullen één of meerdere geschikte bedrijven gevonden moeten worden.

3. Methodologie

Het doel van de methodologie is om verantwoording te geven voor het empirisch onderzoek. Dit wordt gedaan door allereerst een conceptueel ontwerp te schetsen. Vervolgens wordt ingegaan op het technische ontwerp. Als derde wordt ingegaan op de manier van gegevensanalyse. Als laatste wordt verantwoording afgelegd over de validiteit en de betrouwbaarheid.

3.1. Conceptueel ontwerp: keuze van onderzoeksmethode(n)

Het Responsible AI risico framework is op deductieve wijze tot stand gekomen, zijnde dat de theorie de basis vormt van het framework. Of het framework ook daadwerkelijke aansluit met- en bruikbaar is in de praktijk is niet automatisch bewezen omdat het puur gebaseerd is op theorie. Om deze reden dient het framework nog te worden gevalideerd. Het doel van het empirisch onderzoek is om het framework te valideren in de praktijk.

Voor het empirisch onderzoek is het noodzakelijk dat er informatie vanuit de praktijk wordt verzameld. De benodigde informatie is een uitwerking van de risico's die organisaties tegenkomen bij de implementatie van responsible AI en bijbehorende beheersmaatregelen die worden gehanteerd. Met deze informatie zal worden getoetst of de risico's die organisaties zien, indien deze ook terecht zijn aangewezen als risico binnen responsible AI, ook daadwerkelijk in het framework zijn opgenomen. De beheersmaatregel die in het framework is gekoppeld aan het risico zal vervolgens worden vergeleken met de beheersmaatregel van de specifieke organisatie.

Voor het empirisch onderzoek is een afweging gemaakt tussen een kwalitatieve methode of een kwantitatieve methode van onderzoek. Het responsible AI risico framework is algemeen toepasbaar, wat inhoudt dat het framework op alle soorten organisaties van toepassing zou moeten zijn, waardoor het voordelen kan bieden om grotere hoeveelheden organisaties te betrekken in het empirisch onderzoek. Bij een kwantitatieve studie is het doorgaans eenvoudiger om een grotere aantal organisaties mee te nemen in het onderzoek. Een kwantitatieve studie kan hierin dus voordelen bieden op het gebied van volledigheid, omdat meer verschillende soorten bedrijven kunnen worden benaderd. Echter is het van noodzakelijk belang om de risico's van specifieke AI implementaties correct te definiëren. De invulling die gegeven wordt aan een risico kan sterk verschillen per implementatie of organisatie. Hoe risico's van responsible AI binnen een organisatie er precies uitzien kan dus lastig zijn om kwantitatief uit te drukken, er is simpelweg een bepaalde mate van inschattingsvermogen en aandacht voor de specifieke risico nodig. Naar verwachting komt deze diepgang bij een kwalitatief onderzoek erg goed naar voren. De verwachting is dus dat de kwaliteit van het onderzoek bij kwalitatief onderzoek aanzienlijk beter zal zijn waardoor de keuze valt op kwalitatief onderzoek.

Om het risico af te dekken dat de risico's die gelden voor de geïnterviewde organisatie niet dekkend genoeg zijn voor het risico framework, waardoor risico bestaat dat het empirisch onderzoek niet representatief is voor de bevindingen in de literatuurstudie, is het niet voldoende om slechts één organisatie op te nemen in het empirisch onderzoek. Om deze reden is gekozen om een kwalitatief onderzoek uit te vormen in de vorm van een multiple case studie.

Om de het juiste niveau van diepgang te kunnen krijgen, is gekozen om informatie te verzamelen in de vorm van interviews bij de verschillende organisaties in het kader van een AI implementaties die is gedaan. Het heeft de voorkeur om organisaties te interviewen die van mekaar verschillen, of in ieder geval AI implementaties hebben gedaan die verschillen. Dit zal de naar verwachting meer representatief zijn. Kort samengevat, het op kwalitatieve wijze, door middel van het afnemen van interviews bij meerdere case organisaties, zal naar verwachting het juiste niveau van detail kunnen

bieden om het risico framework te valideren in de praktijd en daarbij tevens representatief genoeg kunnen zijn voor het algemeen-toepasbare risico framework.

3.2. Technisch ontwerp: uitwerking van de methode

In vorige paragraaf is besloten dat zal worden gekozen voor een kwalitatief onderzoek doormiddel van interviews bij meerdere case organisaties.

Omdat het risico framework algemeen toepasbaar is, dat wil zeggen voor ieder soort AI implementatie zou moeten gelden, zal het de validiteit van de resultaten ten goede komen als de case organisaties van mekaar verschillen. De reden hiervoor is dat niet alle soorten risico's van toepassing zijn op alle soorten AI implementaties. Er kan dus niet gesteld worden dat wanneer een risico dat bij één van de case organisatie niet speelt, dat deze voor andere case organisaties ook niet geldt.

De stelling wordt ingenomen dat kwaliteit boven kwantiteit verkozen dient te worden om te voorkomen dat bepaalde risico's voor organisaties wordt gemist, of verkeerd ingeschat, in het onderzoek. De kwalitatieve studie dient correct en met de juiste aandacht te worden uitgevoerd. Dit is tevens de reden geweest om voor een kwalitatieve studie te kiezen boven een kwantitatieve studie. Anderzijds is het voor de betrouwbaarheid van de resultaten belangrijk om zoveel mogelijk organisaties te betrekken in het empirisch onderzoek. Hoe meer organisaties uiteindelijk worden gevonden, en hoe meer deze organisaties van mekaar verschillen, hoe beter de resultaten naar verwachting zullen zijn. Om bovengenoemde redenen, waarbij de kwaliteit voorop staat maar wel waarde wordt gehecht aan zoveel mogelijk case organisaties, wordt drie case organisaties gezien al het optimale aantal. Het aantal 'drie' is een subjectief bepaalde, maar wel onderbouwde hoeveelheid. Dit aantal is dekkend genoeg mits er verschillende soorten organisaties worden gekozen, deze hoeveelheid is haalbaar binnen het onderzoek en er kan worden vergelijken tussen de drie case organisaties. Het kiezen voor vier- of meer case organisaties kan potentieel leiden tot problematiek in de haalbaarheid van het onderzoek binnen de gestelde tijd.

Om te waarborgen dat alle beginselen van het risico framework allemaal worden geraakt bij het onderzoek zal specifiek gezocht worden naar passende organisaties. Voor alle bedrijven zal gekeken worden naar alle beginselen, echter speelt niet elk beginsel een even grote rol voor alle organisaties door de verschillende aard van de AI toepassing. De verwachting is dat de beginselen 'aansprakelijkheid' en 'ethiek' voor iedere soort toepassing een rol kan spelen, echter verschilt uiteraard de invulling ervan. Gerelateerd aan het beginsel 'Verantwoordelijkheid' blijkt uit het risico framework dat meerdere risico's betrekking hebben op interacties met mensen. Om deze risico's hieromtrent te kunnen valideren is het daarom wenselijk dat het empirische onderzoek in ieder geval een organisatie bevat die een AI heeft ontwikkeld die opereert in een omgeving waarbij interactie met mensen een rol speelt. Vervolgens zal ook een organisatie worden gezocht waarbij door de AI toepassing omgegaan wordt met persoonsgegevens, gezien het belang in het beginsel 'privacy'. Ten derde wordt gezocht naar een organisatie waarbij uitlegbaarheid een rol speelt. Met deze organisaties zal het empirisch onderzoek naar verwachting dekkend genoeg zijn voor het valideren van het risico framework.

Interviews zullen afgenomen worden bij de case organisaties. Aangezien het risico framework gericht is op de implementatie van AI, zal een contactpersoon worden gezocht die betrokken is bij deze implementatie. Er zal specifiek worden gezocht naar iemand die betrokken is geweest bij de design keuzes en afwegingen hieromtrent. Daarnaast zullen de risico's worden gevalideerd bij de stakeholders die het meeste verband hebben met het beginsel. Welke stakeholder dit is zal af

hangen van de organisatie en zal daarom pas bepaald worden wanneer meer informatie bekend is over de organisaties.

De interviews zullen worden opgenomen en vervolgens worden uitgewerkt. Allereerst wordt ervoor gezorgd dat de interviews een vaste structuur hebben zodat verzekerd kan worden dat de inhoud kan worden vergeleken met het risico framework.

De opzet en structuur van het interview is in detail uitgewerkt in bijlage 4.

Met behulp van de uitkomsten van de interviews met de case-organisaties, wordt achteraf beoordeeld in hoeverre het risico framework toepasbaar is om risico's in te schatten bij de implementatie van responsible AI, of in hoeverre de resultaten van het onderzoek kunnen worden gebruik om het framework aan te vullen of aan te passen. In het geval van aanvulling kan dit leiden tot aanvullingen in het framework zelf en daarmee ontstaat dan een nieuw (gevalideerd) risico framework.

3.3. Gegevensanalyse

Zoals in paragraaf 3.1 is vermeld wordt het kwalitatieve onderzoek deductief uitgevoerd. De gekozen vorm van analyse wordt door (Saunders, 2019) beschreven als Deductive Explanation Building. De gegevens uit het risico framework zullen worden gevalideerd met de gegevens uit het kwalitatieve onderzoek (de interviews). Indien er bevindingen zijn die niet overeenstemmen met het risico framework, dan kan dit leiden tot het aanpassen en iteratief valideren.

De data wordt verzameld door middel van interviews. De analyse van deze data zal gedaan worden conform de richtlijnen die (Saunders, 2019) geeft voor het analyseren van kwalitatieve data. Alle interviews zullen worden opgenomen door middel van audio opnameapparatuur en worden getranscribeerd. In het transcript wordt aandacht besteed aan wat er verbaal wordt gezegd en ook non-verbale aspecten zoals intonatie en context worden meegenomen. Indien de geïnterviewde er geen bezwaar tegen heeft als er beeldmateriaal wordt opgenomen, zal ook dit worden gedaan met als doel de context juist vast te kunnen leggen. Indien er geen beeldmateriaal wordt opgenomen dan zal de interviewer notities maken en het transcript zo kort mogelijk na het interview opstellen. Alle namen en transcripten zullen geanonimiseerd worden zodat deze niet herleidbaar zijn naar de geïnterviewde of de organisatie.

Om te voorkomen dat er transcriptiefouten worden gemaakt zal, conform richtlijnen van Saunders, worden gezorgd voor het proces van data cleaning. Het transcript wordt ter confirmatie tevens naar de geïnterviewde gestuurd.

Om het proces van het transcriberen efficiënt te laten verlopen zal gebruik worden gemaakt van data sampling. Dit houdt in dat de relevante zaken worden losgekoppeld van de niet relevante zaken, en alleen de relevante zaken worden getranscribeerd. Dit wordt gedaan om de tijd van het transcriberen te verkorten. Om potentiële problemen of fouten te voorkomen die samenhangen met deze techniek, zullen de volgende drie punten in acht genomen worden:

1. De gehele audio-opname van het interview wordt ten minste tweemaal volledig teruggeluisterd.
2. De audio-opname wordt gebruikt om het transcript te valideren om te zorgen dat er geen informatie wordt gemist.
3. De secties worden nauwkeurig en met aandacht getranscribeerd.

Er wordt aandacht besteed aan het coderen van de interviews om effectief te kunnen analyseren. Er wordt pattern matching toegepast, waarin het RAI risico framework de uitkomsten op basis van theorie voorspelt, terwijl deze in de praktijk worden gevalideerd in de data. Stukken data worden onder andere gelabeld zodat makkelijk kan worden vergeleken, gegroepeerd en geanalyseerd. Voor de case organisaties worden de risico's en beheersmaatregelen in kaart gebracht die komen kijken bij de AI implementatie. Voor elk van de gevonden risico's wordt getoetst of deze vallen binnen de definitie van responsible AI uit paragraaf 2.2.1 en in hoeverre de beheersmaatregelen volledig zijn voor het afdekken van het specifieke risico. Het RAI risico framework vormt daarom de basis van de labels.

De labels bestaand uit een code van 3 tekens. Het eerste teken (een R of een B) refereert naar respectievelijk een **Risico** of een **Beheersmaatregel**. Het tweede teken (letters E, V, A, P of U) refereert naar de letters van de vijf RAI beginselen. Het derde teken refereert naar het nummer van het risico of de beheersmaatregel. Een stuk data kan ook meerdere labels meekrijgen indien dit betrekking heeft op meerdere groepen in het RAI risico framework. Een voorbeeld van een label kan dus zijn [RE1] wat inhoudt dat het om en risico gaat binnen het beginsel 'Ethiek', specifiek het nummer dat correspondeert met 'maatschappelijke schade door een AI'. De labels zijn dus gelinkt aan het RAI risico framework. Het is ook mogelijk om één van de drie tekens te vervangen met een streep om meer generiek te kunnen verwijzen, bijvoorbeeld [-E1], [RE-] of [-E-]. Door het toepassen van deze labels kunnen meerdere case organisaties met elkaar vergeleken worden, gericht op de specifieke onderdelen van het risico framework. De relevante stukken data uit case study zullen, uitgesplitst per label, in een tabel worden opgenomen om vergelijken en analyseren gemakkelijker te maken. In bijlage 4, de interview opzet, wordt in meer detail ingegaan op deze vormgeving.

3.4. Reflectie t.a.v. validiteit, betrouwbaarheid

In deze paragraaf wordt verantwoording afgelegd voor de keuzes in het onderwerp op de gebieden van betrouwbaarheid en validiteit.

Betrouwbaarheid

Om te zorgen dat de resultaten van het onderzoek betrouwbaar zijn wordt een multiple case studie uitgevoerd, waarbij meerdere organisaties met ander soort AI implementatie de voorkeur heeft. Hoe de organisaties verschillen, hoe representatiever de uitkomst is voor een AI implementatie in algemene zin. Hoe frequenter een vergelijkbaar risico naar voren komt, hoe meer de juistheid ervan wordt bevestigd. Hoe meer de organisaties en implementaties van mekaar verschillen, hoe groter de kans is dat andere risico's worden gevonden.

Het zwakke punt in de opzet is dat niet ieder risico hoeft voor te komen bij de case organisaties. Het is daarom vooraf al de verwachting dat niet ieder risico zal worden gevalideerd in de praktijk. Dit is een zwak punt voor de betrouwbaarheid van het framework aangezien er wellicht meer case organisaties nodig zijn om ieder risico en iedere beheersmaatregel te valideren. Aanvullend zal pas een hoge mate van betrouwbaarheid kunnen worden geconcludeerd indien dezelfde resultaten door meerdere case studies worden bevestigd. Vervolgonderzoek kan de betrouwbaarheid verhogen.

Om toch tot een zo correct mogelijke conclusie te komen is het daarom extra relevant dat de onderzoeker een kritische houding aanneemt en conclusies trekt op basis van solide logica.

Validiteit

Om te zorgen dat de resultaten van het onderzoek valide zijn, is gekozen voor een kwalitatief onderzoek op basis van interviews. Deze vorm van communicatie zorgt ervoor dat kritisch kan

worden gekeken naar de case organisatie door de onderzoeker, wat van belang is om risico's goed te definiëren. Er wordt aandacht besteed aan een duidelijke structuur in het interview waarbij het risico op zowel het missen van risico's, of het onterecht definiëren van een risico, door middel van het ontwerp van het interview zoveel mogelijk wordt voorkomen.

Het zwakke punt in de opzet voor de validiteit is dat de kwalitatieve wijze van informatie verzamelen risico geeft op fouten of inschattingsfouten. De validiteit in het onderzoek is afhankelijk van de conclusies van de onderzoeker en van de informatievoorziening vanuit de organisaties. Een zekere vorm van complexiteit is geen uitzondering bij AI implementaties. Er wordt dus ook niet uitgegaan worden dat de organisatie zelf een volledige of perfecte risico inschatting maakt. De kritische houding van de onderzoeker is daarom van cruciaal belang voor de validiteit. De kennis- en kunde van de geïnterviewde vanuit de case organisatie dient ook goed gewaarborgd te worden. Om deze reden zal een inventarisatie van deze kennis- en kunde worden opgenomen in de interview structuur.

Met een vooraf uitgedachte, duidelijke structuur zoals in bijlage 4 is opgenomen, wordt gepoogd de risico's met betrekking tot validiteit tot een minimum te beperken. Ook het multiple case aspect matigt dit risico omdat bevestiging vanuit meerdere cases kan worden gehaald.

4. Resultaten

Het resultaat van het onderzoek naar de risico's van responsible AI is tweedelig. Het eerste resultaat is het risico framework dat in paragraaf 2.4 vorm heeft gekregen, op basis van het theoretische onderzoek. De in bijlage 3 opgenomen theoretische documentatie van het model is hierbij van belang.

Het tweede resultaat van het onderzoek is de validatie van het framework door middel van het empirisch onderzoek. Het resultaat van de empirische validatie van het risico framework leidt tot een nieuw risico framework, dat in de praktijk is gevalideerd.

De resultaten worden in verschillende onderdelen behandeld, namelijk:

- Het selecteren van de case organisaties
- De uitwerking van de case studies
- Resultaten gegevensanalyse
- Het aangepaste risico framework

4.1. Het selecteren van de case organisaties

Op basis van de in paragraaf 3 opgestelde structuur voor het empirisch onderzoek, en de in paragraaf 3.2 opgestelde ideaal situatie voor het selecteren van AI implementaties binnen organisaties, is gezocht naar case organisaties. Het model is algemeen toepasbaar, dat wil zeggen dat iedere organisatie met een AI implementatie in theorie geschikt is als case organisatie. De keuze is gemaakt om ervoor te zorgen dat de case organisaties van mekaar verschillen zodat de beginselen van risico framework zo volledig mogelijk worden gevalideerd. Daarom is met grote zorg gekeken naar geschikte organisaties. Deze 'ideaal situatie' kende specifieke vereiste voor de case organisaties. Doel van de multiple case studie is, met het oog op de validiteit, om in totaal in ieder geval de volgende situaties te hebben belicht.

- Interactie met mensen
- Omgaan met persoonsgegevens
- Uitlegbaarheid van het decision making proces van de AI

Er is gekozen om kwaliteit te verkiezen boven kwantiteit, waarbij het ideale aantal organisaties is vastgesteld op drie. Meer organisaties wordt wel als beter beschouwd voor de betrouwbaarheid, maar omwille van kwaliteit boven kwantiteit wordt het simpelweg als niet haalbaar gezien om meer case studies met dezelfde aandacht en kwaliteit te behandelen.

De eerst geselecteerde organisatie is een hypotheekverstrekker. Deze organisatie heeft AI in haar bedrijfsprocessen geïmplementeerd om het hypotheekaanvraag proces te verbeteren. Deze AI implementatie is verwerkt in een portaal dat zowel gebruikt wordt door interne afdelingen van de hypotheekverstrekker zelf en door externen zoals een intermediair. In het portaal worden documenten getoetst. Voorbeelden van documenten zijn paspoorten (uit verschillende landen), loonstroken in allerlei verschillende formats en documenten/contracten waarop eventueel een handtekening of paraaf moet staan. Door middel van AI modellen worden hier controles op gedaan zodat direct kan worden teruggekoppeld indien een document niet goed is, om de vertraging te voorkomen van een menselijke controle en daarmee potentieel een afkeuring van het document in

een later stadium van het proces. Hierdoor kan het hypotheek aanvraag proces aanzienlijk versneld worden.

Voor deze case organisatie speelt de interactie met mensen, zoals de intermediair en de interne afdeling, een rol. Ook wordt omgegaan met documenten met de hoogst mogelijke classificatie van gevoelige informatie en persoonsgegevens. Als laatste heeft de organisatie aangegeven dat het terugkoppelen waarom een document is goedgekeurd of afgekeurd, van belang is. Echter is dit geen directe noodzakelijk in het inzichtelijk maken van de 'black box' van AI.

De tweede geselecteerde case studie is een ontwikkelaar van een product, een slimme camera, dat wordt verkocht aan met name organisaties van de overheid. Het product bevat een ecosysteem aan verschillende toepassingen die middels een dashboard inzichtelijk worden gemaakt. In deze camera is veelvuldig gebruik gemaakt van AI modellen, met name gericht op het vakgebied van computer vision. De slimme camera kan bijvoorbeeld de afstand tussen verschillende mensen in een ruimte herkennen. Met deze informatie kan onder andere worden beoordeeld of de door de overheid voorgeschreven afstand tussen personen wordt nageleefd tijdens een pandemie zoals de Covid pandemie. De camera telt ook hoeveel personen er in een gebouw aanwezig zijn, door bij te houden wie een gebouw in- en uit loopt, wat mede wordt ingezet om de veiligheid van verschillende ruimten inzichtelijk te maken. Ook dit tellen van personen is een voorbeeld van een applicatie die tijdens een pandemie relevant kan zijn voor de veiligheid, echter kan dit ook ingezet kan worden ter ondersteuning van een veiligheidsafdeling om andere redenen.

De slimme camera heeft een vorm van interactie van mensen, omdat het gedrag van mensen op beeld een directe invoer is van de AI modellen. Er wordt omgegaan met gevoelige data, zijnde de reeksen van foto's van mensen en wat er ook in de omgeving gebeurt waar de camera wordt opgehangen. De organisatie geeft aan dat het uitleggen wat er binnen de black box van AI gebeurt om tot een uitkomst te komen, totaal irrelevant is voor de case organisatie.

De derde geselecteerde organisatie voor de case studie was op medisch vlak gericht op diagnostiek van patiënten gericht op onderzoek naar kanker. De contactpersoon beheerde meerdere IT systemen waarin onder andere ook AI een rol speelde. De organisatie was uitgekozen voor de uitlegbaarheid van de 'black box' van AI. Echter is besloten om deze case studie niet door te zetten omdat de contactpersoon niet de juiste betrokkenheid heeft gehad bij specifiek de design van de AI implementatie.

4.2. De uitwerking van de case studies

De interviews bij de twee case organisaties zijn afgenomen conform de opzet in bijlage 4. De interviews zijn tevens opgenomen met behulp van opnameapparatuur. Conform de in paragraaf 3.3 genoemde wijze van gegevensanalyse is een transcript gemaakt.

In de praktijk is echter gebleken dat de techniek van 'data sampling', waarvan in paragraaf 3.3 is aangegeven dat deze toegepast zou worden, niet effectief zou zijn. De interviews bevatte veel waardevolle informatie in ieder moment van het interview. Het weglaten van stukken zou het risico opleveren dat bij de gegevensanalyse belangrijke informatie zou worden gemist. Om die reden is de keuze gemaakt om de techniek van data sampling los te laten en de interviews volledig, woord voor

woord, te transcriberen. Op basis van het transcript, en het volledig terugluisteren van de ca. 4 uur aan opnamen, is het proces van data labelen gestart.

Allereerst is gestart met open coderen met als doel om belangrijke stukken data te voorzien van een label. Deze coderingen zijn geplaatst om stukken data te kunnen koppelen aan één of meer beginselen in het RAI risico framework alsmede om een richting te geven aan welk risico of aan welke beheersmaatregel de data betrekking heeft.

De in bijlage 4 beschreven opzet van de interviews, waarin de risico's en beheersmaatregelen in chronologische volgorde zijn besproken, heeft ervoor gezorgd dat de data van de case studies vrij geclusterd was gestructureerd. Deze clusters zijn gekoppeld aan de beginselen in het RAI risico framework. Er is gebleken dat er wel regelmatig overlap plaats heeft gevonden binnen de clusters data, omdat in het gesprek regelmatig een communicatief bruggetje is gemaakt naar de onderdelen die op een ander moment in het interview aan bod zouden komen. Bij het proces van open coderen is deze overlap duidelijk naar voren gekomen, en is richting gegeven aan welke beginselen, welke risico's en welke beheersmaatregel er bij de specifieke stukken data hoort.

Vervolgens is de stap van het axiaal coderen toegepast. Hierbij zijn de open coderingen vervangen door de in paragraaf 3.3 beschreven labels bestaande uit drie tekens. Deze drie tekens zijn verbonden met het RAI risico framework door respectievelijk aan te geven op welk van de vijf beginselen een label betrekking, of het een risico of een beheersmaatregel betreft, en welke volgnummer er in het RAI framework is gebruikt. Tijdens het axiaal coderen van de data is gebleken dat er vaak afwegingen gemaakt moesten worden welk label een bepaald risico zou moeten krijgen, omdat een stuk data ook betrekken kan hebben op meerdere onderdelen van het RAI risico framework. In de gevallen waar er na een kritische afweging nog steeds twijfel bestond, is gekozen om de data van meerdere labels te voorzien. De beslisboom die is opgenomen in afbeelding 4.1 van bijlage 4 is veelvuldig toegepast om tot een uniforme en onderbouwde uitkomst te komen bij het maken van beslissingen. De stap van axiaal coderen heeft ertoe geleid dat alle relevante data van de interviews zowel te linken zijn aan het RAI risico framework alsmede aan andere interviews. Het gelabelde transcript is opgenomen in het separate document 'transcript case studies'.

Als derde stap is selectief coderen toegepast. Dit is gedaan door de inhoud van de gelabelde data te interpreteren en de essentie van de boodschap op te nemen in de tabel voor de gegevensanalyse uit tabel 4.1 van bijlage 4. Deze tabel voegt de resultaten van beide case studies overzichtelijk op één plaats samen. Het maken van een vergelijking is hierdoor mogelijk.

Naast de interpretatie vanuit het transcript is voor ieder risico en iedere beheersmaatregel een separate conclusie getrokken en toegelicht. Het doel hiervan is het bepalen in hoeverre het risico en de beheersmaatregel kan worden gevalideerd, of moet worden aangepast. Deze tabel met de conclusies worden in de volgende paragraaf 'Resultaten gegevensanalyse' behandeld.

Er wordt onder andere gelet op:

- Bevatten de case studies dezelfde resultaten
- Zijn deze resultaten niet in strijd met de theorie
- Zijn de resultaten logisch

Een risico voor de betrouwbaarheid van de resultaten is dat er een discrepantie kan zitten tussen wat de interviewer denkt en de feitelijke situatie. Een voorbeeld kan zijn dat de geïnterviewde denkt dat een bepaald risico niet bestaat, terwijl het eigenlijk het gevolg is geweest van een bepaalde design keuze waarin het risico is voorkomen. Om deze reden zal deze feitelijk waarheid soms

moeten worden geëxtrapoleerd uit hetgeen dat gezegd is alsmede door te kijken naar de reden achter bepaalde beheersmaatregelen die naar voren komen in de case studies.

De kwalitatieve benadering waarbij op basis van logica zal moeten worden geredeneerd vergt daarom de grootste aandacht en een kritische houding, met name met het oog op een mogelijke confirmatie bias. Om dit risico zoveel mogelijk te mitigeren is bevestiging gevraagd aan de geïnterviewde personen.

4.3. Resultaten gegevensanalyse

In deze paragraaf is voor ieder risico en iedere beheersmaatregel van het RAI risico framework een conclusie getrokken op basis van de case studies. De tabellen die zijn opgenomen in bijlage 5 bevatten de interpretatie van de gelabelde data uit de case studies. Deze data vormt de basis waarop de conclusies in deze paragraaf zijn getrokken.

E1 - Ethiek

RE1: Maatschappelijke schade door AI
BE1: Aandacht voor Corporate Social Responsibility

In beide case studies bestaat het risico op maatschappelijke schade door een AI. Het risico is daarom gevalideerd. De beheersmaatregel volgens het framework is ‘aandacht te hebben voor Corporate Social Responsibility (CSR)’. De beheersmaatregel sluit goed aan met wat beide organisaties zeggen, echter wordt dit door de casestudies in bredere en algemenere zin benoemd. Er is naar voren gekomen dat het bespreekbaar maken van het onderwerp ‘ethiek’ in maatschappelijke zin belangrijk is. De conclusie van de case studie is dat de beheersmaatregel in het framework beter kan worden verwoord als ‘Aandacht hebben voor CSR en bespreekbaar maken van ethische dilemma’s’.

E2 - Ethiek

RE2: Morele kwesties uitbesteed aan de AI
BE2: Beoordelen of morele kwesties wel uit zouden moeten worden besteed aan een AI

In beide case studies worden er geen morele kwesties uitbesteed aan de AI. De AI treedt dus niet op als een ‘moral agent’. Echter in de beoordeling of het een bewuste keuze is geweest dat de AI niet als moral agent optreedt, is wel een opmerkelijke gelijkenis in de twee case studies te vinden. In beide organisaties was het mogelijk geweest om de AI wel morele keuzes te laten maken. In de kwestie ‘krijgt een persoon wel of geen hypotheek’ en in mindere mate bij de slimme camera met keuzes als ‘mag een persoon de ruimte nog betreden ook al zit deze ruimte vol’. In beide gevallen is gezegd dat het risico niet speelt, en wordt de exact zelfde redenering gegeven dat de AI informatief werkt, waarin een mens de uiteindelijke beslissing maakt. Een reden voor deze informatieve werking kan logischerwijs zijn dat het als een groot risico werd gezien om de AI als beslissingsorgaan te zien. Dit risico is duidelijk naar voren gekomen vanuit de geïnterviewden. Zowel het risico als de beheersmaatregel kunnen om bovenstaande reden worden gezien als gevalideerd.

E3 - Ethiek

RE3: Inzet voor onethische doeleinden
BE3: Volgen AI Ethics guidelines conform de Asilomar conferentie 2017

In case studie 1 komt geen risico naar voren op het gebied van de inzet van de AI voor onethische doeleinden. Er wordt voor gezorgd dat dit niet het geval is door middel van een objectieve benadering en een uitgedacht proces. Er kan worden gesteld dat dit risico en deze beheersmaatregel voor case studie 1 niet aansluiten met wat er wordt bedoeld in het risico framework, en is opgenomen in de documentatie ervan.

In case studie 2 speelt er een soort ethisch dilemma die wel zeer goed aansluit met het framework, de inzet van een slimme camera om afstand te meten tussen personen tijdens de Covid pandemie. Dit betekent niet dat de inzet van de slimme camera onethisch is, maar wel dat een discussies hierover gevoerd kan worden. Als beheersmaatregelen heeft de organisaties twee maatregelen benoemd. Geen van deze is opgenomen als beheersmaatregel in het risico Ai framework. De beheersmaatregel in het model is dus niet gevalideerd, maar het risico wel.

Enerzijds wordt gezegd dat de afnemer de 'ethische verantwoordelijkheid' heeft. Hierbij is ook raakvlak te vinden met ethsiche Aansprakelijkheid (A1 in het risico framework). Vanuit de theorie in paragraaf 2.2.3 is op basis van onderzoek van (Will Orr, 2019) een voorbeeld genoemd van de structuralist approach. Hier wordt een voorbeeld genoemd waarin 3.000 werknemers van Google een open brief aan de CEO sturen om geen zaken te doen met het Pentagon voor de ontwikkeling van een 'AI Predator Drone'. De case organisatie kiest hier echter voor de context-dependent approach.

Anderzijds wordt gezegd dat de juiste ethische afweging gemaakt dient te worden met de opties die er zijn. Het product kan aangepast worden op een manier dat de ethische kwesties geen, of in mindere mate, meer een probleem vormen. 'Het maken van een ethische afweging' sluit aan bij de theorie van het Future of Life Institute (Future Of Life Institute, 2017) waarbij AI guidelines zijn opgesteld. De omschrijving van de case organisatie kan worden gezien als een meer algemene benaming, en de 'AI Ethics guidelines' zoals die benoemd staan in de beheersmaatregel BE3 kunnen meer worden gezien als het hulpmiddel. Om deze reden kan de omschrijving in het model worden aangepast naar 'Het maken van een juiste ethische afweging' waarbij de documentatie stelt dat hiervoor 'AI Ethics guidelines' beschikbaar zijn.

E4 - Ethiek:

Aanvullende bevindingen, niet in model opgenomen:

In case studie 1 is naar voren gekomen dat een model ongewenste side effects kan hebben wanneer een dataset biased getraind is. Hierbij is specifiek gesproken in de context van mensen. Het biased trainen kan ook ethische implicaties hebben, zoals naar voren is gekomen in de case studies waarin het risico bestaat dat de hypotheekaanvraag langer kan duren voor personen van een bepaald ras. Ondanks dat dit risico niet puur met ethiek te maken heeft, kan de impact hiervan binnen de definitie van responsible AI wel degelijk met ethiek te maken hebben. Het risico op deze 'biased getrainde dataset in relatie tot mensen' is daarom een toevoeging op het risico framework.

De beheersmaatregel ‘een diverse palet aan trainingsdata en een agnostische aanpak’ is wat in de praktijk wordt gehanteerd als beheersmaatregel. Deze beheersmaatregel wordt breed ondersteund door verschillende bronnen die in gaan op de best-practises van AI.

V1 – Verantwoordelijkheid

RV1: Disruptie, verwarring of gevaar door onvoorspelbare handelingen
BV1: Het gebruik van algemene toepasbare designrichtlijnen

In case studie 1 is aangegeven dat disruptie, verwarring of gevaar door onvoorspelbare handelingen niet aanwezig is. In case studie 2 zijn wel degelijk situaties geïdentificeerd waarin verwarring kan ontstaan, Er is benadrukt dat het voornamelijk het menselijk handelen is op basis van de verwarring die de AI potentieel kan scheppen. De beheersmaatregel die wordt gebruikt voor dit probleem zit voornamelijk in de combinatie van transparant zijn over de werking van het model, en de mens de beslissingen laten maken, niet de AI. Dit zorgt ervoor dat de informatie waarop de keuze is gebaseerd, goed wordt geïnterpreteerd. Transparantie over de succes rate, waarover wordt gesproken, kan worden gezien als onderdeel van een goed design. Onderzoek van (H. J. Wilson & Daugherty, 2018) bevestigd dat deze samenwerking tussen AI en mens de verantwoordelijkheid ten goede komt. De beheersmaatregel kan hiermee worden aangepast naar ‘Samenwerking tussen mens en AI en het gebruik van algemene toepasbare designrichtlijnen’.

V2 – Verantwoordelijkheid

RV2: Negatieve gevolgen bij vervangen van mens (professionaliteit & betrouwbaarheid)
BV2: AI inzetten als aanvulling op de mens, niet als volledig autonome vervanging

In beide case studies komt duidelijk naar voren dat het vervangen van de mens door een AI duidelijke risico’s met zich mee brengt. Echter wordt nadrukkelijk gesproken in de context van ‘beslissingen door de AI’ in beide case studies, wat ook door de theorie wordt benadrukt. Dit kan worden verduidelijkt in de omschrijving van het risico in het framework. De beheersmaatregel, AI inzetten als aanvulling op de mens, komt exact overeen met de uitkomst van zowel case studie 1 als case studie 2 en is daarmee gevalideerd.

V3 – Verantwoordelijkheid

RV3: Schade aan de omgeving door AI
BV3: AI laten opereren in een afgeschermde (micro-) omgeving

Risico op schade aan de omgeving is voor case studie 2 geheel niet van toepassing. Echter voor case studie 1 is een interessante bevinding. In het model is schade aan de omgeving opgenomen, echter is hierin het besef niet direct geweest dat dit ook schade binnen de interne organisatie kan zijn op een meer emotioneel vlak bij de eigen werknemers. Als aanvulling op het model is hierom verduidelijkt door het risico ‘schade aan de omgeving of interne organisatie door AI’.

In de case studie 1 is besproken in hoeverre een afgeschermde omgeving de risico's beperkt, waarbij wel is gezegd dat er duidelijke kaders zijn. Waar meer nadruk ligt is het risico wanneer de mens wordt vervangen door AI, en dat een samenwerking beter is. Dit wordt tevens door het onderzoek van (H. J. Wilson & Daugherty, 2018) bevestigd.

A1 - Aansprakelijkheid

RA1 Onduidelijkheid m.b.t. aansprakelijk op het gebied van ethiek en compliance
BE1 Veiligheidsanalyses, juiste frameworks hanteren & nadenken over toewijzing aansprakelijkheid

In beide case studies is duidelijk naar voren gekomen dat het vraagstuk met betrekking tot aansprakelijkheid van toepassing is en dat er goed is nagedacht over het aspect van toewijzing van aansprakelijkheid wanneer het op compliance aan komt. Op ethisch gebied was dit vraagstuk met name bij case studie 2 relevant. De manieren van het toewijzen van aansprakelijk volgens (Sara Suárez-Gonzalo, 2019) zijn tijdens de case studies naar voren gekomen. In case studie 1 is meer gekozen voor de structuralist approach en in case studie 2 juist voor de context dependent approach. De beheersmaatregel sluit voor beide case organisaties aan met het model.

A2 - Aansprakelijkheid

RA2 Onverantwoord gebruik van AI door gebrek aan publieke regulatie
BE2 Zelf definiëren en hanteren van een set van standaarden

In case studie 1 wordt het risico door de geïnterviewde niet herkend. Er wordt met name gekeken naar de wettelijke standaarden. In case studie 2 komt naar voren dat de wet minder voor schrijft dan dat het bedrijf zelf belangrijk vindt. De wet stelt bepaalde eisen, en daarbovenop bestaan er bepaalde (niet wettelijke) standaarden zoals de ISO-27001, die aanvullende richtlijnen geven. Hierbij komt duidelijk naar voren dat de 'set van standaarden' waarover in de beheersmaatregel in het framework gesproken wordt, niet door een organisatie zelf gedefinieerd hoeft te zijn. Er kunnen ook publiekelijk beschikbare standaarden bestaan. In de beheersmaatregel wordt daarom weggelaten dat deze set van standaarden zelf gedefinieerd moet zijn, door de beheersmaatregel aan te passen naar 'Hanteren van niet-wettelijke standaarden'. De case studie bevestigd hiermee het risico en de beheersmaatregel in het framework.

A3 - Aansprakelijkheid

RA3 Niet compliant zijn met wetgeving
BE3 In het design nadenken over compliance (en ethics)

In beide case studies speelt de GDPR wetgeving een zeer belangrijke rol. Beide case studies geven aan dat deze compliance in het design is geïncorporeerd. Een opvallende bevinding in beide case studies is dat er specialistische kennis wordt betrokken wanneer het om compliance gaat. In case studie 1 ging dit om een legal team, en in case studie 2 om de product owner. Door de bevestiging

vanuit beide case studies en de logica achter deze beheersmaatregel kan deze beheersmaatregel aan het risico framework worden toegevoegd.

P1 - Privacy & Veiligheid

RP1: Inzet AI voor doeleinden die onethisch zijn op gebied van privacy
BP1: Volgen van AI guidelines conform de Asilomar conferentie 2017

In case studie 1 is naar voren gekomen dat er op het gebied van privacy geen onethische doeleinden zijn, maar de inzet puur zakelijk is. In case studie 2 zijn wel dilemma's naar voren gekomen in een meer ethische hoek. Het risico in het framework wordt herkend door de organisatie en op dat gebied worden afwegingen gemaakt door het organisatie. De geïnterviewde heeft hierin ook een vergelijking gemaakt van wat er mogelijk zou zijn met een slimme camera, door te wijzen op de smart camera's die door de Chinese regering worden ingezet voor handhaving van de wet. Het risico in het framework is gevalideerd door het bevestigen van dit risico door de organisatie. De organisatie in case studie 2 heeft er bewust voor gekozen om de ethische afwegingen zelf te maken en het product onpersoonlijk te houden, zonder nadrukkelijk de AI guidelines in het Asilomar conferentie te gebruiken. Het aanpassen van de beheersmaatregel naar een meer generaliserende omschrijving 'Maken juiste ethische afweging en bespreekbaar maken ethische dilemma's' is daarom een terechte aanpassing ten goede van het risico framework.

P2 - Privacy & Veiligheid

RP2: Complexiteit op het gebied van privacy
BP2: Nadenken over risico's van data (datastromen, locaties, mogelijkheid datalekken/manipulatie)

In beide case studies komt duidelijk naar voren dat er veel complexiteit is op het gebied van privacy, met een duidelijke richting van data veiligheid. De omschrijving van het risico kan bijgesteld worden naar 'Complexiteit op het gebied van data & privacy' om dit meer passend te maken met zowel de theorie als de beide case studies.

Beide organisaties hanteren duidelijke authenticatie en autorisatie protocollen. Het incorporeren van best-practises op het gebied van cyber security komt naar voren bij beide case studies en het theoretisch framework. Of hier een aparte afdeling bij betrokken moet worden, zoals in case studie 1 naar voren is gekomen, is meer gericht op de keuzes van een individuele casus en organisatie en hoeft niet algemeen te gelden. De nieuwe omschrijving van de beheersmaatregel, die past bij de theorie en de bevindingen in de twee case studies, is 'Data security in design incorporeren & authenticatie en autorisatie protocollen.'.

P3 - Privacy & Veiligheid

RP3: Onjuist omgaan met persoonsgegevens
BP3: Privacy van persoonsgegevens waarborgen in het design en toepassen van best-practises

Beide case organisaties weken met gevoelige data en onderstrepen de risico's hiervan. Dit risico in het framework is door deze bevestiging van de case organisaties gevalideerd.

Beide case studies betrekken personen met kennis op legal gebied om te zorgen dat de geldende wetgeving op een juiste manier wordt nageleefd. Case studie 1 doet dit door middel van een legal afdeling, case studie 2 door middel van de product owner. De betrokken partij op legal gebied dient de wetgeving te kennen. Beide case studies onderstrepen dat zij de relevante wetgeving in het design hebben geïncorporeerd, en dat zo strikt mogelijk hebben gedaan. Daarnaast benadrukt case studie 2 nog de extra standaarden die de wet niet voorschrijft, zoals de ISO-27001, waaraan wordt voldaan. De betrokkenheid van kennis op legal gebied kan worden toegevoegd aan de beheersmaatregel op basis van de bovenstaande conclusie.

U1 - Uitlegbaarheid

RU1: Complexiteit in regulatie van AI bij gebrek aan begrip van werking
BU1: Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)

Beide case studies hebben aangegeven dat zij niet verplicht zijn om de werking van de modellen toe te lichten. Echter is vanuit de theorie bedoeld dat er begrip zou moeten zijn van hoe de black box van AI werkt zodat de juiste keuze gemaakt kunnen worden (Robbins, 2020) en (Accenture, 2018).

Op gebied van uitlegbaarheid is in case studie 1 duidelijk naar voren gekomen dat dit begrip van de werking essentieel is, omdat de AI anders in de praktijk niet goed zal werken. Hier is eigenlijk het risico genoemd dat het model 'blind draait'. Het risico kan beter worden verwoord naar 'Onbedoelde gevolgen van een AI model bij complexiteit of gebrek aan begrip over werking' om aansluiting bij zowel de theorie als de praktijk te behouden,.

De manier hoe hier mee omgegaan wordt, is veelzijdiger dan initieel in het risico framework is aangegeven. Het toepassing van logging, monitoren en het geven van feedback aan de gebruiker is volgens de organisatie van belang. Aanvullend kan door middel van het inbouwen van een feedback loop, gezorgd worden dat het model continue kan blijven leren zodat de succes rate kan worden blijven verbeterd. De beheersmaatregel in het model is om deze reden aangepast naar: 'Logging, monitoring en terugkoppeling en het blijven verbeteren van het model.'

U2 - Uitlegbaarheid

RU2: Wettelijke verplichting tot inzicht in decision-making proces
BU2: Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)

Beide case studies hebben aangegeven dat zij geen wettelijk verplichting hebben tot het uitleggen waarom een bepaalde keuze is gemaakt. Case studie 2 heeft aangegeven dat het ook niet relevant is voor ze waarom een bepaalde uitkomst op die manier tot stand komt, zolang de succes rate maar

onderbouwd kan worden. In case studie 1 wordt wel de wens tot inzicht geven in het decision making proces genoemd, in tegenstelling tot het hebben van de verplichting. Met uitlegbaarheid op het gebied van AI, vanuit de literatuur, wordt in het framework puur het uitleggen van de black box in het AI model bedoeld. Deze 'wens' van case studie 1 kan dezelfde technische oplossing hebben als de 'wettelijke verplichting'. Los van de mate waarin uitlegbaarheid wordt nagestreefd, is het einddoel hetzelfde, het uitleggen van deze black box. Uit de case studies is gebleken dat de uitsplitsing tussen 'wens' en 'verplichting' als zijnde twee verschillende risico's, voor verwarring kan zorgen. In het model kan het risico worden herschreven naar 'Wettelijke verplichting of wens tot inzage in de black box van AI'

In case studie 1 wordt dit risico echter niet aangepakt door het verklaren van de black box van AI. De organisatie heeft ervoor gekozen om niet de black box, maar juist de input van het model te beoordelen met behulp van zelf opgestelde tools en frameworks. Omdat er al logischerwijs kan worden verwacht dat de uitkomst foutief bij een onjuist input, zoals bij het uploaden van een verkeerd document, is dit voldoende voor de case organisatie. De beheersmaatregel die de organisatie hanteert kan gezien worden als een praktische manier. In case studie 2 wordt de opmerking gemaakt dat de inzet van tools en software een logische manier zou zijn om de black box van AI te achterhalen, omdat de modellen zodanig complex kunnen worden dat andere manieren simpelweg niet haalbaar zijn.

U3 - Uitlegbaarheid

RU3: Vereisten om decision-making uit te leggen. Essentieel bij invloed op mensenlevens
BU3: Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)

Voor casestudie 1 is geconcludeerd dat het niet essentieel is om het decision making proces uit te leggen voor de organisatie. Voor case studie 2 is geconcludeerd dat een 'vereiste' om decision making uit te leggen wel erg lijkt op 'wettelijke verplichting' om decision making uit te leggen. Een duidelijkere omschrijving, waarin de twee verschillende risico's niet separaat worden genoemd, kan verwarring voorkomen. Gezien de overlap van dit risico met het met risico van 'RU2' is er voor gekozen om U2 en U3 samen te voegen op basis van de uitkomsten van de case studies.

4.4. Het aangepaste risico framework

Op basis van de conclusies in de case studies is een aangepaste risico framework opgesteld. De verschillende risico's en beheersmaatregelen zijn met behulp van de case studies gevalideerd met de praktijk, waardoor een gevalideerd RAI risico framework is ontstaan.

Er zijn omschrijvingen gewijzigd wanneer uit de praktijk is gebleken dat deze beter passend zijn, met de eis dat deze nog in lijn met de theorie waren. Ook is er een nieuw risico toegevoegd en zijn er risico's samengevoegd. De redenering en de aansluiting met de theorie zijn onderbouwd voor de gemaakte keuzes. Gewijzigde omschrijvingen komen het meest frequent voor. Het belang van deze gewijzigde omschrijvingen dient onderstreept te worden vanwege de positieve impact die deze soort wijzigingen heeft op hoe volledig, accuraat of duidelijk het RAI risico framework is.

In de onderstaande lijst zijn alle verschillen tussen het theoretische model en het gevalideerde model weergegeven. Om herhaling van de conclusies te voorkomen fungeert de onderstaande opsomming puur om de wijzigingen overzichtelijk te presenteren. De onderbouwing van iedere keuze is in paragraaf 4.3 gegeven. De visuele weergave van het gevalideerde model wordt vervolgens ook gegeven.

Ethiek

Op het gebied van Ethiek is de beschrijving van beheersmaatregelen 1 en 3 aangepast en is daarnaast een nieuw risico en beheersmaatregel toegevoegd, gelabeld met het nummer E4.

E1 - De oude omschrijving van beheersmaatregel 1 was 'Aandacht voor Corporate Social Responsibility'. Deze omschrijving is gewijzigd naar 'Aandacht hebben voor CSR en bespreekbaar maken van ethische dilemma's'.

E3 - De oude omschrijving van beheersmaatregel 3 was 'Volgen AI Ethics guidelines conform de Asilomar conferentie 2017'. Deze omschrijving is gewijzigd in 'Het maken van een juiste ethische afweging'

E4 – Er is een risico en beheersmaatregel toegevoegd aan het model, namelijk:
Risico**:** Een dataset dat biased is getraind in relatie tot mensen
Beheersmaatregel**:** Een diverse palet aan trainingsdata en een agnostische aanpak

Verantwoordelijkheid

Op het gebied van verantwoordelijkheid is de omschrijving van de risico's aangepast voor risico's 2 en 3. De omschrijving van de beheersmaatregelen is aangepast voor beheersmaatregel 1 en 3.

V1 – Aan de huidige omschrijving van de beheersmaatregel is een toevoeging gedaan door de omschrijving te wijzigen van 'Het gebruik van algemene toepasbare designrichtlijnen' naar 'Samenwerking tussen mens en AI en het gebruik van algemene toepasbare designrichtlijnen'.

V2 – De omschrijving van het risico is aangepast naar een meer generaliserende vorm van 'Negatieve gevolgen bij vervangen van mens (professionaliteit & betrouwbaarheid)' naar 'Negatieve gevolgen bij vervangen van mens als beslissingsorgaan'

V3 – Het risico is aangepast van 'Schade aan de omgeving door AI' naar 'Schade aan de omgeving of interne organisatie door AI' en de beheersmaatregel is aangepast van 'AI laten opereren in een

afgeschermde (micro-)omgeving' naar 'Samenwerking tussen mens en AI. AI laten opereren in een afgeschermde omgeving'

Aansprakelijkheid

Op het gebied van aansprakelijkheid zijn de omschrijvingen van beheersmaatregel 2 en 3 aangepast.

A2 – De omschrijving van de beheersmaatregel is aangepast van 'Zelf definiëren en hanteren van een set van standaarden' naar 'Hanteren van niet-wettelijke standaarden'

A3 – De omschrijving van de beheersmaatregel is aangepast van 'In het design nadenken over compliance (en ethics)' naar 'Betrokkenheid specialistische teams bij het design op gebied van compliance'

Privacy & Veiligheid

Op het gebied van Privacy en veiligheid is de omschrijving van risico 2 aangepast en zijn de omschrijvingen van beheersmaatregelen 1, 2 en 3 aangepast.

P1 – De omschrijving van de beheersmaatregel is aangepast van 'Volgen van AI guidelines conform de Asilomar conferentie 2017' naar 'Maken juiste ethische afweging en bespreekbaar maken ethische dilemma's'

P2 – De omschrijving van het risico is aangepast van 'Complexiteit op het gebied van privacy' naar 'Complexiteit op het gebied van data & privacy'. De omschrijving van de beheersmaatregel is aangepast van 'Nadenken over risico's van data (datastromen, locaties, mogelijkheid datalekken/manipulatie)' naar 'Data security in design incorporeren. Authenticatie en autorisatie protocollen'

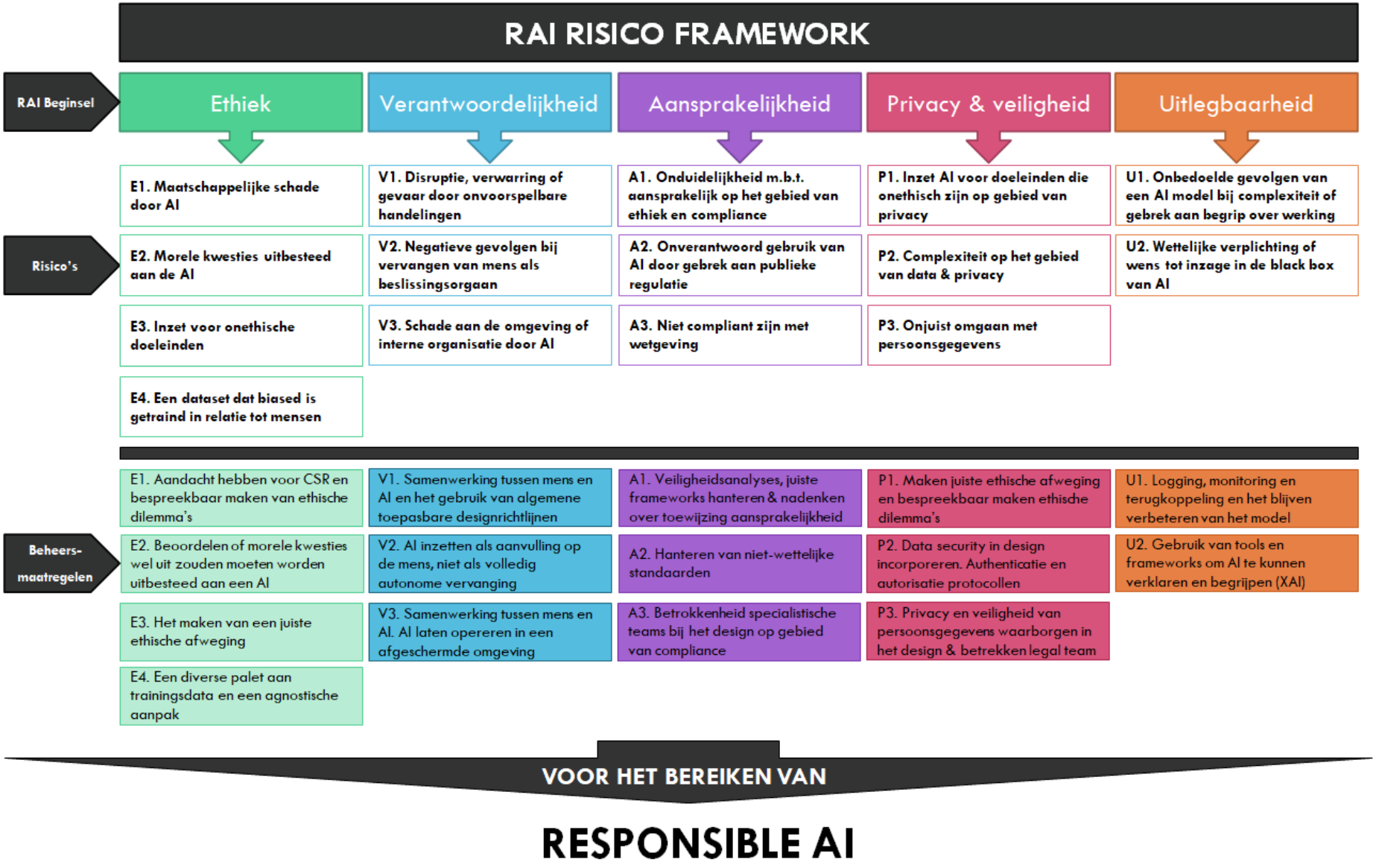
P3 – De beheersmaatregel is aangepast van 'Privacy van persoonsgegevens waarborgen in het design en toepassen van best-practises' naar 'Privacy en veiligheid van persoonsgegevens waarborgen in het design & betrekken legal team'.

Uitlegbaarheid

Op het gebied van uitlegbaarheid zijn zowel de omschrijving van het eerste risico alsmede de eerste beheersmaatregel aangepast. Risico 2 en 3 zijn samengevoegd tot één risico.

U1 – De omschrijving van het risico is aangepast van 'Complexiteit in regulatie van AI bij gebrek aan begrip van werking' naar 'Onbedoelde gevolgen van een AI model bij complexiteit of gebrek aan begrip over werking'. De beheersmaatregel is aangepast van 'Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)' naar 'Logging, monitoring en terugkoppeling en het blijven verbeteren van het model'.

U2 & U3 – Risico's 2 en 3 zijn samengevoegd tot één omschrijving. Risico 2 was omschreven als 'Wettelijke verplichting tot inzicht in decision-making proces' en risico 3 was omschreven als 'Vereisten om decision-making uit te leggen. Essentieel bij invloed op mensenlevens'. De omschrijvingen zijn samengevoegd tot 'Wettelijke verplichting of wens tot inzage in de black box van AI'. De beheersmaatregel van risico 2 en 3 waren dezelfde, en zijn niet aangepast.


RAI RISICO FRAMEWORK
RAI Beginsel
Ethiek
Verantwoordelijkheid
Aansprakelijkheid
Privacy & veiligheid
Uitlegbaarheid
Risico's
E1. Maatschappelijke schade door AI
V1. Disruptie, verwarring of gevaar door onvoorspelbare handelingen
A1. Onduidelijkheid m.b.t. aansprakelijk op het gebied van ethiek en compliance
P1. Inzet AI voor doeleinden die onethisch zijn op gebied van privacy
U1. Onbedoelde gevolgen van een AI model bij complexiteit of gebrek aan begrip over werking
E2. Morele kwesties uitbesteed aan de AI
V2. Negatieve gevolgen bij vervangen van mens als beslissingsorgaan
A2. Onverantwoord gebruik van AI door gebrek aan publieke regulatie
P2. Complexiteit op het gebied van data & privacy
U2. Wettelijke verplichting of wens tot inzage in de black box van AI
E3. Inzet voor onethische doeleinden
V3. Schade aan de omgeving of interne organisatie door AI
A3. Niet compliant zijn met wetgeving
P3. Onjuist omgaan met persoonsgegevens
E4. Een dataset dat biased is getraind in relatie tot mensen
Beheers-maatregelen
E1. Aandacht hebben voor CSR en bespreekbaar maken van ethische dilemma's
V1. Samenwerking tussen mens en AI en het gebruik van algemene toepasbare designrichtlijnen
A1. Veiligheidsanalyses, juiste frameworks hanteren & nadenken over toewijzing aansprakelijkheid
P1. Maken juiste ethische afweging en bespreekbaar maken ethische dilemma's
U1. Logging, monitoring en terugkoppeling en het blijven verbeteren van het model
E2. Beoordelen of morele kwesties wel uit zouden moeten worden uitbesteed aan een AI
V2. AI inzetten als aanvulling op de mens, niet als volledig autonome vervanging
A2. Hanteren van niet-wettelijke standaarden
P2. Data security in design incorporeren. Authenticatie en autorisatie protocollen
U2. Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)
E3. Het maken van een juiste ethische afweging
V3. Samenwerking tussen mens en AI. AI laten opereren in een afgeschermde omgeving
A3. Betrokkenheid specialistische teams bij het design op gebied van compliance
P3. Privacy en veiligheid van persoonsgegevens waarborgen in het design & betrekken legal team
E4. Een diverse palet aan trainingsdata en een agnostische aanpak
VOOR HET BEREIKEN VAN
RESPONSIBLE AI

5. Discussie, conclusies en aanbevelingen

In de theorie zijn verschillende artikelen te vinden die ingaan op specifieke risico's van Responsible AI. Deze artikelen hebben een sterke focus op een specifiek onderdeel van RAI. Om waarde toe te kunnen voegen aan de literatuur is in dit onderzoek gekozen voor een aanpak die anders is. In tegenstelling tot artikelen met deze ingezoomde focus richting specifieke risico's binnen RAI, is in voor dit onderzoek gekozen voor een meer holistische benadering.

Er is gekeken naar hoe de al bestaande literatuur, artikelen met een sterke focus op een kleiner onderdeel van de risico's van RAI, past binnen het holistische geheel. Dit is gedaan door middel van het RAI risico framework. Het RAI risico framework in dit onderzoek is de toevoeging van dit onderzoeksrapport aan de bestaande literatuur.

In paragraaf 1.4 is de hoofdvraag gedefinieerd als 'Hoe kan een framework worden vormgegeven dat gebruikt kan worden om met risico's om te gaan in de context van Responsible AI?'. In dit rapport is systematisch toegewerkt naar het beantwoorden van deze hoofdvraag. In dit hoofdstuk wordt gereflecteerd op het onderzoek zelf, de resultaten van het onderzoek en de conclusies die hieruit getrokken kunnen worden. Er worden daarnaast aanbevelingen gedaan voor de praktijk en voor vervolgonderzoek.

5.1. Discussie – reflectie

Het belangrijkste resultaat van het onderzoek is het RAI risico framework en de validatie hiervan in de praktijk. Dit framework kan worden omschreven als een visuele representatie van de risico's van Responsible AI, en de maatregelen die genomen kunnen worden om deze risico's te mitigeren.

Op gebied van de betrouwbaarheid van het onderzoek is vooraf de ideaal situatie vastgesteld, een multiple case studie bestaande uit drie organisaties die gezamenlijk alle vijf beginselen vertegenwoordigen. Meer organisaties zou de betrouwbaarheid ten goede komen, maar omwille van kwaliteit boven kwantiteit als niet haalbaar gezien om alle case studies met dezelfde aandacht te behandelen. In de praktijk zijn twee case studies uitgevoerd. De derde case studie is niet uitgevoerd vanwege sterke twijfels over de mate van betrokkenheid over specifiek de AI componenten van de contactpersoon bij het bedrijf.

Ondanks de kleine sample van case studies is in de praktijk gebleken dat voor ieder beginsel een validatie plaats kon vinden, en dat voor ieder beginsel waardevolle conclusies konden worden getrokken. Er was tevens voldoende informatie beschikbaar om te kunnen vergelijken tussen de twee case studies. In de gegevensanalyse is vooraf aangegeven dat er validiteitsrisico's konden bestaan, en dat een kritische houding en solide logica van essentieel belang was voor de validiteit. Een kritische houding, het onderbouwen met de theorie, zo onderbouwd mogelijk redeneren en validatie achteraf bij de geïnterviewden, heeft de focus gehad tijdens het empirisch onderzoek. Door deze genoemde aanpak is de betrouwbaarheid en validiteit van het onderzoek niet in het geding gekomen, ondanks de minder dan vooraf bepaalde aantal case studies van 2.

Bij het opstellen van het framework is aandacht besteed aan algemene toepasbaarheid, dat wil zeggen dat het framework ingezet kan worden voor elke organisatie die bezig is met een implementatie in de context van Responsible AI. Het framework is tijdens de multiple case studie

voor beide case organisaties succesvol toegepast, wat de gedachte van de algemene toepasbaarheid bevestigt.

In het risico framework staan de vijf beginselen centraal. De volledigheid van deze beginselen is in de praktijk getest door de risico's van twee case organisaties te inventariseren en deze te toetsen aan de definitie van Responsible AI. Alle risico's konden worden gecategoriseerd binnen ten minste één van de vijf beginselen. Er is wel veel overlap gevonden in de vorm van risico's die bij meerdere categorieën tegelijk passen. Deze overlap tussen de bevindingen is logisch bevonden.

Wat tevens uit de case studies is gebleken, is dat de interpretatie van het framework kan verschillen omdat een organisatie met haar AI implementatie zeer specifieke risico's zou kunnen hebben. De verschillende specifieke risico's kunnen daarmee anders zijn, maar alsnog onder dezelfde hoofdcategorisering vallen. Andere risico's kunnen daarmee vanuit het framework geïdentificeerd worden voor organisatie, op basis van dezelfde globale omschrijving van het risico, zonder dat deze interpretatie in strijd hoeft te zijn met de theorie. Een andere interpretatie hoeft dus niet foutief te zijn. Tijdens de multiple case studie is in meerdere gevallen gebleken dat een risico of een beheersmaatregel te specifiek was beschreven, en daarmee niet algemeen toepasbaar was. Dit is tijdens de case studies naar voren gekomen en hierop aangepast in het gevalideerde framework.

Een andere limitatie van het RAI risico framework, en de algemene toepasbaarheid hiervan, is dat de mate van impact van een risico in niet wordt meegenomen. De impact kan namelijk verschillen voor iedere toepassing. Alle gevonden risico's worden op eenzelfde manier gepresenteerd, terwijl dat niet hoeft te betekenen dat de impact van ieder risico, voor iedere AI implementatie, even impactvolle gevolgen heeft. De impact van een risico is sterk afhankelijk van de details van de specifieke AI implementatie. Er kan bij het volgen van het RAI risico framework geen waardeoordeel worden gegeven aan de mate van risico dat een organisatie loopt indien niet wordt voldaan aan één of meer van de beginselen. In de praktijk betekent dit dat een organisatie zelf een inschatting zal moeten maken van het belang dat een organisatie moet hechten aan een specifiek risico of een beheersmaatregel.

Het is ook goed mogelijk dat bepaalde risico's nog niet zijn opgenomen omdat deze zijn gemist in het onderzoek. Ook kunnen er beheersmaatregelen zijn die niet in het model zijn opgenomen, voor risico's die al wel zijn opgenomen in het framework. Tijdens de case studies zijn in verhouding, overduidelijk meer aanpassingen aan de beheersmaatregelen naar voren gekomen. Dit kan te maken hebben met het feit dat in de praktijk veel nagedacht dient te worden over bepaalde problemen om tot een beheersbaar product te komen.

Na praktijkonderzoek om de theorie in het framework te valideren, is vastgesteld dat er binnen alle beginselen van het RAI framework gelijkenissen te vinden zijn tussen het theoretische kader, en de bevindingen in de case studies. Op basis van de multiple case studie, waarin de sample size als acceptabel wordt gezien in het kader van de betrouwbaarheid en validiteit, is het RAI risico framework gevalideerd in de praktijk.

5.2. Conclusies

Om de hoofdvraag 'Hoe kan een framework worden vormgegeven dat gebruikt kan worden om met risico's om te gaan in de context van Responsible AI?' te kunnen beantwoorden, en daarmee tot een framework te komen dat gebruikt kan worden om met risico's om te gaan in de context van Responsible AI, is allereerst theorie verzameld. In de bestaande theorie is in hoofdlijnen duidelijk wat RAI is, maar van een uniforme gebruikte definitie is geen sprake. Hierom is het van belang geweest om op basis van de bekende definities vanuit de theorie, tot een eigen definitie te komen die wordt gehanteerd in het kader van dit onderzoek. Deze definitie is in paragraaf 2.2.1 vastgesteld op 'Een verantwoorde manier van implementeren van AI binnen organisaties in de vorm van een governance framework'.

Met in acht neming van deze definitie, is het mogelijk geweest om risico's te verzamelen die voldoen aan de definitie van RAI. Om tot een duidelijke structuur te kunnen komen is een uitsplitsing gemaakt van verschillende beginselen die met RAI te maken hebben. Deze beginselen zijn tot stand gekomen op basis van de theorie. De vijf beginselen bestaan uit: ethiek, verantwoordelijkheid, aansprakelijkheid, privacy & veilig en uitlegbaarheid.

De onderverdeling in vijf beginselen staat centraal in het RAI risico framework als een soort kapstok om risico's van AI die in wetenschappelijke bronnen naar voren komen, op te hangen. Voor ieder beginsel is vervolgens een eigen stuk theoretisch kader opgesteld. Op basis van het theoretische kader zijn risico's en beheersmaatregelen in het RAI risico framework opgenomen. Hieronder zijn enkel de belangrijkste conclusies per beginsel opgenomen.

Op ethisch gebied is geconcludeerd dat er een reëel risico kan bestaan op schade aan de maatschappij door AI. Een maatregel om dit risico te mitigeren is om aandacht te hebben voor Corporate Social Responsibility en door ethische dilemma's bespreekbaar te maken. Een ander risico op ethisch vlak zijn de ethische kwesties die kunnen voortkomen wanneer morele kwesties worden uitbesteed aan een AI. Dit kan gematigd worden door de afweging te maken of morele kwesties überhaupt zouden moeten worden uitbesteed aan een AI. Als derde risico op het gebied van ethiek kan worden geconcludeerd dat het risico kan bestaan dat AI wordt ingezet voor ethisch onaanvaardbare doeleinden. Het maken van een juiste ethische afweging is hiervoor van belang als een beheersmaatregel. Een vierde risico is op basis van het empirisch risico gevonden, namelijk het risico op een biased getrainde dataset in relatie tot mensen. Dit risico kan worden gematigd door een divers palet aan trainingsdata te gebruiken en een agnostische aanpak te hanteren.

Op het gebied van verantwoordelijkheid kan worden geconcludeerd dat een mens onvoorspelbaar kan zijn voor een AI, wat negatieve gevolgen met zich mee kan brengen. Op basis van zowel de theorie als de case studies is duidelijk naar voren gekomen dat het risico's met zich mee kan brengen wanneer een mens als beslissingsorgaan wordt vervangen door een AI. Een samenwerking tussen mens en AI, waarbij met name de impactvolle beslissingen door de mens wordt gemaakt, is een manier om dit risico te mitigeren. In de literatuur en in beide case studies is naar voren gekomen dat een samenwerking tussen mens en AI zowel tot de beste resultaten leidt en een meer verantwoorde keuze is dan het vervangen van een mens door een AI.

Op het gebied van aansprakelijkheid is het voldoen aan wetgeving een belangrijk aspect voor organisaties. Beide case studies hebben heb belang aangegeven van de betrokkenheid van specialistische teams op het gebied van compliance, zodat compliance op een juiste manier in het design van een AI kan worden geïncorporeerd. Er is tevens naar voren gekomen dat het hanteren

van niet-wettelijke richtlijnen ook positieve invloed kan hebben op een verantwoord gebruik, ondanks dat de wet dit (nog) niet voorschrijft. Uit de case studies is helder naar voren gekomen dat deze niet-wettelijk standaarden niet zelf ontwikkeld hoeven te worden, maar dat standaarden al publiekelijk beschikbaar kunnen zijn. Een ander aspect van aansprakelijkheid is het vraagstuk wie er aansprakelijk is voor de acties van een AI, waarvan naar voren is gekomen dat dit niet altijd duidelijk hoeft te zijn voor organisaties. Vanuit de theorie zijn twee mogelijkheden geboden, namelijk de structuralist approach en de context-driven approach. In de case studies zijn beide mogelijkheden van toewijzing van aansprakelijkheid duidelijk naar voren gekomen.

Op het gebied van privacy & veiligheid is naar voren gekomen dat de complexiteit van data en privacy een risico is voor organisaties. Dit is door beide case organisaties onderstreept. De beheersmaatregel die naar voren zijn gekomen om dit risico te matigen zijn het incorporeren van data security en het nadenken over authenticatie en autorisatie. Een ander risico dat is gevonden in zowel de theorie als de praktijk, is het omgaan met persoonsgegevens. De noodzaak om compliant zijn met wetgeving zoals de AVG/GDPR is hier onderdeel van. Als belangrijkste beheersmaatregel is naar voren gekomen om privacy te waarborgen in het design, waarbij vanuit het empirisch onderzoek de aanvulling is gedaan dat de betrokkenheid van specialistische teams of personen op het gebied van deze wetgeving een risico mitigerende maatregel is.

Op het gebied van uitlegbaarheid is naar voren gekomen dat een AI model onbedoelde uitkomsten kan hebben als niet duidelijk is hoe het model werkt door de mate van complexiteit. Vanuit de case studies is naar voren gekomen dat logging, monitoring, feedback loops alsmede het blijven verbeteren van het model, belangrijk zijn om dit risico te mitigeren. Daarnaast is naar voren gekomen dat er in specifieke situaties een wettelijke verplichting, of een wens, kan zijn om de black box van AI te verklaren. Bepaalde tools en frameworks, waarbij een link kan worden gelegd met het domein van Explainable AI (XAI), kunnen hierbij helpen.

Met de vijf beginselen van RAI als basis is een framework gecreëerd dat fungeerde als een metaforische kapstok om risico's en beheersmaatregel aan te koppelen. Op basis van de bevindingen uit het literatuuronderzoek is dit framework gevuld. Door middel van de multiple case studie zijn de bevindingen in de praktijk gevalideerd. De hoofdvraag 'Hoe kan een framework worden vormgegeven dat gebruikt kan worden om met risico's om te gaan in de context van Responsible AI?' is hiermee beantwoord.

5.3. Aanbevelingen voor de praktijk

Het RAI risico framework is opgesteld met als doel om bruikbaar te zijn voor organisaties in de praktijk. Het framework kan als een hulpstuk fungeren om risico's te identificeren bij AI implementaties in de context van RAI, met behulp van de vijf beginselen. Ook wordt richting geboden in de vorm van beheersmaatregelen om met deze risico's om te gaan.

In de praktijk dient meegenomen te worden dat het framework algemeen toepasbaar is, wat betekent dat risico's en beheersmaatregelen breed te interpreteren kunnen zijn, en kunnen verschillen per toepassing. Een kritische houding is daarom nodig bij het gebruiken van het framework.

Daarnaast dient te worden opgemerkt dat alle beginselen en risico's in eenzelfde mate van impact en belang worden gepresenteerd. Dit hoeft in de praktijk niet te betekenen dat alle risico's even impactvol zijn voor de organisaties die het gebruikt. Deze impact verschilt per organisatie en dient daarom kritisch door de organisatie die het framework gebruikt, te worden beoordeeld.

Als laatste dient benadrukt te worden dat dit onderzoek geen volledigheid van het RAI risico framework garandeert. Het framework is opgesteld op basis van meerdere bronnen uit de literatuur en is getoetst aan de praktijk door middel van twee case studies. Dit betekent niet dat alle risico's die er kunnen bestaan allemaal zijn volledig zijn afgedekt door het framework. Het kan voorkomen dat er in de praktijk risico's spelen voor een organisatie die niet naar voren komen in het framework. Dit biedt echter weer kansen voor vervolgonderzoek.

Om bovengenoemde redenen is een kritische houding van belang om te zorgen dat het gebruik van het framework goed aansluit bij de AI implementatie waar deze voor wordt ingezet. Dit waarborgt de interpretatie, het afwegen van het belang van individuele risico's en de volledigheid.

5.4. Aanbevelingen voor verder onderzoek

In de conclusie is aangegeven dat een onderverdeling vijf beginselen centraal staat in het RAI risico framework, als een metaforische kapstok om risico's en beheersmaatregelen binnen RAI aan op te hangen. In verder onderzoek kunnen, zowel op inductieve als op deductieve wijze, nieuwe risico's of beheersmaatregelen worden gevonden om aan deze metaforische kapstok te hangen.

Middels een proces van toetsing van het framework met de praktijk, om vervolgens aanpassingen te maken of aanvullingen te doen, en deze opnieuw te toetsen met de praktijk, kan op iteratieve wijze een steeds volledigere, kwalitatief beter en betrouwbaarder framework tot stand komen.

Zoals aangegeven in de conclusie kan bij het volgen van het framework geen waardeoordeel worden gegeven aan de mate van impact van een risico voor een organisatie. Vervolgonderzoek kan op dit gebied wellicht een aanvulling doen aan het framework doen door de impact van verschillende risico's te bepalen. Dit kan een waardevolle toevoeging aan het framework zijn in de praktijk voor organisatie, omdat sturing op basis van de meest belangrijke risico's dan eenvoudiger wordt gemaakt.

Referenties

Accenture (2018). "RESPONSIBLE AI: A Framework for Building Trust in Your AI Solutions."

Anton Saveliev, D. Z. (2020). "Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China." Kybernetes **50**(3): 656-675.

Barredo Arrieta, A., et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion **58**: 82-115.

Clarke, R. (2019). "Principles and business processes for responsible AI." computer law & security **35**(4): 410-422.

Doorn (2021). "Artificial intelligence in the water domain: Opportunities for responsible use." Science of the Total Environment **755**.

European Commission (2016). "Rules for business and organisations."

European Commission (2018). Ethics Guidelines for Trustworthy AI, European Commission.

Future Of Life Institute (2017). "ASILOMAR AI PRINCIPLES." from https://futureoflife.org/ai-principles/.

Google (2021). "Explainable AI." Retrieved 31-5-2021, 2021.

Hatfield (2019). "Professionally Responsible Artificial Intelligence." Arizona State Law Journal **51**(3).

IBM Research (2021). "AI Explainability 360." Retrieved 31-5-2021, 2021.

Microsoft (2021). "Responsible AI - Microsoft AI principles." Retrieved 2021-04-21, 2021.

Nyholm, S. R. (2018). The ethics of crashes with self-driving cars: a roadmap I. Eindhoven University of Technology.

Robbins (2020). "AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines." AI & SOCIETY **35**(2): 391.

Saleema Amershi, et al. (2019). "Guidelines for Human-AI Interaction." CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Sara Suárez-Gonzalo, L. M.-M., Frederic Guerrero-Solé (2019). "Tay is you. The attribution of responsibility in the algorithmic culture." Observatorio **13**(2): 1-14.

Sullins, J. P. (2006). "When Is a Robot a Moral Agent? ." International Review of Information Ethics **6**.

Unesco (2018). "Towards a global code of ethics for artificial intelligence research." The Unesco Courier **3**.

Will Orr, J. L. D. (2019). "Attributions of ethical responsibility by Artificial Intelligence practitioners." Information, Communication & Society **23**(5): p719-735.

Wilson, H. J. D., Paul R. (2018). "Collaborative Intelligence: Humans and AI Are Joining Forces." Harvard Business Review.

Saunders (2019). Research methods for business students, Pearson Education Limited.

Bijlage 1 – Literatuur zoekstrategie

Om op een systematische wijze naar literatuur te zoeken voor het onderzoek, zijn vooraf kernwoorden gedefinieerd die zijn gebruikt voor het zoeken naar literatuur. Hieronder staat welke kernwoorden zijn gebruikt en hoeveel resultaten dit heeft opgeleverd. Ook zijn de gevonden artikelen opgenomen in tabel 1.

Wat is responsible AI?

Kernwoorden	**Extra filters**	**Aantal hits**
Responsible AI & Principles	-	8
Responsible AI & Framework	-	7
Responsible AI	Date = 2011-2021 Source = 'Academic Journals'	36
Responsible Artificial Intelligence	Date = 2011-2021 Source = 'Academic Journals'	13
Responsible use of AI		8
Responsibility & artificial intelligence & Business	Date = 2011-2021 Source = 'Academic Journals'	50

De zoekquery is uitgevoerd in april 2021. De uitkomst is te vinden in tabel 1.

De gevonden artikelen zijn in de onderstaande tabel weergegeven. Per artikel is tevens de relevantie voor het onderzoek opgenomen.

Artikel	**Auteur**	**Source**	**Relevantie**
Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.	Barredo Arrieta, Alejandro1 (AUTHOR) Díaz-Rodríguez, Natalia2 (AUTHOR) Del Ser, Javier1,3,4 (AUTHOR) javier.delser@tecnalia.com Bennetot, Adrien2,5,6 (AUTHOR) Tabik, Siham7 (AUTHOR) Barbado, Alberto8 (AUTHOR) Garcia, Salvador7 (AUTHOR) Gil-Lopez, Sergio1 (AUTHOR) Molina, Daniel7 (AUTHOR) Benjamins, Richard8 (AUTHOR) Chatila, Raja6 (AUTHOR) Herrera, Francisco7 (AUTHOR)	Information Fusion. Jun2020, Vol. 58, p82-115. 34p.	In de abstract wordt ingegaan op de frameworks Explainable AI, Responsible AI en de verhouding tussen de twee.
Principles and business processes for responsible AI.	Clarke, Roger	Computer Law & Security Review. Aug2019, Vol. 35 Issue 4, p410-422. 13p.	In de abstract staat '*This second article discusses how an organisation can manage **AI** responsibly, in order to protect its own interests, but also those of*

			its stakeholders and society as a whole.'
Regulatory alternatives for AI.	Clarke, Roger	Computer Law & Security Review. Aug2019, Vol. 35 Issue 4, p398-409. 12p.	Dit artikel gaat in op hoe AI kan worden ingezet op een verantwoordelijke manier.
AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines.	Robbins, Scott[1] (AUTHOR) *scott@scottrobbins.org*	AI & Society. Jun2020, Vol. 35 Issue 2, p391-400. 10p.	Artikel gaat in op hoe AI gereguleerd kan worden voor veilig gebruik in de huidige wereld.
Academics as leaders in the cancer artificial intelligence revolution.	Kochanny, Sara E.1 (AUTHOR) Pearson, Alexander T.1 (AUTHOR) *apearson5@medicine.bsd.uchicago.edu*	Cancer (0008543X). 3/1/2021, Vol. 127 Issue 5, p664-671. 8p. *(can't be found for free)*	Artikel gaat in op standarden voor responsible use van AI binnen de medische sector
Artificial intelligence in the water domain: Opportunities for responsible use.	Doorn, Neelke[1] (AUTHOR) *N.Doorn@tudelft.nl*	Science of the Total Environment. Feb2021:Part 1, Vol. 755, pN.PAG-N.PAG. 1p.	Gaat in op responsible AI in de water industrie
Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China.	Saveliev, Anton[1] (AUTHOR) *anton.saveliev@gmail.com* Zhurenkov, Denis[1] (AUTHOR) *dzhurenkoff@mail.ru*	Kybernetes. 2021, Vol. 50 Issue 3, p656-675. 20p.	Design en methodologie voor de inzet van AI in de context van CSR en responsible use.
Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability.	London, Alex John (AUTHOR)	Hastings Center Report. Jan2019, Vol. 49 Issue 1, p15-21. 7p.	In de abstract wordt ingegaan op AI in de medicijnen industrie. De link wordt gemaakt naar verantwoordelijkheid wanneer redenen voor decision-making niet altijd duidelijk zijn.
IT Governance Knowledge: From Repositories to Artificial Intelligence Solutions.	Chergui, Meriyem[1] *chergui.meriyem@gmail.com* Chakir, Aziza[2]	Journal of Engineering Science & Technology Review. 2020, Vol. 13 Issue 5, p67-76. 10p.	Artikel gaat in op het design en implementatie van AI systemen voor IT Governance.
New 'Corpus Juris' from Artificial Intelligence.	Dardani, Arianna Perotti[1] Costabel, Attilio M.[2]	Journal of Multidisciplinary Research (1947-2900). Spring2021,	Artikel gaat in op de 'legal' kant van AI. Dit kan zeker relevantie hebben met een Governance

		Vol. 13 Issue 1, p31-58. 28p.	framework voor verantwoorde implementatie.
Professionally Responsible Artificial Intelligence.	Hatfield, Michael1	Arizona State Law Journal. Fall2019, Vol. 51 Issue 3, p1057-1122. 66p.	Artikel gaat in op 'Professional responsibility' bij het gebruik van AI in de tax law sector
How Competition Law Should React in the Age of Big Data and Artificial Intelligence	Hayashi, Shuya[1] Arai, Koki[2] *koki.arai@nifty.ne.jp*	Antitrust Bulletin. Sep2019, Vol. 64 Issue 3, p447-456. 10p.	Artikel gaat onder andere in op de 'Responsibility distribution' van AI tussen verschillende stakeholders.
Is Watson for Oncology Unreasonably Dangerous?: Making A Case for How to Prove Products Liability Based on a Flawed Artificial Intelligence Design.	Frank, Xavier	American Journal of Law & Medicine. 2019, Vol. 45 Issue 2/3, p273-294. 22p. *(can't be found for free)*	Artikel gaat in op de risico's, en de gevolgen daarvan zoals onder andere aansprakelijkheid, van AI van een specifieke case.
Tay is you. The attribution of responsibility in the algorithmic culture.	Suárez-Gonzalo, Sara1 Mas-Manchón, Lluís1 Guerrero-Solé, Frederic1	Observatorio (OBS*). 2019, Vol. 13 Issue 2, p1-14. 14p.	Gaat onder andere in op vraag '*Who should be held responsible for non-human beings' actions, particularly when the consequences of these actions are negative?*'
Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs.	Dauvergne, Peter[1] (AUTHOR) *peter.dauvergne@ubc.ca*	Review of International Political Economy. Sep2020, p1-23. 23p.	De positieve impact van Corporate Social Responsilibity door de inzet van AI wordt in dit artikel in twijfel getrokken.
A Generalized Framework for Moral Dilemmas Involving Autonomous Vehicles: A Commentary on Gill.	Novak, Thomas P[1] (AUTHOR) *novak@gwu.edu*	Journal of Consumer Research. Aug2020, Vol. 47 Issue 2, p292-300. 9p. 1 Diagram.	Artikel gaat in op morele/ethische kant van AI. Specifiek bij het gebruik van Autonome voertuigen.

Tabel 1: resultaten literatuuronderzoek

Bijlage 2 – Verschillende invalshoeken van Responsible AI

Op basis van onderzoek is in tabel 2.1 weergegeven welke invulling verschillende bronnen geven aan het begrip 'Responsible AI'.

Barredo Arrieta, Díaz-Rodríguez et al.	Clarke	Accenture	Microsoft	Doorn
Eerlijkheid	Ethiek	Ethiek	Fairness	Justice & fairness
Explainability	Risicobeheersing	Transparantie	Reliability & safety	Responsibility
Verantwoordelijkheid		Verantwoordelijk gebruik	Privacy & security	Accountability
			Inclusiveness	Privacy
			Transparancy	Non-maleficense
			Accountability	

Op basis van tabel 1 kunnen in hoofdlijnen drie basisprincipes worden gedefinieerd welke als uitgangspunten voor dit onderzoek worden gebruikt, zijnde:

1. Ethiek
2. Verantwoordelijkheid
3. Transparantie

De link tussen de verschillende invalshoeken en de basisprincipes is in tabel 2.2 weergegeven

Basisprincipe	Barredo Arrieta, Díaz-Rodríguez et al.	Clarke	Accenture	Microsoft	Doorn
Ethiek	Eerlijkheid	Ethiek	Ethiek	Fairness; inclusiveness	Justice & fairness
Verantwoordelijkheid	Verantwoordelijkheid	Risicobeheersing	Verantwoordelijk gebruik	Reliability & safety;	Responsibility; Non-maleficense
Aansprakelijkheid				Accountability	Accountability
Privacy & security				privacy & security;	Privacy
Explainability	Explainability		Transparantie	Transparancy	

Bijlage 3 – Identificatie van risico's en beheersmaatregelen

In deze bijlage zijn de risico's en beheersmaatregelen van responsible AI geïdentificeerd en onderbouwd gebaseerd op de theorie van paragraaf 2.2.3. De bijlage fungeert als de documentatie van het risico framework dat is opgenomen in paragraaf 2.2.4. De nummers corresponderen tevens met de nummers in het risico framework.

Voor het aspect 'ethiek' is onderscheid te maken in de volgende risicofactoren met de daarbij passende beheersmaatregelen:

1. Volgens (Anton Saveliev, 2020) kan een AI mogelijk schade hebben op de maatschappij. Volgens (Unesco, 2018) is er nog geen dekkend legal framework om dit tegen te gaan. Aandacht voor Corporate Social Responsibility is hierom gewenst.
2. Volgens (Sullins, 2006) kan een AI voor morele kwesties komen te staan indien de AI interacties heeft die verwant zijn aan interacties met mensen. De AI zou volgens Sullins dan een 'Moral agent' zijn. Sullins geeft aan dat onder experts, de meningen zijn verdeeld of een AI überhaupt een een moral agent zou moeten zijn.
3. Volgens de (Future Of Life Institute, 2017) is er een risico dat een AI wordt ingezet voor ethisch onaanvaardbare doeleinden. Tijdens de Asilomar conferentie is daarom en set van AI Guidelines opgesteld om ethisch onaanvaardbaar handelen te kunnen vermijden.

Nummer	**Aspect**	**Risico**	**Beheersmaatregel**
E1	Ethiek	Maatschappelijke schade door AI	Aandacht voor Corporate Social Responsibility
E2	Ethiek	Morele kwesties uitbesteed aan de AI	Beoordelen of morele kwesties wel uit zouden moeten worden besteed aan een AI
E3	Ethiek	Inzet voor onethische doeleinden	Volgen AI Ethics guidelines conform de Asilomar conferentie 2017

Voor het aspect 'verantwoordelijkheid' is onderscheid te maken in de volgende risicofactoren:

1. Volgens (Saleema Amershi et al., 2019) kan, wanneer een AI interactie heeft met een mens, disruptie, verwarring of gevaar ontstaan door het onvoorspelbaar handelen van de AI. Dit risico kan beperkt worden door het gebruik van algemene toepasbare designrichtlijnen.
2. Volgens (Hatfield, 2019) kan het vervangen van mensen met expertise door een AI in een organisatie, een negatieve invloed hebben op de professionaliteit en betrouwbaarheid. (H. J. Wilson & Daugherty, 2018) heeft uit onderzoek geconcludeerd dat AI het beste resultaat behaalt wanneer deze wordt ingezet als aanvulling op de mens.
3. Volgens (Robbins, 2020) kan een AI schade aan de omgeving aanrichten wanneer de AI niet in een afgeschermde omgeving opereert. Deze schade kan worden beperkt door de AI te laten opereren in een afgeschermde omgeving.

Nummer	Aspect	Risico	Beheersmaatregel
V1	Verantwoordelijkheid	Disruptie, verwarring of gevaar door onvoorspelbare handelingen	Het gebruik van algemene toepasbare designrichtlijnen
V2	Verantwoordelijkheid	Negatieve gevolgen bij vervangen van mens (professionaliteit & betrouwbaarheid)	AI inzetten als aanvulling op de mens, niet als volledig autonome vervanging
V3	Verantwoordelijkheid	Schade aan de omgeving door AI	AI laten opereren in een afgeschermde (micro-)omgeving

Voor het aspect aansprakelijkheid is onderscheid te maken in de volgende risicofactoren:

1. Volgens (Will Orr, 2019) is het niet altijd transparant wie aansprakelijk is op het gebied van compliance en ethiek verantwoordelijk in het geval dat een AI niet correct handelt. Van belang in zulke situaties is dat aantoonbaar de juiste veiligheidsanalyse wordt gemaakt en het juiste frameworks wordt toegepast. Daarnaast kan vooraf nagedacht worden over dit vraagstuk. Mogelijkheden zijn volgens (Sara Suárez-Gonzalo, 2019) onder andere de Structuralist approach of de context-dependent approach.
2. Volgens (Clarke, 2019) is een effectieve vorm van regulatie noodzakelijk voor het verantwoord kunnen benutten van AI. Volgens (Will Orr, 2019) is de regulatie nu te los en niet voldoende meegegaan met de groei van het gebied van AI, waardoor veel organisaties eigen standaarden ontwikkelen
3. Het is verplicht om als bedrijf te voldoen aan geldende wetgeving. Volgens de (European Commission, 2016) is het bijvoorbeeld verplicht voor organisaties om aan de GDPR wetgeving te voldoen. Op het gebied van compliance is dit potentieel een risico. Volgens (Will Orr, 2019) dient over compliance (en ethics) nagedacht te worden in het design.

Nummer	Aspect	Risico	Beheersmaatregel
A1	Aansprakelijkheid	Onduidelijkheid m.b.t. aansprakelijk op het gebied van ethiek en compliance	Veiligheidsanalyses, juiste frameworks hanteren & nadenken over toewijzing aansprakelijkheid
A2	Aansprakelijkheid	Onverantwoord gebruik van AI door gebrek aan publieke regulatie	Zelf definiëren en hanteren van een set van standaarden
A3	Aansprakelijkheid	Niet compliant zijn met wetgeving	In het design nadenken over compliance (en ethics)

Voor het aspect privacy & veiligheid is onderscheid te maken in de volgende risicofactoren:

1. Volgens (Unesco, 2018) bestaat het risico dat AI wordt ingezet voor doeleinden die de privacy van mensen schendt. Er zijn AI guidelines opgesteld tijdens de Asilomar Conferentie (Future Of Life Institute, 2017) om dit tegen te gaan.
2. Volgens (Microsoft, 2021) brengt AI extra complexiteit mee op het gebied van privacy. Hiermee kan omgegaan worden door na te denken over datastromen, waar modellen gedraaid worden, hoe data uit kan lekken en hoe het gemanipuleerd kan worden. Kortom, nadenken over de verschillende aspecten en risico's van data.
3. Volgens (Accenture, 2018) is het van belang om onder andere privacy in het design van de AI te incorporeren. Hier kunnen volgens Accenture aspecten bij komen kijken zoals compliance met de GDPR wetgeving (European Commission, 2016) en best-practises op het gebied van toegang en beheer van persoonsgegevens door individuen. Het onjuist omgaan met persoonsgegeven vormt potentieel een risico.

Nummer	Aspect	Risico	Beheersmaatregel
P1	Privacy & veiligheid	Inzet AI voor doeleinden die onethisch zijn op gebied van privacy	Volgen van AI guidelines conform de Asilomar conferentie 2017
P2	Privacy & veiligheid	Complexiteit op het gebied van privacy	Nadenken over risico's van data (datastromen, locaties, mogelijkheid datalekken/manipulatie)
P3	Privacy & veiligheid	Onjuist omgaan met persoonsgegevens	Privacy van persoonsgegevens waarborgen in het design en toepassen van best-practises

Voor het aspect uitlegbaarheid is onderscheid te maken in de volgende risicofactoren:

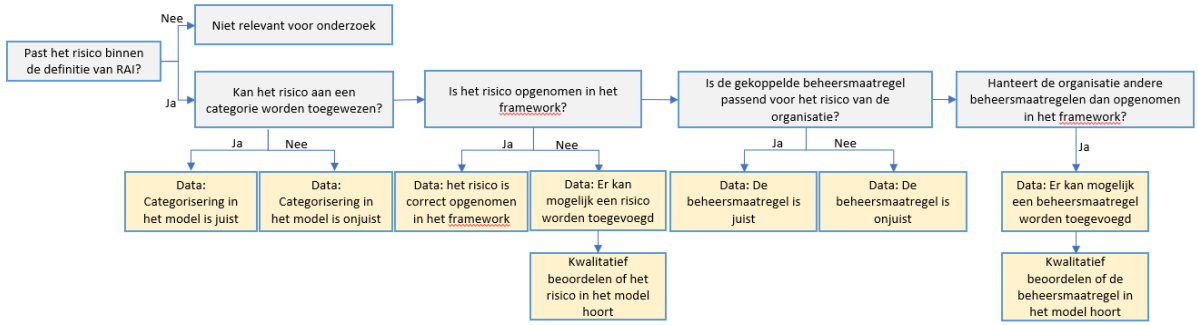
1. Volgens (Robbins, 2020) en (Accenture, 2018) is het nodig om te snappen hoe AI algoritmen werken, specifiek op gebied van training data, inputs, outputs en functies, zodat verantwoorde keuzes met betrekking tot regulatie kunnen worden gemaakt. Gebrek aan begrip en transparantie over de werking is daarmee een risico voor regulatie van de AI.
2. Volgens Accenture is het in bepaalde situaties volgens de wet verplicht om te kunnen uitleggen waarom tot een bepaalde keuze is gekomen, waar een keuze door een AI dus ook onder valt. Er bestaat dus potentieel risico van niet compliant zijn indien het decision-making proces niet voldoende kan worden uitgelegd.
3. Volgens (Barredo Arrieta et al., 2020) is de mogelijkheid tot het begrijpen en uitleggen van de keuzes van een AI van cruciaal belang wanneer een AI invloed heeft op mensenlevens. Specifiek worden de vakgebieden medicijnen, recht en defensie genoemd. Aanvullend is volgens de AI Guidelines opgesteld tijdens de Asilomar conferentie (Future Of Life Institute, 2017) het op het gebied van ethiek belangrijk om te snappen waarom een AI op een bepaalde manier gehandeld heeft in het geval dat er schade is aangericht door de AI en zou het daarnaast voor een competent mens altijd mogelijk moeten zijn om te achterhalen waarom een beslissing door een autonoom systeem is gemaakt. Hierdoor kan worden geconcludeerd dat gebrek aan uitlegbaarheid een risico vormt m.b.t. ethiek.

Nummer	Aspect	Risico	Beheersmaatregel
U1	Uitlegbaarheid	Complexiteit in regulatie van AI bij gebrek aan begrip van werking	Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)
U2	Uitlegbaarheid	Wettelijke verplichting tot inzicht in decision-making proces	Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)
U3	Uitlegbaarheid	Vereisten om decision-making uit te leggen. Essentieel bij invloed op mensenlevens	Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)

Bijlage 4 – opzet af te nemen interviews

De opzet van de interviews in de kwalitatieve studie bestaat uit zes chronologische stappen:

1) De context van de organisatie en de rol van de contactpersoon zijn belangrijk. Informatie en achtergrond met betrekking tot de specifieke AI implementatie wordt ook besproken.
2) Naar verwachting heeft de kennis van de geïnterviewde over het begrip Responsible AI een positieve invloed op de relevantie van de antwoorden. Om deze reden wordt de definitie van responsible AI toegelicht aan de case organisatie.
3) Voordat wordt ingegaan op het risico framework wordt de geïnterviewde gevraagd om de risico's van de AI toepassing, die te maken hebben met responsible AI, in de case organisatie te noemen. Dit wordt gedaan om te voorkomen dat teveel gestuurd wordt op de vijf beginselen uit het risico framework.
4) De specifieke risico's in het risico framework worden aan de case organisatie voorgelegd aan, waarbij nadrukkelijk onderscheid wordt gemaakt tussen de risico's binnen de vijf beginselen gedefinieerd in paragraaf 2.2.2.
5) Met behulp van een standaard vragenlijst waarbij iedere beginsel en elk risico belicht wordt, zal onderzocht worden welke mogelijke risico's van toepassing zijn voor de organisatie. Voor ieder risico zal worden geïnventariseerd welke beheersmaatregelen worden gebruikt om de gevonden risico's af te dekken. Ook wordt gekeken of voor de organisatie risico's spelen die niet zijn afgedekt in het model. Om structuur te geven aan deze stap en uniformiteit tussen de verschillende interviews te kunnen behalen wordt de onderstaande structuur gevolgd.



Afbeelding 4.1: beslisboom interview

6) De gegeven antwoorden en getrokken conclusies worden met de geïnterviewde besproken. In het geval dat er bepaalde informatie is gemist, of ter gedachte komt bij de interviewer of geïnterviewde, is er ruimte om nog te herzien. Dit is tevens de afronding van het interview.

De volgende vragen zouden beantwoord moeten kunnen worden na het interview:

Vraag 1: Is de organisatie, en de geïnterviewde, daadwerkelijk bezig geweest met een AI implementatie die relevant is binnen de definitie van Responsible AI?

Beantwoord door: Stap 1 (van bijlage 4)

Deze eerste vraag, met betrekking tot het beoordelen van de achtergrond van de organisatie en de geïnterviewden, is al grotendeels afgedekt door vooronderzoek in het vinden van de juiste persoon voor het interview. Echter zal tijdens de case studie toch nog gecheckt worden om zeker te zijn dat de bevindingen uit het interview bruikbaar zijn.

Vraag 2: Spelen er risico's bij de organisatie welke niet in het AI Framework aanwezig zijn opgenomen?

Beantwoord door: Stap 3 en stap 5

Deze tweede vraag, om te kijken in hoeverre de vijf beginselen voldoen, is ook ter validatie waarbij het antwoord 'ja' zal voldoen. Indien het antwoord 'nee, want..' is zal dit leiden tot het her evalueren van de beginselen van het model.

Vraag 3: Zijn de 5 beginselen van het AI risico framework voldoende dekkend voor de organisatie op het gebied van Responsible AI?

Beantwoord door: Stap 3

Vraag 4: In hoeverre sluiten de risico's die voor de organisatie van toepassing zijn, aan met het AI risico framework?

Beantwoord door: Stap 5; door middel van de decision tree

Vraag 5: In hoeverre sluiten de beheersmaatregelen die de organisatie hanteert, aan met het AI risico framework?

Beantwoord door: Stap 5; door middel van de decision tree

De resultaten zullen in de onderstaande tabel worden ingevuld. Elk beginsel, en elk risico, zal dus een eigen rij krijgen in de onderstaande tabel.

Initiële Risico/beheersmaatregel	**Risico(s) gevonden d.m.v. case studies**	**Beheersmaatregel gevonden d.m.v. case studies**
Risico en beheersmaatregel volgens het framework. **Voorbeeld**: **E1 - Ethiek** **RE1:** Maatschappelijke schade door AI **BE1:** Aandacht voor Corporate Social Responsibility	Het risico dat voor de case organisatie speelt. **Voorbeeld**: In de case organisatie speelt maatschappelijke schade door AI precies op de wijze dat is beschreven in het risico framework.	De beheersmaatregel toegepast door de case organisatie **Voorbeeld**: Om maatschappelijke schade te voorkomen gebruikt de case organisatie …… als beheersmaatregel.

Tabel 4.1: De tabel die wordt gehanteerd voor de gegevensanalyse

Voor ieder van de 5 beginselen zal vervolgens worden beoordeeld in hoeverre de case organisatie aansluit met het responsible AI framework. Hier zal als conclusie uit moeten komen welke risico's en welke beheersmaatregelen aansluiten met de praktijk en welke aanvullingen of aanpassingen er gemaakt zouden kunnen worden op basis van de uitkomst van de case studie. Dit kan ook betrekken hebben op de documentatie van het model.

Bijlage 5 – Tabel onderzoeksresultaten

De gegevens van de case studies zijn samengevoegd in één tabel om het vergelijken van de data mogelijk te maken. De aanduiding welke data uit welke casestudie afkomstig is, wordt als volgt gemaakt:

[CS-1] – Betreft case studie 1 – De hypotheekverstrekker

[CS-2] – Betreft case studie 2 – De ontwikkelaar van de slimme camera

De resultaten zijn onderverdeeld conform de risico's en beheersmaatregelen in het RAI risico framework.

Ethiek

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
E1 - Ethiek **RE1:** Maatschappelijke schade door AI **BE1:** Aandacht voor Corporate Social Responsibility	[CS1] Maatschappelijke schade als mensen onterecht geen, of juist een te hoge, hypotheek krijgen. [CS2] Applicaties die potentieel schade aan de maatschappij kunnen brengen bij verkeerd gebruik, zoals smart camera's in publieke ruimten.	[CS1] Dit wordt afgedekt door de menselijke check wanneer informatie op verschillende documenten niet matcht. Tevens zijn duidelijke toetsingskaders van belang. Ook moet er worden voldaan aan de vereiste op ethisch en sociaal gebied, wat raakvlakken heeft met 'Corporate Social Responsibility' [CS2] Discussies aangaan over de ethische dilemma's. Gevoelige applicaties kunnen bijv. voor openbare ruimten worden vermeden, maar wel voor een lokale groepen worden gebruikt waar geen ethische grenzen worden overschreden.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
E2 - Ethiek **RE2:** Morele kwesties uitbesteed aan de AI **BE2:** Beoordelen of morele kwesties wel uit zouden moeten worden besteed aan een AI	[CS1] Er worden geen morele kwesties uitbesteed [CS2] Morele kwesties zijn er niet aan de orde in de applicaties. De AI heeft een puur informerende rol.	[CS1] De AI is eigenlijk informatief. Het menselijke besluit is uiteindelijk de ultieme autoriteit. Het AI model doet in die zin alleen een suggestie. Maar uiteindelijk is het nooit het beslissingsorgaan [CS2] De AI wordt ingezet om te meten, en daarmee wordt een gebruiker geïnformeerd. 'Is het bureau vrij', 'Wordt afstand gehouden', 'er zijn X personen binnen'. De actie ligt uiteindelijk bij de persoon die hier zijn eigen interpretatie aan kan geven.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
E3 - Ethiek **RE3:** Inzet voor onethische doeleinden **BE3:** Volgen AI Ethics guidelines conform de Asilomar conferentie 2017	[CS1] Er is risico op onethische impact, maar het doel wordt door de case organisatie niet als onethisch gezien. [CS2] Er zijn applicaties die een ethisch dilemma voortbrengen, namelijk een slimme 1.5 meter sensor. Is het wel/niet ethisch.	[CS1] Er spelen concreet vastgestelde toetsingskaders om onethische impact zoveel mogelijk te voorkomen. [CS2] Ten eerste: Een gedeelte van de verantwoordelijkheid wordt bij de afnemer neergelegd. Rekening houden met wensen van afnemers. Ten tweede: Ethische afwegingen maken met de opties die er zijn, zoals geen waanzinnig alarm af laten gaan wanneer bij 2 personen de 1.5 meter wordt geschonden, maar achteraf op een onpersoonlijke manier kijken ‘hoe veilig was de ruimte vandaag’. Waarop je kunt sturen ‘wellicht moeten er meer developers thuis werken want die ruimte was niet veilig’

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
E4 - Ethiek: Aanvullende bevindingen, niet in model opgenomen:	[CS1] Ongewenste side effects, misschien iemand met een donkere huidskleur die in het contrast wegvalt, minder hoge succes rate heeft bij de paspoorten van donkere mensen. waardoor hun hypotheekaanvraag langer duurt. (een biased dataset of model)	[CS1] meer agnostische aanpak. Dat je dus meerdere features meeneemt, zoals de contouren van een pasfoto. een zo divers mogelijke test sample. Een zo divers mogelijk pallet aan input zodat ook in situaties die net afwijken, er misschien wel enkele herkenningspunten zijn. Dit zorgt ervoor dat je het model zo min mogelijk biased traint.

Verantwoordelijkheid

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
V1 – Verantwoordelijkheid **RV1:** Disruptie, verwarring of gevaar door onvoorspelbare handelingen **BV1:** Het gebruik van algemene toepasbare designrichtlijnen	[CS1] Dit is minder van toepassing. Er zit weinig controversie in de output. [CS2] Het risico wanneer het fout gaat, is vrij beperkt. ; Onvoorspelbare handelingen van mensen kan wel leiden tot foutieve uitkomsten, die mogelijk gevaarlijk zijn, of tot verwarring kunnen leiden. Bijvoorbeeld als mensen met een paraplu een gebouw binnenkomen, dan worden ze wellicht niet gezien, kan dat leiden tot te volle gebouwen wat tijdens de pandemie niet veilig is. Of tot verwarring bij een ontruiming. ; Bij de uitkomst van een informatieve AI is het de interpretatie van de uitkomst die het risico vormt, niet de uitkomst zelf.	[CS2] Transparant zijn over de succes rate van de modellen, op een juiste manier aangeven wat de uitkomst een richtlijn is en de uitkomst van niet als een harde waarheid te zien. Laat de beslissingen over aan de mens, niet aan de AI.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
V2 – Verantwoordelijkheid **RV2:** Negatieve gevolgen bij vervangen van mens (professionaliteit & betrouwbaarheid) **BV2:** AI inzetten als aanvulling op de mens, niet als volledig autonome vervanging	[CS1] Het risico dat het model tot een onjuiste uitkomst komt, waardoor iemand onterecht geen hypotheek krijgt toegewezen. [CS2] Wanneer de uitkomst van de AI wordt ingezet om beslissingen te maken dan kan dat in het geval van de smart camera leiden tot gevaarlijke situaties. Denk bijvoorbeeld bij een maximaal toegestaan aantal personen in een ruimte vanwege Corona. Als de telling van het model onjuist is, dan kan dat ertoe leiden dat er teveel mensen in het gebouw zijn wat gevaarlijke situaties oplevert.	[CS1] De AI is eigenlijk informatief. Het menselijke besluit is uiteindelijk de ultieme autoriteit. Het AI model doet in die zin alleen een suggestie. Maar uiteindelijk is het nooit het beslissingsorgaan. [CS2] De AI is geen vervanging van de mens, de AI kan dingen die een mens niet kan doen zoals 1.5 meter afstand meten. Een mens heeft geen laserogen dus zou dat niet kunnen doen. Bij beslissingen wordt dit risicovol. De AI inzetten op een informatieve wijze dekt dit af. in de case studie wordt dat gedaan door middel van een dashboard, waar duidelijk wordt gezet dat het om een indicatie gaat. Er wordt geen beslissing gemaakt door de AI zelf, dit wordt altijd door een mens gedaan.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
V3 – Verantwoordelijkheid **RV3:** Schade aan de omgeving door AI **BV3:** AI laten opereren in een afgeschermde (micro-) omgeving	[CS1] Disruptie voor de interne organisatie. De AI is in grote delen bedoeld ter ondersteuning bij de mensen die er al werken. Wat die eigenlijk feitelijk voor hen kan betekenen is dat er een applicatie komt die een deel van hun werk gaat vervangen. Voor die werknemers hangt daar natuurlijk een bepaald sentiment bij. Dat kan disruptief voor hun leven en/of voor de continuiteit van hun carriere. [CS2] Schade aan de omgeving is niet van toepassing.	[CS1] dat je mensen niet het gevoel moet geven dat zij vervangen gaan worden door een stukje software.; Waar je de nadruk vooral op moet leggen is dat het echt de bedoeling moet zijn dat het hun werk makkelijker maakt, waardoor zij meer tijd overhouden voor taken die wat uitdagender zijn. [CS1] De Ai opereet in een gecontroleerde omgeving, wat bijvoorbeeld bij chatbots niet altijd het geval is. In onze applicaties zijn overal kaders aangebracht om risico's te voorkomen. [CS2] Alle berekeningen in het model worden in het device zelf gedaan. De camera beweegt verder niet, en heeft geen manier om de buitenwereld te beïnvloeden.

Aansprakelijkheid

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
A1 - Aansprakelijkheid **RA1** Onduidelijkheid m.b.t. aansprakelijk op het gebied van ethiek en compliance **BE1** Veiligheidsanalyses, juiste frameworks hanteren & nadenken over toewijzing aansprakelijkheid	[CS1] Wat dit voor ons ook inhield is dat als de intermediair een paspoort uploadt, dat zij ook al geen BSN nummer hadden mogen sturen. Hier zie je dus eigenlijk een soort shift in aansprakelijkheid. Want wie er verantwoordelijk. De intermediair krijgt het document van de klant. Hier zou de intermediair al tegen de klant kunnen zeggen, je hebt je BSN niet doorgestreept. Als de intermediair dit wel doorstuurt naar het AI model, zou je hier kunnen zeggen dat hij dit had kunnen voorkomen door dit zelf te checken. [CS2] Er is nagedacht over de aansprakelijkheid, de gebruiker van de camera is aansprakelijk voor hoe het wordt ingezet, en alle gevolgen ervan, ook op ethisch gebied. Als de output foutief is en daar acties aan worden verbonden ligt dit risico ook bij de gebruiker. Voor de werking van het product zelf heeft de designer een bepaalde aansprakelijkheid, maar dat is niet gericht op het AI model maar meer op het fysieke product zelf.	[CS1] In ons geval is dat in de meeste gevallen heel duidelijk gedefinieerd. De aansprakelijkheid ligt bij de maker van het model. Wij hebben ervoor gekozen om te intermediair te informeren in dat geval. Maar zij mogen ook geen BSN verwerken. Wij hebben ook een defensieve aanpak gekozen om aan de hand van computer vision te herkennen of er een BSN in staat ja of nee.; preventieve defensieve maatregel om te zorgen dat wij de aansprakelijkheid niet hebben. [CS2] Er worden de juiste veiligheidsanalyses gedaan, en er is nagedacht over het aansprakelijkheid vraagstuk.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
A2 - Aansprakelijkheid **RA2** Onverantwoord gebruik van AI door gebrek aan publieke regulatie **BE2** Zelf definiëren en hanteren van een set van standaarden	[CS1] De focus ligt voornamelijk op wat voor de wet wel- of niet mag. Daar wordt reactief op gereageerd, bijvoorbeeld wanneer de wet vanaf een bepaalde datum besluit om verwerking van BSN nummers te verbieden. [CS2] De wet schrijft minder voor, dan dat het bedrijf belangrijk vindt voor kwaliteitswaarborging. Dat is grotendeels preventief voor onverantwoord gebruik.	[CS1] We hanteren wel optionele monitoring, die wel noodzakelijk zijn, maar dit niet wettelijk verplicht zijn. Dit voorkomt wellicht wel dat er ooit is gebeurt dat tegen de wet in gaat. Maar voor zover ik weet hebben we geen vaste lijst met richtlijnen. [CS2] Er wordt gezocht voor kwaliteitswaarborging, veel strenger dan dat de wet voorschrijft. Hierbij wordt rekening gehouden met niet-wettelijke normen, zoals de ISO-27001. Ook zijn er CE-certificeringen. Er worden ook eigen ‘extra stappen’ genomen zoals het on-the-device draaien. Voor veiligheid wordt graag een stapje extra gezet.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
A3 - Aansprakelijkheid **RA3** Niet compliant zijn met wetgeving **BE3** In het design nadenken over compliance (en ethics)	[CS1] Wat bijvoorbeeld in 2021 relevant werd is dat banken geen BSN nummers mogen verwerken. Dit is wetgeving die veranderd was. [CS2] Er is een risico op het breken van de privacy wetgeving.	[CS1] Je kan er natuurlijk wel op anticiperen maar het is heel reactief op veranderende wetgeving. Er is nagedacht over de manier hoe omgegaan wordt met deze compliance, in het geval de BSN nummer is gekozen om de AI hier proactief naar te laten zoeken in het design. [CS1] Wat hier ook aan aanvulling voor is, is de betrokkenheid van een legal team. Want moet de developer gaan bepalen of de applicatie aan de wet voldoet. Daar heb je ook een specialistisch team voor nodig. Een Legal team. Eigenlijk zouden zij het voortouw moeten nemen en het development team informeren. [CS2] In het design zit dit aardig dicht getimmerd. De verantwoordelijkheid ligt bij de product owner, die de wetgeving precies dient te kennen.

Privacy & veiligheid

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
P1 - Privacy & Veiligheid **RP1**: Inzet AI voor doeleinden die onethisch zijn op gebied van privacy **BP1**: Volgen van AI guidelines conform de Asilomar conferentie 2017	[CS1] Als er een document wordt geüpload waar klantdata in staat, gaat de AI op zoek naar die klantdata. Dat is niet onethisch, maar het is heel zakelijk. [CS2] Inzet van een camera die voor mensen op ethisch gebied ongewenst kan zijn. Sommige mensen willen niet gefilmd worden, of in de gaten worden gehouden of ze wel 1.5 meter afstand houden. De link naar het gebruik van een smart camera in China is ook gemaakt, waar ditzelfde wellicht wel wordt geschonden. Het is een bewuste keuze geweest om het eigen product zo onpersoonlijk mogelijk te maken.	[CS1] Er is nagedacht over de risico's van de data waarbij gekeken wordt naar data stromen. Wie kan er bij en waar sla je dat op, op welke locaties. Waarbij dus ook het risico op datalekken wordt beperkt en de manipulatie van data wordt voorkomen. [CS2] De afnemer wordt verantwoordelijk gehouden voor de wijze van inzet. Daarnaast wordt zelf een ethische afweging gemaakt, zonder gebruik gemaakt te hebben van de AI guidelines. Een bewuste keuze is geweest om het eigen product zo onpersoonlijk mogelijk te houden.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
P2 - Privacy & Veiligheid **RP2**: Complexiteit op het gebied van privacy **BP2**: Nadenken over risico's van data (datastromen, locaties, mogelijkheid datalekken/manipulatie)	[CS1] Datalekken; dat iedereen zomaar bij je data kan. Het risico op data breaches is enorm groot en kan gigantisch complex worden om te herkennen, denk aan de recente ontwikkelingen in de cyber security space die te maken hebben met Log4J. omgaan met persoonsgegevens, dat op het gebied van privacy veel GDPR komt kijken [CS2] Risico bij het data innemen, afbeeldingen verwerken waar personen op staan, afbeeldingen van de omgeving.; risico op hacks Risico m.b.t. wie toegang tot de trainingsdata, en de live camera's heeft	[CS1] authenticatie en autorisatie protocollen Identity Access Management; Alleen via het intranet toegankelijk. Er zijn teams in de plaats om dit soort zaken preventief te herkennen omdat het zo complex kan worden. software engineer en cyber security zijn gewoon hele andere takken van sport. Het development team kijkt eigenlijk altijd vanuit het oogpunt van een business user. Het cyber security team kijkt vanuit het oogpunt van een malafide gebruiker. Zij kijken heel erg 'het interessert mij niets wat de software functioneel doet. Maar hoe kan ik binnendringen.' Zij gaan gegarandeerd zaken vinden die een development team niet gaat vinden. [CS2] Beveiligde verbindingen; coderen van data; Verwerking van data op het device zelf draaien; Nadenken wie er bij de data moet kunnen. Juiste risico afwegingen maken.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
P3 - Privacy & Veiligheid **RP3**: Onjuist omgaan met persoonsgegevens **BP3**: Privacy van persoonsgegevens waarborgen in het design en toepassen van best-practises	[CS1] Werken met data met de hoogst persoonlijke classificatie. Gevoelige data als input gebruiken om te trainen [CS2] Er wordt gewerkt met persoonlijke data van de klant (camera beeldingen/reeksen foto's).	[CS1] Betrokkenheid Legal afdeling voor formele goedkeuring. Wat je altijd doet is dat je het voorstel toetst aan een legal team. Opnemen van de relevante GDPR wetgeving in het design. [CS2] Hanteren van beleidsdocumenten zoals de ISO-27001 standaarden, die geven precies aan hoe je ermee om moet gaan. De verantwoordelijkheid ligt bij de product owner, die de wetgeving precies dient te kennen. Die moet ervoor zorgen dat de developer het juist integreert in het design. Verwerking van persoonsgegevens wordt behoorlijk dicht afgedekt in het design.

Uitlegbaarheid

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
U1 - Uitlegbaarheid **RU1**: Complexiteit in regulatie van AI bij gebrek aan begrip van werking **BU1**: Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)	[CS1] Het risico is dus dat je blind runt.	[CS1] We loggen al die zaken die de intermediair door stuurt, zodat je kunt zien hoe vaak het wel wordt aangeleverd, en koppelen dat ook terug. Zo'n feedback loop is enorm belangrijk zodat je kunt blijven verbeteren. Je draagt bij dat het steeds dichter bij een nauwkeurig antwoord komt. Het wordt steeds duidelijker en accurater. Aanvullend, als het uitlegbaar is waarom tot een keuze is gekomen, dan maakt het de werking wel een stuk fijner voor de intermediair [CS2] De designer zegt zelf niet te weten hoe je de black box uit kan leggen, zonder dit met behulp van tools en software te doen. Wat overeenkomt met de beheersmaatregel.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
U2 - Uitlegbaarheid **RU2:** Wettelijke verplichting tot inzicht in decision-making proces **BU2:** Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)	[CS1] Er is geen wettelijke verplichting tot inzicht geven in decision making. Maar wel een gewenste. Er kan een discrepantie ontstaan tussen wat het model als acceptabel ziet, en een mens/intermediair. [CS2] De organisatie hoeft het decision making proces niet uit te leggen. Het voegt ook niets toe.	[CS1] De gebruiker van goede feedback voorzien over de input. Als het uitlegbaar is waarom tot een keuze is gekomen, dan maakt het de werking wel een stuk fijner voor de intermediair, want het doel is ook om de intermediair te helpen en het proces te versnellen. De uiteindelijke keuze om het toch door te zetten ligt uiteindelijk altijd bij de intermediair, het model gaat dat niet overrulen. [CS2] De designer zegt zelf niet te weten hoe je de black box uit kan leggen, zonder dit met behulp van tools en software te doen. Wat overeenkomt met de beheersmaatregel.

Initiële Risico/beheersmaatregel	Risico(s) gevonden d.m.v. case studies	Beheersmaatregel gevonden d.m.v. case studies
U3 - Uitlegbaarheid **RU3:** Vereisten om decision-making uit te leggen. Essentieel bij invloed op mensenlevens **BU3:** Gebruik van tools en frameworks om AI te kunnen verklaren en begrijpen (XAI)	[CS2] Naast dat de organisatie niet wettelijk verplicht is, is er ook geen enkele wens om het decision making proces uit te leggen, want het voegt niets toe. Het maakt niet uit dat niemand er meer “een bal begrijpt” van wat er in de black box gebeurt, zolang de succes rate maar wordt gehaald.	[CS2] De designer zegt zelf niet te weten hoe je de black box uit kan leggen, zonder dit met behulp van tools en software te doen. Wat overeenkomt met de beheersmaatregel.