MASTER'S THESIS

Transparantie en artificiële intelligentie: “...en zij leefden nog lang en gelukkig”?

Meuwissen, M.

Award date:
2021

Transparantie en artificiële intelligentie: “...en zij leefden nog lang en gelukkig”?

Over het effect van het gebruik van een uitleg als mechanisme van transparantie op begrip en vertrouwen rondom AI.

Opleiding	Open Universiteit, faculteit Management, Science & Technology Masteropleiding Business Process Management & IT
Programma	Open University of the Netherlands, faculty of Management, Science & Technology Master Business Process Management & IT
Cursus	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student	Mariska Meuwissen
Identiteitsnummer	
Datum	01-04-2021
Afstudeerbegeleider	Dr. Laury Bollen
Meelezer	Prof. Dr. Ir. Remko Helms
Derde beoordelaar	n.v.t.
Versie nummer	1.0
Status	definitief

Abstract

Transparantie kan mogelijk het begrip van en het vertrouwen in algoritmische beslissingssystemen beïnvloeden. Het openen van de ‘black box’ is hiervoor misschien niet altijd nodig. Inzichten uit de sociale wetenschappen met betrekking tot het geven van een uitleg zullen worden gebruikt om een online experiment op te zetten, waarin onderzocht wordt of een uitleg kan zorgen voor meer begrip en vertrouwen, en wat voor soort uitleg het meest bijdraagt aan begrip voor en het vertrouwen in de beslissingen die voortkomen uit systemen die een algoritmische grondslag hebben.

Sleutelbegrippen

Artificiële intelligentie, governance, algoritmische beslissingssystemen, transparantie, vertrouwen, uitleg

Samenvatting

Artificiële intelligentie is bezig aan een opmars. Steeds meer organisaties zetten algoritmes in om beslissingen te maken over een velerlei aan onderwerpen, of om complexe problemen op te lossen. Aan de ene kant brengt deze nieuwe technologie kansen met zich mee op het gebied van onder andere economische ontwikkeling, maatschappelijk belang en wereldgezondheid; aan de andere kant brengt het echter ook nieuwe uitdagingen met zich mee op het gebied van ethiek, regelgeving, standaarden, toezicht, veiligheid en privacy. Daarnaast groeit ook de zorg over potentieel misbruik en het wantrouwen met betrekking tot de inzet van AI, wat belemmeringen kan opleveren in de acceptatie en daarmee het succes van deze nieuwe technologie. Om dit soort risico's in kaart te brengen en te mitigeren, is governance van AI onontbeerlijk. Hoewel er de laatste jaren veel aandacht is voor governance, staan de ontwikkeling en implementatie ervan nog in zijn kinderschoenen. Een belangrijk onderdeel van een volwassen governance is het creëren van transparantie en daarmee meer verantwoording en vertrouwen met betrekking tot AI. Transparantie is echter geen eenduidig begrip: zo zijn er verschillende niveaus en manieren van transparantie te onderscheiden, er zijn meerdere stakeholders bij betrokken, en er kleven ook nadelen aan bepaalde vormen van transparantie. Volledige transparantie (het openen van de 'black box') is niet zonder meer de oplossing. In deze studie wordt middels een experiment onderzocht hoe transparantie een mogelijke bijdrage kan leveren aan meer begrip van en vertrouwen in besluitvormende AI-systemen door niet-experts. De resultaten suggereren dat transparantie in de vorm van een uitleg een positief effect heeft op het begrip, maar niet op het vertrouwen in (de uitkomst van) een AI-systeem.

Summary

Artificial intelligence is on the rise. Ever more organizations are using algorithms to make decisions about a wide variety of subjects or to solve complex problems. On the one hand, this new technology offers opportunities in areas such as economic development, social importance and global health. On the other hand, however, it also brings new challenges in the areas of ethics, regulations, standards, supervision, security and privacy. In addition, there is also growing concern about potential abuse and mistrust with regard to the use of AI, which can pose a threat to the acceptance and thus the success of this new technology. To map out and mitigate these types of risks, AI governance is indispensable. Although there has been a lot of attention for governance in recent years, its development and implementation are still in its infancy. An important part of mature governance is the creation of transparency and thus more accountability and trust with regard to AI. However, transparency is not an unambiguous concept: different levels and ways of transparency can be distinguished, multiple stakeholders are involved, and there are also disadvantages to certain forms of transparency. Full transparency (opening the "black box") is not necessarily the solution. This study uses an experiment to investigate how transparency can contribute to greater understanding of and confidence in decision-making AI systems by non-experts. The results suggest that transparency in the form of an explanation has a positive effect on the understanding, but not on trust in (the outcome of) an AI system.

Inhoudsopgave

1. Introductie

Gaat kunstmatige intelligentie de mensheid overnemen? In de film 'Minority report', waarin een systeem genaamd 'Precog' voorspelt of iemand een misdaad gaat plegen, en mensen al veroordeeld worden voordat ze de misdaad hebben begaan, wel. Dit lijkt misschien een vergezochte dystopie, maar in de internationale strijd tegen terrorisme wordt kunstmatige intelligentie al volop ingezet om toekomstige terreurdaden te kunnen voorspellen, en mensen die hiermee in verband worden gebracht, op te pakken. Ook in Nederland werkt de Nationale Politie met het Criminaliteit Anticipatie Systeem; daarmee worden risicogebieden in kaart gebracht en incidenten voorspeld. De voorspellende kracht van deze nieuwe, baanbrekende technologie blijft toenemen, en met de komst van big data en 'deep learning' technieken lijken de mogelijkheden eindeloos. Maar hoe weten we of deze systemen wel betrouwbaar zijn? Wie controleert of een voorspelling juist is? En wie is er verantwoordelijk voor een beslissing die gemaakt wordt op basis van artificiële intelligentie? Met andere woorden: Hoe kunnen we vertrouwen op deze wereld veroverende techniek, die voor de meesten van ons een 'black box' is?

1.1 De noodzaak tot governance van AI

Hoewel de term Artificiële Intelligentie (AI) voor het eerst werd gebezigd in 1956 door John McCarthy toen hij de eerste academische conferentie hield over dit onderwerp[1], zien we pas de laatste jaren een opmars van de toepassingen van AI. Dit heeft mede te maken met het feit dat de populariteit van big data gestegen is in het afgelopen decennium, waardoor er meer data beschikbaar is voor analyses (Allam, 2019). Daarnaast zien we een toenemende complexiteit van de toepassingen; denk bijvoorbeeld aan 'deep learning' technieken, waarbij een artificieel neuraal netwerk zelflerende capaciteiten heeft. Desondanks is er tot op heden geen consensus over de exacte definitie van AI. Een algemene, bruikbare definitie is die van Allan Dafoe in zijn stuk "AI Governance: A research agenda" (2018), waarin hij AI definieert 'als de ontwikkeling van machines die in staat zijn om gesofisticeerd (intelligent) informatie te verwerken.' Veelal wordt er daarnaast een onderscheid gemaakt tussen 'zwakke' en 'sterke' AI. Het eerste wordt gekenmerkt door een duidelijke probleemstelling, waarbij AI op basis van regels en logica een correcte oplossing kan vinden. Een groot nadeel van zwakke AI is dat alle oplossingsmogelijkheden vooraf bekend moeten zijn om een probleem op te kunnen lossen. Sterke AI, waaronder ook 'deep learning' valt, biedt daarentegen meer mogelijkheden. Dit type AI veronderstelt een grote hoeveelheid data (big data), en het is over het algemeen moeilijker om uit te leggen hoe er tot een oplossing van een probleem is gekomen. Sterke AI is in staat om zeer complexe problemen op te lossen, zoals het voorspellen van menselijk gedrag.

De toepassingsmogelijkheden van sterke AI lijken bijna eindeloos en het ziet ernaar uit dat het gebruik ervan de komende jaren alleen maar zal toenemen. Door sommigen wordt de transformatie die AI teweeg zal brengen in deze eeuw zelfs beschouwd als een transitie die we kunnen vergelijken met de industriële revolutie[2][3]. Deze nieuwe technologie brengt naast een

[1] https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf

[2] http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity

[3] http://lukemuehlhauser.com/industrial-revolution/

schare aan beloftes op allerlei gebieden, van onder andere zelfrijdende auto's en nieuwe medische inzichten en toepassingen tot meer welvarendheid in het algemeen, echter ook nieuwe uitdagingen met zich mee op het gebied van ethiek, regelgeving, standaarden, toezicht, veiligheid en privacy (Butcher, 2019; Brundage, 2018). Butcher en Beridze (2019) beschrijven enkele van de bezwaren die deze vooruitstrevende techniek met zich meebrengt, zoals het verlies van banen door verregaande automatisering, mogelijk kwaadaardig gebruik en inzet van AI, verminderde verantwoording en aansprakelijkheid, kans op algoritmische bias en het gebrek aan transparantie, en de onbedoelde effecten die het gebruik van AI teweeg kan brengen. Hierbij kun je bijvoorbeeld denken aan een zelfrijdende auto die een dodelijk ongeluk veroorzaakt, of aan een verkeerde beslissing over een persoon die is genomen op basis van een automatisch beslissingssysteem. De impact van een niet goed functionerend algoritme kan enorm zijn.

Om dit soort risico's in kaart te brengen en te kunnen mitigeren is governance onontbeerlijk (Butcher, 2019), maar de governance van AI staat nog in de kinderschoenen (Wang, 2018). Governance van AI is een vakgebied dat bestudeert hoe de mensheid de transitie naar geavanceerde AI-systemen het beste in banen kan leiden (Dafoe, 2018). Hoewel er gepleit wordt voor wereldwijde standaarden om het gebruik van AI te coördineren en beheersen en de risico's die AI met zich meebrengt te mitigeren (Cihon, 2019), is er tot op heden geen duidelijkheid over hoe deze standaarden eruit moeten zien, of wie deze standaarden op moet stellen. Op regionaal niveau biedt de General Data Protection Regulation (GDPR) een beperkte oplossing, met name op het gebied van privacy, waarbij organisaties verplicht worden gesteld om in ieder geval het achterliggende datagebruik uit te kunnen leggen bij het maken van geautomatiseerde beslissingen. Echter biedt dit niet voldoende soelaas, omdat de GDPR op zichzelf niet voldoende is om een besluit geheel te rechtvaardigen of een (verkeerde) beslissing uit te kunnen leggen (Wachter, 2017). Gasser en Almeida (2017) spreken van grofweg drie uitdagingen op het gebied van de governance van AI. Ten eerste is er het 'black box' fenomeen, zo genoemd omdat de algoritmes die ten grondslag liggen aan AI zo complex zijn dat ze voor (de overgrote meerderheid van de) mensen onbegrijpelijk zijn. De uitlegbaarheid van de algoritmes is gering, waardoor er een grote kennisasymmetrie bestaat tussen AI-experts, AI-gebruikers en beleidsmakers. De tweede uitdaging ligt op het ethische vlak, waarbij criteria en principes een rol spelen. Als laatste noemen Gasser en Almeida het sociale en juridische aspect. Hierbij kan men denken aan regulatie en certificatie van AI-systemen. Op de korte termijn zal de eerste uitdaging de meest eminente zijn, omdat zij ten grondslag ligt aan de ethische en sociaaljuridische vraagstukken.

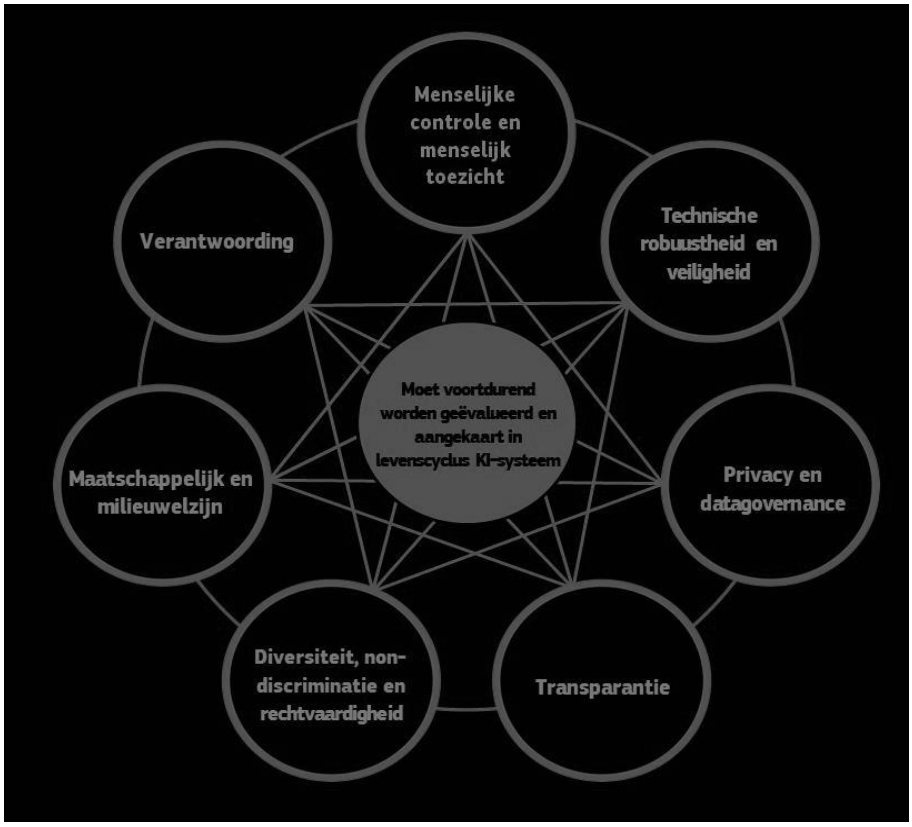
De verantwoording, het begrip en vertrouwen in AI, en de daarmee gerelateerde transparantie van algoritmes is van groot belang voor een eerste stap tot een volwassen governance van AI (Magrani, 2018; Vedder, 2017; Buiten, 2019). Maar hoe komen we tot een verantwoording die iedereen tevredenstelt? Is het mogelijk en haalbaar om de 'black box' die AI heet te openen voor een breed publiek? En brengt dat uiteindelijk wel de oplossing waar we naar zoeken? De mogelijkheid tot uitlegbaarheid (van de uitkomst) van algoritmes (transparantie) wordt door velen gezien als een van de voornaamste uitdagingen om AI breder in te zetten (Magrani, 2018; Diakopoulos, 2016).

1.2 Probleemstelling

In Europa, en specifiek ook in Nederland waar ons onderzoek zal plaatsvinden, wordt het enthousiasme over het gebruik van big data, algoritmes en artificiële intelligentie groter. Echter, het relatieve aandeel van Nederlandse publicaties over AI is de afgelopen jaren gezakt

van 4% tot minder dan 2% van de publicaties wereldwijd, vooral omdat andere landen veel meer zijn gaan doen. (SAPAI, 2019). Europa en daartoe behorend Nederland heeft een inhaalslag te maken om de grootmachten China en de VS bij te houden op het gebied van de technologische ontwikkelingen omtrent AI, zoals te lezen is in het Strategisch Actieplan voor Artificiële Intelligentie[4] dat door de rijksoverheid in 2019 is gepubliceerd. Maar het gebruik van AI brengt ook risico's met zich mee op het gebied van privacy, veiligheid en mensenrechten (Cihon, 2019). Als er geen maatregelen worden genomen om deze risico's in kaart te brengen, en te mitigeren, zal dit een grootschalig gebruik van deze technologie in de weg staan. De High Level Expert Group on AI (AI HLEG) van de Europese Commissie heeft recentelijk de 'ethics guidelines for trustworthy Artificial Intelligence' (AI HLEG, 2019) gepubliceerd, waarin zij het belang van rechtmatige, ethische en robuuste AI-systemen vanuit een technisch en sociaal perspectief onderstrepen. Meer specifiek noemen zij zeven vereisten voor betrouwbare AI: menselijke controle en menselijk toezicht, technische robuustheid en veiligheid, privacy en data governance, transparantie, diversiteit, non-discriminatie en rechtvaardigheid, maatschappelijk en milieuwelzijn en verantwoording. De AI HLEG onderkent dat transparantie een belangrijk onderdeel vormt van de governance van AI. In figuur 1 wordt schematisch in beeld gebracht hoe de governance van AI eruit zou moeten zien volgens deze groep van experts. Uit deze figuur valt tevens te lezen dat alle onderdelen met elkaar verbonden zijn, en dat wil ik hier nogmaals onderschrijven. Zonder transparantie is menselijk begrip en controle bijvoorbeeld niet mogelijk, maar ook de verantwoording van een AI-systeem zal uitermate moeilijk blijken als dit systeem verre van transparant is. Transparantie speelt daarmee ook een belangrijke rol als het gaat om het vertrouwen dat mensen hebben in een AI-systeem (Rossi, 2018; AI HLEG, 2019). De vraag is echter hoe transparantie eruit moet zien, om daadwerkelijk een gewenst effect te hebben. Kroll et al. (2016) benadrukken dat het openen van de 'black box' die AI heet een grote bedreiging vormt voor de marktposities van organisaties, omdat de algoritmes die ze gebruiken bedrijfsgeheimen zijn, die bij openbaring voor iedereen beschikbaar zijn. Daarnaast is het maar de vraag of men überhaupt zou begrijpen wat zo'n algoritme nu eigenlijk doet, omdat de broncode voor vrijwel de meeste mensen, en vooral voor een leek, onbegrijpelijk zou zijn. Er zal dus gekeken moeten worden naar een andere manier waarop (een vorm van) transparantie een bijdrage kan leveren aan meer begrip van en vertrouwen in AI.

[4] https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie



Figuur 1: Governance van AI, schematische weergave uit 'Ethische richtsnoeren voor betrouwbare KI', HLEG (2019)

1.3 Onderzoeksvraag

Met dit onderzoek wordt getracht een bijdrage te leveren aan de vraag hoe transparantie in een bepaalde vorm wellicht een bijdrage aan een oplossing tot een van de 'black box' problemen kan bieden. Dit doen we door voor een bepaalde doelgroep, de eindgebruiker, een uitleg als mechanisme van transparantie te beschouwen. Onderzocht wordt specifiek of en in welke vorm een uitleg kan bijdragen aan een tweetal gewenste effecten van transparantie, namelijk meer begrip en vertrouwen. In hoofdstuk 2 zal eerst dieper ingegaan worden op het begrip transparantie, en wat hier allemaal bij komt kijken, alvorens we de diepte ingaan, en transparantie vanuit een bepaalde invalshoek, zoals hierboven beschreven, nader zullen bekijken.

Transparantie is een breed begrip dat in verschillende contexten een andere invulling behoeft (Felzmann, 2019). In dit onderzoek zullen twee beoogde effecten van transparantie van geautomatiseerde beslissingssystemen centraal staan: begrijpelijkheid en vertrouwen. We zullen ons tevens beperken tot een bepaalde doelgroep waarvoor juist deze twee effecten belangrijk zijn en die, zoals we zullen zien, in de huidige onderzoeken onderbelicht is gebleven: de (eind)gebruikers of de personen die de beslissing aangaat. De onderzoeksvraag die hierbij gesteld wordt, luidt:

In welke vorm kan transparantie bijdragen aan een beter begrip van en vertrouwen in (de uitkomsten van) AI-systemen bij de personen die deze uitkomsten aangaan?

1.4 Relevantie van het onderzoek

Zoals we in voorgaande paragrafen hebben gezien speelt het begrip transparantie een cruciale rol binnen de governance van AI. Belangrijke besluiten en processen worden steeds meer gestuurd door algoritmische beslissingssystemen, waardoor er meer belang gehecht wordt aan de transparantie van dit soort systemen (Diakopoulos, 2016; Pasquale, 2015). Transparantie kan ervoor zorgen dat informatie die normaal gesproken onzichtbaar is voor een individu zichtbaar gemaakt wordt, waardoor het kan bijdragen aan meer begrip omtrent hoe en waarom een systeem functioneert. Echter, er is nog weinig onderzoek gedaan naar hoe transparantie nu eigenlijk eruit moet zien voor gebruikers om hun begrip van en vertrouwen in algoritmische beslissingssystemen te vergroten. Een empirisch onderzoek dat gedaan is op dit gebied heeft aangetoond dat transparantie in een bepaalde vorm en context kan bijdragen aan het vertrouwen in AI-systemen (Kizilcec, 2016). In dit onderzoek werd tevens gevonden dat meer transparantie niet altijd meer vertrouwen betekent, maar dat te veel transparantie ook een averechts effect kan hebben. In dit onderzoek is echter niet gekeken naar het mechanisme waarmee transparantie meer vertrouwen sorteert, en daardoor blijft het onduidelijk op welke manier transparantie nu precies bijdraagt aan vertrouwen. Grimmelikhuijsen en Meijer (2014) hebben ook onderzoek gedaan naar de relatie tussen transparantie en vertrouwen. Hoewel dit onderzoek zich niet specifiek richt op AI, maar op de relatie tussen transparantie en het waargenomen vertrouwen in de overheid, zijn de resultaten van dit onderzoek wel relevant. Ook hier werd gevonden dat transparantie een relatie heeft met vertrouwen, en meer specifiek vonden zij dat de predispositie tot vertrouwen, en de voorkennis met betrekking tot het onderwerp van onderzoek een modererend effect heeft op gepercipieerd vertrouwen. Hier dienen we rekening mee te houden als we de relatie tussen transparantie en vertrouwen willen onderzoeken. Er lijkt dus sprake van een relatie tussen transparantie en vertrouwen, maar de manier waarop transparantie verschaft wordt en de context waarin dit gedaan wordt, lijkt van invloed te zijn op de resultaten die gevonden worden. In het volgende hoofdstuk zullen we zien dat transparantie een breed begrip is, en dat het ‘black box’ probleem niet één oplossing kent, dat er meerdere stakeholders zijn, en dat er ook beperkingen aan transparantie zijn. Transparantie in de vorm van uitlegbaarheid van een algoritme stimuleert de acceptatie en het vertrouwen van gebruikers in een algoritmisch systeem (Biran, 2017). Het is echter onduidelijk welk soort uitleg nu eigenlijk het meest geschikt is. Vanuit de hoek van de sociale wetenschappen zijn er onderzoeken bekend over hoe mensen onderling communiceren en hoe zij elkaar dingen uitleggen. Deze kennis zou van waarde kunnen zijn voor het onderzoek naar transparantie van AI-systemen, maar tot nu toe is hier weinig onderzoek naar gedaan (Miller, 2019). Onderzoeken naar transparantie hebben zich tot op heden vooral gefocust op het openen van de ´black box’, maar hierbij stoot men onherroepelijk op de bezwaren die volledige transparantie met zich meebrengt. Eigenlijk kunnen we ook wel spreken van meerdere ‘black boxes’, omdat het probleem zo divers is dat er niet gezocht moet worden naar een enkele oplossing voor alle contexten en stakeholders.

1.5 Opbouw van het rapport

In de volgende hoofdstukken zal het onderzoek gestructureerd worden opgebouwd. Hoofdstuk 2 vormt het theoretisch kader; hierin zal worden besproken wat er tot nu toe al bekend is in de wetenschappelijke literatuur met betrekking tot de geformuleerde onderzoeksvraag. Daarnaast zullen er hypotheses opgesteld worden voor het vervolgonderzoek. In hoofdstuk 3 komt de onderzoeksmethode aan bod, en wordt er stilgestaan bij de opbouw van het onderzoek, en

welke afwegingen daarbij gemaakt zijn. In hoofdstuk 4 staat de bespreking van de resultaten centraal, en in hoofdstuk 5 tenslotte volgt er een conclusie en een discussie naar aanleiding van de gevonden uitkomsten. Tevens wordt er gereflecteerd op het onderzoek.

2. Theoretisch kader

In dit hoofdstuk wordt getracht een antwoord te formuleren op de onderzoeksvraag die in het eerste hoofdstuk aan de orde is gekomen. Dit wordt gedaan op basis van een literatuuronderzoek. Voor dit literatuuronderzoek is er een zoekstrategie opgesteld om relevante wetenschappelijke literatuur te vinden. Deze zoekstrategie is terug te vinden in bijlage 1. Er zal in dit hoofdstuk worden besproken wat er tot nu toe in de literatuur bekend is op dit gebied, welke onderzoeken er al gedaan zijn, en welke antwoorden daarin al gegeven zijn. Daarnaast is er ruimte om de vragen die hierna onbeantwoord zijn gebleven, te herformuleren, en hypotheses op te stellen voor het vervolgonderzoek.

2.1 Waarom transparantie?

Artificiële intelligente wordt steeds meer gebruikt en toegepast in ons dagelijks leven en de verwachting is dat deze groei de komende jaren alleen maar toe zal nemen. Een voor veel mensen sprekend voorbeeld hiervan is de zelfrijdende auto. Maar op dit moment wordt er bijvoorbeeld door de belastingdienst al veel gebruik gemaakt van zelflerende systemen die fraude helpen opsporen. Ook lokale overheden maken gebruik van algoritmes om bijvoorbeeld bijstandsfraude te kunnen voorkomen. Om toezicht te kunnen houden op de ontwikkelingen die plaatsvinden op dit gebied, is governance van kunstmatige intelligentie onontbeerlijk. Dat betekent tevens dat de onderliggende algoritmes uitlegbaar moeten zijn. Als we immers niet weten hoe een systeem tot bepaalde beslissing of resultaat komt, hoe kunnen we dan de rechtmatigheid of rechtvaardigheid van zo'n systeem bepalen? 'Computer says no' is geen geldige verklaring. We hebben menselijke intelligentie nodig om bijvoorbeeld algoritmische vooringenomenheid, oftewel bias op te sporen, om te achterhalen wanneer en waarom de beslissingen van een model juist of verkeerd zijn en om individuele beslissingen te kunnen uitleggen. Dat kunnen we doen door de modellen te inspecteren, door statistische analyse en door geautomatiseerde besluiten te controleren op hun juistheid. Dit vereist een transparante benadering van deze geautomatiseerde besluitvormingssystemen.

Transparantie is een onderwerp dat meer en meer publieke aandacht krijgt, ook in Nederland. Zeer recentelijk (februari 2020) deed de rechtbank in Den Haag een uitspraak inzake het systeem Systeem Risico Indicatie (SyRi)[5]. SyRi is een wettelijk instrument dat door de Nederlandse overheid gebruikt wordt ter voorkoming en bestrijding van fraude. Het betreft een technische infrastructuur en bijbehorende procedures (algoritmes) waarmee in een beveiligde omgeving anoniem data kunnen worden gekoppeld en geanalyseerd, teneinde risicomeldingen te genereren. In deze uitspraak valt te lezen dat de rechtbank van oordeel is dat *"de Syri-wetgeving onvoldoende inzichtelijk is en controleerbaar is voor de conclusie dat de inmenging die de inzet van SyRi in het recht op respect voor het privéleven mee kan brengen noodzakelijk, evenredig en proportioneel is in verhouding tot de doelen die de wetgeving dient."* Het transparantiebeginsel, zoals vastgelegd in de AVG (AVG, 6.27-6.34) is volgens de rechtbank onvoldoende in acht genomen. Maar wat is dat eigenlijk, transparantie? Transparant voor wie? In deze zaak, voor de rechter? Voor een groep deskundigen? Voor de ontwikkelaar van SyRi? Voor iedereen? In de AVG staat:

"Het transparantiebeginsel vergt toegankelijke en begrijpelijke informatie, communicatie en eenvoudig taalgebruik..."

[5] http://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBDHA:2020:865

Het lijkt erop dat iedereen het moet kunnen begrijpen. Maar hoe doe je dat dan? En wat precies moet er transparant gemaakt worden? Om deze vragen te kunnen beantwoorden zal ik eerst ingaan op de onderliggende structuren van een systeem zoals SyRi, een zelflerend, besluitvormend of ondersteunend systeem. Daarna zal het begrip 'transparantie' worden uitgelegd, en waarom transparantie zo'n belangrijke rol speelt in de hedendaagse discussies rondom AI-systemen. We zullen zien dat transparantie een complex begrip is, en dat de tegenpool van transparantie - ondoorzichtigheid - ten grondslag ligt aan het zogenaamde 'black box probleem'. Ook zullen we kort stilstaan bij de uitdagingen die transparantie met zich meebrengt. Transparantie heeft meerdere stakeholders, en het lijkt erop dat met name het wetenschappelijk onderzoek naar de vraag hoe transparantie eruit moet zien voor een bepaalde groep stakeholders - de groep van eindgebruikers achterblijft.

2.2 Een korte introductie van algoritmes en besluitvormende systemen

Het concept 'algoritme' is aan verschillende interpretaties onderhevig, maar het voert te ver om daar hier op in te gaan (voor een uitgebreid relaas, zie Kitchin, 2017). In enge zin bestaan algoritmes uit procedures die stap voor stap numerieke input verwerken tot output. Algoritmes kunnen worden gekarakteriseerd als ondoorzichtige en niet-neutrale menselijke constructen: Mensen zijn verantwoordelijk voor het programmeren en trainen van algoritmes. Door mensen gemaakte keuzes in de ontwerpfase spelen een niet te onderschatten rol in de uitkomst van een algoritme; het zijn keuzes met betrekking tot de manier waarop een algoritme een bepaald gewicht of prioritering geeft aan data, waardoor de uitkomst grof beïnvloed kan worden. Doordat algoritmes primair menselijke constructen zijn, kunnen maatschappelijke ongelijkheden, of de vooroordelen en waarden van programmeurs of opdrachtgevers, in een algoritme worden opgenomen. Hierdoor kan een algoritme een bepaalde bias laten zien, die kan zorgen voor problematische ongelijkheid, zoals vormen van discriminatie, in de uitkomsten. Met name omdat algoritmes steeds ingewikkelder en zelflerend worden, aan elkaar gekoppeld worden, en ze steeds vaker ingezet worden in big data-omgevingen, blijken zelfs experts niet altijd meer in staat om de werking ervan te begrijpen (Buiten, 2019). Algoritmes worden hierdoor vaak gezien als een 'black box': de input en output van het algoritme zijn bekend, maar hoe het tussenliggende proces functioneert is lastig te doorgronden.
Algoritmes werken meestal niet op zichzelf, maar functioneren binnen een bepaalde omgeving, een AI-systeem. Een AI-systeem bestaat uit algoritmes, de onderliggende data en de bedrijfsmodellen (Zarski, 2016). Maar ook een AI-systeem functioneert niet als een op zichzelf staand geheel, de context waarin het gebruikt wordt is onlosmakelijk verbonden met de techniek. In deze scriptie zullen we met name een soort groep van AI-systemen onder de loep nemen, namelijk de geautomatiseerde beslissingssystemen. Deze systemen kenmerken zich doordat ze (een deel van) het besluitvormingsproces van de mens overnemen. Omdat deze systemen gebruik maken van algoritmes, worden zij vaak gekenmerkt door dezelfde ondoorzichtige structuur. Het is niet altijd mogelijk om de herkomst van een besluit goed te duiden. Kan transparantie een oplossing bieden?

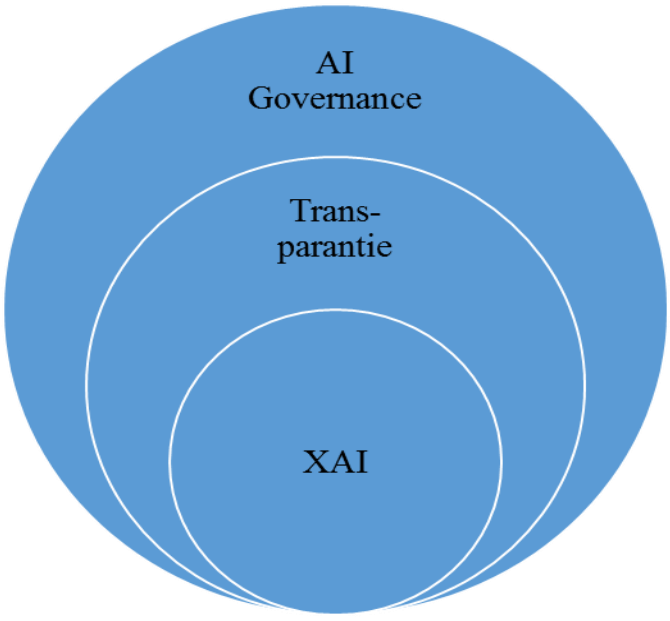
2.3 Wat is transparantie?

Het toenemende gebruik van (zelflerende) algoritmes om beslissingen te nemen en andere complexe problemen op te lossen, zorgt voor een immer toenemende vraag naar transparantie van deze systemen (Diakopoulos, 2016). Ondoorzichtigheid ligt ten grondslag aan het 'black box probleem', waarbij het AI-systeem de black box representeert, waar zich allerlei zaken afspelen die voor het menselijke oog verborgen blijven. Het is een probleem dat allerlei

uitdagingen met zich meebrengt op het gebied van privacy, juridische gronden, praktische en theoretische overwegingen. Op juridisch gebied speelt bijvoorbeeld de vraag wie er verantwoordelijk is wanneer een zelflerend systeem een verkeerde beslissing maakt; ondoorzichtigheid brengt ook vraagstukken met zich mee wat betreft eindgebruikers en hun rechten op het gebied van onder andere privacy onder de huidige GDPR.

Transparantie wordt in het algemeen gedefinieerd als het principe om het publiek informatie te verschaffen over operaties en structuren van een bepaalde entiteit (Heald, 2006). Transparantie wordt daardoor vaak als synoniem gezien met openheid, maar er is een subtiel verschil: openheid kan worden gezien als een karakteristiek van een organisatie, terwijl transparantie tevens impliceert dat er een ontvangende partij is die de geopenbaarde informatie verwerkt: Er zijn dus stakeholders bij betrokken. Het woord transparantie is afgeleid van het Latijnse woord ‘transparere’, dat doorschijnen betekent. Figuurlijk gezien betekent het ‘met een goede doorzichtigheid’. Dit lijkt te impliceren dat transparantie ons in de ‘black box’ kan laten kijken; maar wat treffen we daar dan aan?

Transparantie lijkt de afgelopen decennia een echt modewoord geworden te zijn. Overheden, winstgevende bedrijven en consumenten, iedereen lijkt op jacht te zijn naar meer transparantie. Maar wat is dat nu eigenlijk, waar we met zijn allen zo naarstig naar op zoek zijn? Is transparantie de heilige graal? Of slechts een utopie? Het begrip transparantie bestaat eigenlijk al heel lang, in diverse contexten. Ik zal mij hier beperken tot de betekenis van transparantie in het kader van technologische ontwikkelingen, en dan met name op het gebied van artificiële intelligentie en het aanverwante gebruik van algoritmes. Alvorens dieper in te gaan op transparantie wil ik eerst kort stil staan bij het vakgebied waar we het over gaan hebben. Het vakgebied waar het hier om draait wordt ook wel ‘explainable AI’ genoemd, of kortweg XAI. Zoals de naam wellicht al verraadt, gaat het hier om de uitlegbaarheid van de toepassingen van kunstmatige intelligentie. Hierbij kun je denken aan automatische besluitvormingssystemen die gebruik maken van algoritmes om tot hun oordeel te komen. Omdat de achterliggende algoritmes steeds meer complex worden, is het voor de mens vaak niet meer transparant hoe een dergelijke beslissing gemaakt is, en waarom. Een algemeen geaccepteerde definitie is er vooralsnog niet, maar de term XAI ‘lijkt veelal te refereren naar de initiatieven en inspanningen die gedaan worden als antwoord op de toenemende vraag naar transparantie en de vertrouwensissues die non-transparantie met zich mee brengt’ (Adadi, 2018). Het doel van XAI is om mensen in staat te stellen om door middel van een geschikte uitleg het begrip van en het vertrouwen in AI te waarborgen (Gunning, 2017). De roep om



Figuur 2: Schematische weergave van de begrippen AI governance, transparantie en XAI

transparantie neemt toe met het groeiende gebruik van artificiële intelligentie, big data en daarmee immer meer complexe algoritmes die automatische besluitvorming ondersteunen. Er bestaat steeds meer onduidelijkheid over de vraag op basis waarvan een bepaald besluit genomen is of hoe een voorspelling tot stand is gekomen. Transparantie is een complex en veelomvattend concept. In dit hoofdstuk zal ik eerst het begrip transparantie proberen te plaatsen in de huidige wetenschappelijke context en een antwoord geven op de vraag wat transparantie betekent in de context van informatiesystemen, zoals bijvoorbeeld een automatisch besluitvormingssysteem of een voorspellend systeem. In de volgende paragrafen zal achtereenvolgens uiteengezet worden wat het nu precies is dat transparant gemaakt dient te worden (wat), waarom het zo belangrijk is dat dit gebeurt (waarom), wie de stakeholders zijn (voor wie), en welke moeilijkheden en beperkingen transparantie met zich meebrengt. Hierna zullen we dieper ingaan op wat transparantie nu precies inhoudt voor de ontvangende partij en waarom het belangrijk is om transparantie als een multidisciplinair onderwerp te beschouwen, waarbij context een cruciale positie inneemt (hoe).

Om transparantie in zijn volledigheid te kunnen vatten, moeten we een onderscheid maken tussen prospectieve en retrospectieve transparantie, ook wel ex ante c.q. post hoc transparantie genoemd. Prospectieve transparantie informeert gebruikers vooraf over datagebruik en verwerking. Het beschrijft de werking van een (AI) systeem en hoe de uitkomst van een algoritme tot stand is gekomen. Prospectieve transparantie is hierdoor onder andere een vereiste voor de verantwoording van een systeem (Zerilli, 2018). We komen hier later nog op terug als we het belang van transparantie bespreken. Retrospectieve transparantie daarentegen refereert aan de uitleg van een beslissing achteraf. Het laat zien hoe en waarom een bepaalde beslissing tot stand is gekomen. Het vereist een mate van inspecteerbaarheid van een systeem. In deze hoedanigheid is retrospectieve transparantie onder andere van belang voor audits, maar ook voor uitlegbaarheid van beslissingen achteraf. Een ander onderscheid dat in de literatuur gemaakt wordt, is het verschil tussen objectieve versus subjectieve transparantie (Zhao, 2019). Objectieve transparantie, bezien vanuit het systeemperspectief, is de mate waarin een systeem informatie geeft over wat het doet en waarom. Subjectieve transparantie daarentegen is de mate van transparantie gezien vanuit het gebruikersperspectief: de mate waarin gebruikers de informatie dat een systeem prijsgeeft over het hoe en waarom waarnemen, en de mate waarin het beschikbaar is voor deze gebruikers.
De discussie rondom het belang van transparantie in het kader van technologische ontwikkelingen, waarvan artificiële intelligentie een groot en belangrijk onderdeel vormt, speelt zich voornamelijk af op het gebied van de zogenaamde verifieerbare aanpak van transparantie (Felzman, 2019). Binnen deze stroming wordt transparantie gezien als de openbaarmaking van informatie, zowel in kwalitatieve als kwantitatieve vorm. Door het openbaar maken van informatie zou de onderliggende entiteit zichtbaar worden, en kan er verificatie plaatsvinden met betrekking tot de waarheid, accuraatheid en rechtvaardigheid van de entiteit (Albu, 2019). Hier tegenover staat de performatieve aanpak van transparantie, waarbij een meer holistische aanpak van transparantie wordt nagestreefd, en waarbij transparantie in zijn volledige context beschouwd dient te worden. De definitie van transparantie binnen deze aanpak heeft meer verstrekkende gevolgen dan de verifieerbare vorm, waarbij er louter gekeken wordt naar de openbaarmaking van de informatie zelf: transparantie wordt gezien als meer dan alleen de informatie die overgebracht wordt, het is een geheel van complexe en dynamische communicatieprocessen. Transparantie wordt gezien als

relationeel concept, waarbij niet alleen een algoritme of de onderliggende data (de entiteit) onder de loep genomen wordt, maar waarbij de volledige verzameling van menselijke en niet-menselijke actoren evenals de context waarin zij opereren als een relationeel geheel bezien moet worden (Ananny, 2016; Felzman, 2019). Dit impliceert dat het louter beschikbaar maken van informatie niet voldoende is om transparantie te garanderen; het openmaken van de 'black box' is niet afdoende. Transparantie is meer, het heeft ook onbedoelde effecten en nadelen, en het is niet alleen een technisch concept, maar juist ook een sociaal fenomeen, dat tevens een culturele context kent (Kemper, 2018). Dit maakt performatieve transparantie een veelomvattend concept, waarbij er een centrale rol is weggelegd voor de gebruikers van een AI-systeem. In de volgende paragrafen wordt getracht om een overzicht te schetsen van de zaken die komen kijken bij performatieve transparantie.

2.4 Transparantie van wat?

Zoals we hebben gezien is transparantie een veelomvattend en complex begrip. Hierbij komt dat nog veel onduidelijkheid bestaat over hetgeen er transparant moet zijn. Is dat de ruwe data die door het algoritmisch systeem gebruikt wordt? De werking van het algoritme zelf? Of hebben we het hier ook over andere zaken? Koene (2019) geeft mijns inziens een volledig beeld van alle zaken waarover transparantie vereist kan zijn (voor diverse stakeholders). In de volgende tabel is te zien dat het niet slechts over data en algoritmes gaat, maar ook over de doelen van een algoritmisch systeem, de uitkomsten ervan, de naleving, invloed en het gebruik ervan.

Data	Transparantie van de data die verzameld wordt, en gebruikt wordt door het algoritmisch systeem, kan verwijzen naar de ruwe data, maar ook naar reeds geanalyseerde of geschoonde data, en de methodes om dit te realiseren.
Algoritmes	Transparantie van de algoritmes zelf, de wiskundige constructen, die samen met de data het AI-systeem maken.
Doelen	Transparantie met betrekking tot de doelen van een AI-systeem. Dit is typisch een prospectieve vorm van transparantie.
Uitkomsten	Voor vervaardigers van AI-systemen kan het nodig zijn om transparant te zijn met betrekking tot de uitkomsten van een systeem, of de tussenliggende interne staten van een berekening, om te kunnen aantonen dat de uitkomsten valide zijn. Dit gebeurt met name post hoc.
Naleving	Transparantie over de naleving van bepaalde regels of vereisten kan nodig zijn om aan te tonen dat een systeem voldoet aan de voorwaarden.
Invloed	Voor het publiek kan het interessant zijn om te weten of er invloeden van buitenaf zijn gebruikt om tot een bepaalde uitkomst te komen. Zo kan er bijvoorbeeld betaald zijn door een opdrachtgever om bepaalde uitkomsten uit te lichten.
Gebruik	Voor gebruikers van een systeem is transparantie met betrekking tot het gebruik van hun persoonlijke data onontbeerlijk. Zij willen de mogelijkheid hebben om deze data aan te passen naar hun eigen wensen, of wellicht zijn ze van mening dat het gebruik hun privacy schendt.

Tabel 1: Transparantie van wat? Naar het onderscheid dat Koene (2019) maakt.

Er is in de laatste jaren relatief veel onderzoek gedaan naar transparantie, en dan met name naar welke eisen er gesteld worden aan systemen om als transparant genoemd te worden, maar deze onderzoeken hebben zich voornamelijk beperkt tot hoe de informatie beschikbaar en toegankelijk wordt gemaakt op basis van ethische en juridische gronden (verifieerbaarheid). Bovendien lijken vooral de data en de algoritmes (en soms ook het modelgebruik) het onderwerp van discussie; de andere onderwerpen zijn onderbelicht. In mijn ogen zou het concept transparantie in zijn volledige breedte erkend moeten worden. Het is onmogelijk om transparantie als geheel te onderzoeken, omdat het te veelomvattend is, dus mijn voorstel zou zijn om bepaalde categorieën van transparantie (bijvoorbeeld op basis van stakeholders) te onderscheiden en deze apart te onderzoeken. Diverse stakeholders vergen een andere categorie van transparantie. Hoewel er onderzoek gedaan is naar volledige (performatieve) transparantie, en er ook al modellen zijn om dit te verwezenlijken (Morley, 2019), is er nog maar weinig bekend en onderzocht over de effecten en vereisten aan de kant van de ontvanger van deze informatie. Deze groep van stakeholders vraagt om een ander soort van transparantie, maar als we transparantie blijven bekijken vanuit een eenzijdig, technisch perspectief, zal het voor deze groep geen oplossing bieden. Een meer holistische beschouwing van het begrip transparantie is nodig om ervoor te zorgen dat *kwalitatieve informatie* beschikbaar is in een *betekenisvolle* en *bruikbare* vorm bij het *juiste publiek* (Hosseini, 2018).

2.5 Het belang / de rechtvaardiging van transparantie

In deze paragraaf wil ik ingaan op het belang van transparantie. Waarom is transparantie belangrijk, en hoe kan transparantie bijdragen aan een oplossing voor het 'black box' probleem? Zonder volledig te willen zijn in alle argumenten die meespelen bij dit belang, zal ik proberen de belangrijkste zaken op een rij te zetten. Vertrouwen, verantwoording en transparantie lijken onlosmakelijk met elkaar verbonden:

"Verantwoording is cruciaal voor het scheppen en behouden van vertrouwen van gebruikers in KI-systemen. Dat betekent dat processen transparant moeten zijn, dat de capaciteiten en het doel van KI-systemen openlijk kenbaar moeten worden gemaakt en dat beslissingen - voor zover mogelijk - verklaarbaar moeten zijn aan degenen die er direct of indirect de gevolgen van ondervinden." (AI HLEG, 2019)

De uitspraak in de zaak SyRi onderstreept nogmaals het belang van transparantie voor de verantwoording van en daarmee het vertrouwen in geautomatiseerde systemen. Indien we niet in staat zijn om de uitkomsten van een AI-systeem te verklaren, dan is het onmogelijk om de uitkomsten of beslissingen te vertrouwen of te verantwoorden. De immense toename van het gebruik van algoritmes legt extra druk op de verantwoording van deze geautomatiseerde besluitvorming (Koene, 2019). Dit maakt transparantie een onmisbaar element in de governance van AI-systemen, zoals we in hoofdstuk 1 al zagen.
Transparantie speelt daarnaast ook een rol bij het vertrouwen van consumenten om bepaalde systemen/applicaties te gebruiken. Organisaties en gebruikers willen AI-systemen die transparant, uitlegbaar, ethisch verantwoord en zonder bias zijn (Rossi, 2018). Zonder transparantie zullen consumenten geen vertrouwen hebben in de technologie en zal de adoptie ervan niet of niet volledig tot stand komen en zal er geen gebruik kunnen worden gemaakt van alle mogelijkheden die de technologie met zich meebrengt, aldus Rossi (2018). Rechtvaardigheid is een belangrijk onderdeel van geautomatiseerde beslissingssystemen (AI

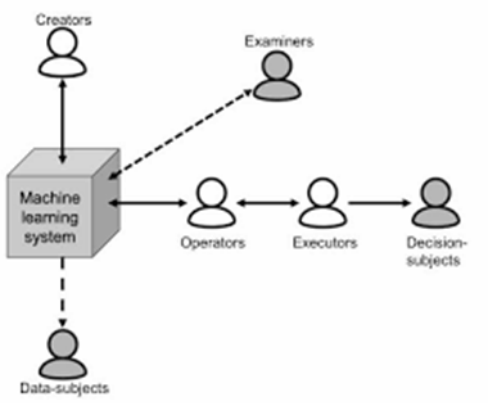
HLEG, 2019). Zarski (2016) beargumenteert dat transparantie ervoor kan zorgen dat geautomatiseerde systemen tot meer rechtvaardige besluiten komen dan besluiten die door de mens worden gemaakt, mits er aan bepaalde voorwaarden wordt voldaan. In het verlengde hiervan kan transparantie een rol spelen bij het verminderen of weghalen van (historische) bias, waardoor discriminatie of vooroordelen tegengegaan worden. Bias wordt vaak gezien als een van de grootste nadelen van AI-systemen, omdat het kan zorgen voor een oneerlijke beslissing over bepaalde individuen of groepen. Door meer inzichtelijk te maken hoe bepaalde besluiten tot stand zijn gekomen, is het mogelijk om bias te verminderen of uit te sluiten (Zarski, 2016; Rossi, 2019). Dit kan op zijn beurt ook weer bijdragen aan meer vertrouwen.
Transparantie kan ook als middel worden ingezet voor het zogenaamde 'crowdsourcing' (Zarski, 2013). Indien de werking van bepaalde algoritmes bij een breder publiek bekend worden gemaakt, zou dit kunnen leiden tot een verbetering van een algoritme, omdat er zo feedbackmogelijkheden worden gecreëerd die niet mogelijk zijn als er sprake is van ondoorzichtigheid van de ingezette methodes. Zodra er met meer ogen gekeken wordt, ziet men meer, en experts buiten een organisatie kunnen zo hun input geven en er kunnen mogelijke verbeteringen aangebracht worden. Op het gebied van de ontwikkeling van software kennen we al het 'open source' fenomeen: de broncode van software wordt publiekelijk geopenbaard, teneinde de doorontwikkeling te stimuleren en faciliteren. In de context van transparantie van AI-systemen ligt dit echter wat ingewikkelder, omdat de onderliggende algoritmes immer gecompliceerder worden, en voor de leek, en zelfs voor een expert, moeilijk te vatten zijn. Het vergt dus wellicht een wat andere vorm van transparantie. Echter, de grootste uitdaging die crowdsourcing met zich meebrengt, is het niveau van transparantie dat nodig is. Om werkelijk succes te hebben, zal volledige transparantie op alle onderdelen van een AI-systeem nodig zijn, en juist dat zorgt ervoor dat er ook risico's aan verbonden zijn zoals we in paragraaf 2.7 zullen zien. Een mogelijke oplossing hiervoor zou kunnen zijn om volledige transparantie te beperken tot bepaalde groepen van experts (Zarski, 2016; Weller, 2019).
In de privacywetgeving is er ook een rol voor transparantie weggelegd. Met de komst van de EU General Data Protection Regulation[6] (GDPR) worden organisaties wettelijk verplicht om een data subject erop te wijzen dat zijn gegevens gebruikt worden (het datacollectie proces). Echter, waarvoor deze gegevens gebruikt gaan worden, welke analyses erop plaats zullen vinden, daarover wordt de consument niet geïnformeerd. Weet een consument eigenlijk wel waarvoor hij toestemming geeft, als het niet transparant is hoe en waartoe zijn data gebruikt wordt? Je kunt je dus afvragen of er zo'n geval wel sprake is van geïnformeerd consent, zoals gespecificeerd in artikel 7 van de GDPR. Verwant aan de discussie rondom privacy is die met betrekking tot de autonomie van de betrokken individuen. De autonomie van de betrokkenen is een belangrijk speerpunt om transparantie te promoten. Als individuen worden geraakt door een automatische beslissing of voorspelling, lijkt het logisch dat zij recht hebben om te weten waarom; zij verdienen enige vorm van transparantie met betrekking tot de beslissing die hen aangaat, zoals welke beslissingscriteria er zijn gehanteerd en de logica achter een besluit. Zo'n besluit of voorspelling op basis van bijvoorbeeld profiling kan verstrekkende gevolgen hebben voor iemands leven. In dit geval voldoet alleen volledige transparantie; op het niveau van dataverzameling, analyse en de toegepaste modellen. De grote vraag blijft echter hoe dit eruit moet zien om betekenisvol en relevant te zijn voor de betrokkene. In de GDPR is vastgelegd dat individuen recht hebben op het verkrijgen van betekenisvolle informatie van de betrokken logica indien geautomatiseerde besluitvorming plaatsvindt met legale of anderszins verwante

[6] http://ec.europa.eu/justice/data-protection

effecten op het individu. Echter, uit de GDPR is weinig af te leiden wat dit betekent voor de transparantie van algoritmes (Wachter, 2017). Er is nog geen wetgeving die specifiek dit onderwerp behandelt. Uit bovenstaande mag duidelijk zijn dat dit wel gewenst is.

2.6 Stakeholders en transparantie

Inzicht in de werking en de uitkomst van een AI-systeem of algoritme is zoals we hebben gezien om verschillende redenen van belang. Echter, dit belang loopt uiteen voor diverse belanghebbenden (Weller, 2019). In deze paragraaf wordt uiteengezet wie de belanghebbenden zijn en we zullen zien dat de diverse groepen stakeholders wellicht om een ander type of niveau van transparantie vragen (Weller, 2019; Zednik, 2019). Het is daarom belangrijk om deze verschillende groepen van stakeholders te onderkennen. Tomsett et al. (2018) introduceren een Machine Learning ecosysteem dat erg bruikbaar is voor het onderscheiden van de stakeholders binnen een AI-systeem (figuur 3). De creatoren in dit systeem zijn de ontwikkelaars van het AI-systeem. Dit zijn de mensen die verantwoordelijk zijn voor het ontwerpen van de algoritmes en het creëren van het AI-systeem als geheel. De examinatoren zijn degenen die het systeem onderwerpen aan veiligheids- en compliance-testen, en bijvoorbeeld het uitvoeren van audits. Operatoren en uitvoerenden zijn beide gebruikers van het systeem, met dit verschil dat de eerste het systeem voeden met input en een output verkrijgen, en de laatste de eindverantwoordelijke is voor het maken van een data-gedreven besluit. De data subjecten zijn de personen wiens persoonlijke data wordt verzameld en gebruikt wordt in het AI-systeem om tot toekomstige voorspellingen en besluiten te komen. De laatste groep is de groep van personen over wie een beslissing genomen wordt. Aangezien deze verschillende groepen op een andere manier interacteren met een AI-systeem, hebben zij andere wensen ten aanzien van transparantie (Zhao, 2019). Weller (2019) combineert het onderscheid tussen de diverse stakeholders met het onderscheid tussen verschillende vormen van transparantie, die elk een ander soort uitleg vereisen. Hij onderkent daarbij 8 types transparantie, zoals in tabel 2 beschreven.



Figuur 3: Het ML ecosysteem; overgenomen uit Tomsett et al. (2018)

Type 1	Voor een ontwikkelaar, om te begrijpen hoe het systeem werkt, zodanig dat hij het kan debuggen of verbeteren.
Type 2	Voor een gebruiker, om hem te voorzien van een idee wat het systeem doet en hoe, zodanig dat de gebruiker voorspellingen kan doen over toekomstige uitkomsten, en om vertrouwen in het systeem te hebben.
Type 3	Voor de gemeenschap in het algemeen, om vertrouwd te raken met het systeem en diens voordelen en beperkingen, om de angst voor het onbekende weg te nemen.
Type 4	Voor een gebruiker, om te begrijpen waarom een bepaalde uitkomst of beslissing wordt gemaakt, zodanig dat er een check kan plaatsvinden of het systeem naar behoren functioneert, en om een beslissing aan te kunnen vechten.
Type 5	Om een expert de mogelijkheid te geven om een voorspelling of beslissing in detail te

	auditen, vooral wanneer er is mis is gegaan. Hiervoor kan het nodig zijn datastromen te monitoren en elke stap te traceren, en dit kan tevens de verantwoording en de juridische aansprakelijkheid ten goede komen.
Type 6	Om het monitoren en testen voor veiligheidsstandaarden te faciliteren
Type 7	Om ervoor te zorgen dat het publiek zich comfortabel en vertrouwd voelt met een voorspelling of beslissing, zodanig dat zij het systeem zal blijven gebruiken.
Type 8	Om een gebruiker (het publiek) te sturen naar bepaald gedrag of een actie.

Tabel 2: Types transparantie en hun stakeholders, zoals beschreven door Weller (2019).

In de huidige wetenschappelijke literatuur inzake xAI blijft het onderzoek naar de wensen van de leek, oftewel de data subjecten en degenen waarover besluiten worden genomen achter. In deze scriptie wil ik een bijdrage leveren aan het onderzoek naar transparantie en de daarmee onlosmakelijk verbonden uitlegbaarheid voor deze groep van stakeholders.

2.7 Beperkingen van transparantie

Is volledige transparantie een haalbare kaart? Transparantie wordt vaak gezien als wenselijk, omdat het inzicht geeft en bijdraagt aan de governance van kunstmatige intelligentie. Er wordt uitgegaan van de aanname dat de waarheid besloten ligt de volledige ontsluiting van een algoritme en de onderliggende data. Maar is dat eigenlijk wel zo? Of hebben we hier te maken met “de transparantie illusie” (Ananny, 2016)? Vragen we eigenlijk niet te weinig als we zeggen dat we in de ‘black box’ willen kijken om te zien wat daar precies gebeurt? Ananny en Crawford (2016) suggereren dat transparantie onmogelijk onderdeel kan zijn van een algoritmisch model. De ondoorzichtigheid van algoritmes zou daarentegen moeten worden gezien in de context van hun gebruik. Transparantie is volgens hen een socio-technisch construct tussen algoritmes en mensen. In de literatuur zijn diverse voorbeelden te vinden die indiceren dat volledige transparantie voor alle betrokken stakeholders een utopie is (Etzioni, 2016; Ananny, 2016; Felzman, 2019). Het is niet mijn bedoeling om alle beperkingen in kaart te brengen, maar om de belangrijkste zaken eruit te lichten en te laten zien waarom deze het ‘black box’ probleem tot een ingewikkeld probleem maken. Daarnaast wil ik laten zien dat deze problemen voor verschillende stakeholders wellicht een andere oplossing vergen. De bezwaren die er gemaakt worden zijn grofweg te verdelen in vier categorieën, te weten privacy, de mogelijke averechtse effecten van volledige transparantie, het verlies van concurrentievoordeel en de inherente ondoorzichtigheid van algoritmes. In deze paragraaf zal ik deze bezwaren verder toelichten.

Als we denken aan volledige transparantie van de datasets die ten grondslag liggen aan het gebruik van algoritmes, wordt het meteen duidelijk dat privacy en transparantie soms elkaars tegenpolen zijn. Transparantie vereist immers de inzichtelijkheid in de onderliggende datasets en daar kan privacygevoelige informatie in opgeslagen zijn. Dit kan zelfs voor een en dezelfde persoon een contradictie teweegbrengen; denk bijvoorbeeld aan een gebruiker van een systeem die graag wil dat zijn data privé blijft, maar daarentegen ook graag een uitleg wil van hoe datzelfde systeem (de algoritmes en de onderliggende data) werkt. Een ander bezwaar voor volledige transparantie is het zogenaamde ‘gaming the system’, waarbij transparantie als nadelig effect heeft dat mensen misbruik kunnen maken van het kennen van de onderliggende structuren. Als men weet hoe een bepaald systeem (denk hierbij bijvoorbeeld aan een fraude-opsporingssysteem zoals Syri) tot een beslissing over verdacht gedrag komt, is het mogelijk

om dit te omzeilen. Tevens kunnen bedrijven hun concurrentievoordeel verliezen als de exacte werking van hun algoritmes blootgesteld worden. We kunnen dus ook wel spreken van een transparantie-paradox[7]: Hoewel meer informatie of transparantie over AI zeker zijn voordelen heeft, brengt het ook risico's met zich mee. Ten laatste wil ik de inherente ondoorzichtigheid van algoritmes noemen. Algoritmes en AI-systemen worden steeds intelligenter en daarmee neemt hun ingewikkeldheid toe. Soms lijkt het zelfs voor experts onmogelijk om een beslissing of voorspelling van een AI-systeem achteraf precies te kunnen verklaren.
Uit bovenstaande blijkt wel dat het 'black box' probleem geen eenzijdig probleem met een eenduidige oplossing. Het lijkt erop dat het transparant maken van een AI-systeem in technische zin dit probleem niet zomaar op kan lossen. Transparantie kent vele facetten en er zijn diverse actoren bij betrokken, daarom is het van belang deze facetten en actoren te onderkennen, multidisciplinair te onderzoeken en voor elk van hen te bekijken welke oplossingsmogelijkheden er zijn. Volledige transparantie lijkt vooralsnog in veel gevallen wel technisch haalbaar (Zednik, 2019) en wellicht ook nodig te zijn om te zorgen voor onder andere verantwoording, rechtvaardig gebruik en juridische aspecten, maar dit lijkt slechts voorbehouden voor een bepaalde doelgroep: een expertgroep en/of toezichthoudende instantie die alle ins en outs kent (de Laat, 2018). Transparantie in deze vorm voor het grote publiek beschikbaar stellen lijkt niet aan te raden omdat het onbegrijpelijk is voor de leek en er tevens risico's aan verbonden zijn. Maar stel voor het moment dat we met zijn allen besluiten dat we deze verantwoordelijkheid in handen leggen van een selecte groep, die zorg draagt voor een ethische en verantwoorde aanpak, zoals we ook doen met veel andere wetenschappelijke vragen, dan blijft de vraag hoe het zit met de uitlegbaarheid van een automatisch genomen besluit aan de leek, of degene die het besluit betreft. In de volgende paragrafen wordt hier dieper op ingegaan.

2.8 Transparantie en vertrouwen

Zoals we eerder hebben gezien is een van de beoogde effecten van transparantie het verhogen van het vertrouwen in AI-systemen. Vertrouwen is essentieel om de sociale acceptatie van het gebruik van artificiële intelligentie te waarborgen; zonder vertrouwen zal de adoptie van systemen die gebruik maken van artificiële intelligentie stagneren (Wachter, 2017; Rossi, 2018). Vertrouwen is een lastig concept. Als we het over vertrouwen hebben, hebben we het altijd over vertrouwen in een bepaalde context (Siau, 2018). Een arts vertrouwen we als hij ons een middel voorschrijft om ons van onze klachten af te helpen, maar we zouden deze arts niet vertrouwen om bijvoorbeeld onze belastingaangifte in te vullen. Hierdoor kan vertrouwen in verschillende situaties anders uitpakken. Vertrouwen in computersystemen in het algemeen en intelligente beslissingssystemen in het bijzonder is een onderwerp dat met de recente groei en verwachte toename van het gebruik van dit soort systemen van eminent belang is: Als een gebruiker een dergelijk systeem niet vertrouwt, kan dit ernstige gevolgen hebben voor het gebruik ervan, of voor de gebruiker zelf die geconfronteerd wordt met een beslissing die hij niet vertrouwt. Vertrouwen is algemeen gezien een complex begrip, maar hier zal de definitie van McAllister (1995) gebruikt worden, die toegespitst is op het onderwerp van dit onderzoek:

> *De mate waarin een gebruiker zeker is van, en gewillig om te acteren op basis van, de aanbevelingen, acties, en beslissingen van een artificieel intelligent hulpmiddel voor het nemen van beslissingen.*

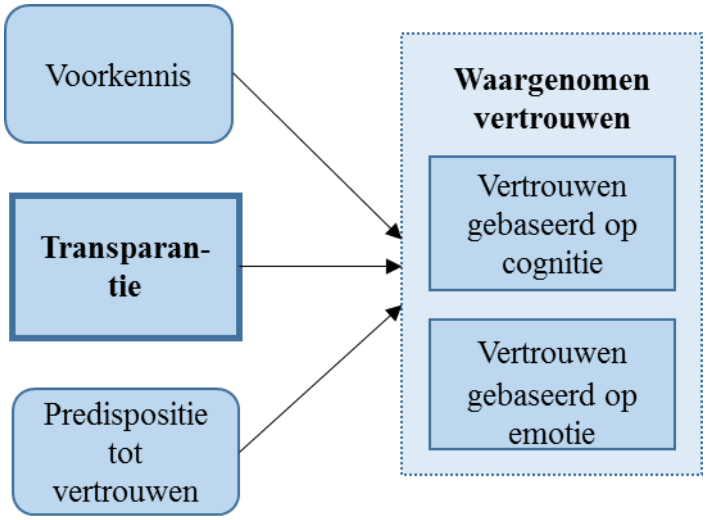
[7] https://hbr.org/2019/12/the-ai-transparency-paradox

Siau en Wang (2018) benadrukken de rol die context speelt bij vertrouwen. Zij benoemen een aantal factoren die van invloed zijn op het vertrouwen dat mensen hebben in technologie in het algemeen, en stellen dat het vertrouwen in AI-systemen in het bijzonder globaal afhankelijk is van drie factoren: de menselijke karakteristieken, de omgeving en de technologische karakteristieken. Onder de menselijke karakteristieken valt de predispositie tot vertrouwen, wat neerkomt op de algemene bereidheid van iemand tot het vertrouwen van een bepaalde technologie, in dit geval een AI-systeem. Daarnaast speelt ook de kennis met betrekking tot het onderwerp van vertrouwen een rol. Transparantie is volgens Siau en Wang een van de technologische kenmerken die een rol spelen bij het vertrouwen in een AI-systeem. Naast transparantie noemen zij bijvoorbeeld nog de prestaties van het systeem en het doel dat het systeem dient.

De relatie tussen transparantie en vertrouwen is al eerder onderzocht. Grimmelikhuijsen en Meijer (2014) vinden in hun onderzoek naar de relatie tussen transparantie en vertrouwen een positief effect van transparantie op het vertrouwen dat mensen hebben in een overheidsinstelling. Hierbij dient wel opgemerkt te worden dat de context in deze studie een andere is dan die waarin we hier geïnteresseerd zijn. Sinha et al. (2002) vinden een verband tussen transparantie en het vertrouwen dat mensen hebben in aanbevelingssystemen die gebruik maken van algoritmes. Transparantie wordt door hen gemeten als de mate van begrip dat iemand heeft voor de aanbeveling die hij of zij krijgt (middels de vraag: ‘Begrijp je waarom het systeem deze aanbeveling voor jou heeft gedaan?’). Uit deze studie blijkt dat gebruikers meer vertrouwen hebben in een aanbeveling naarmate ze deze aanbeveling beter begrijpen en volgens Sinha et al. dus meer transparant achten. Er lijkt dus een verband te zijn tussen transparantie en vertrouwen, maar de manier waarop transparantie wordt benaderd in deze studies loopt uiteen. Er blijft onduidelijkheid over op welke manier transparantie nu precies bijdraagt aan vertrouwen.

Grimmelikhuijsen (2012) vindt in een eerder experiment dat het effect van transparantie op vertrouwen gering is, en dat voorkennis en de predispositie tot vertrouwen (in dit geval in een overheidsinstelling, dus niet één op één te vertalen naar ons experiment) een meer belangrijke factor blijkt voor het vertrouwen dat mensen hebben. Voorkennis en predispositie tot vertrouwen moet volgens Grimmelikhuijsen meegenomen worden in de theoretische modellen van de relatie tussen transparantie en vertrouwen. In latere experimenten wordt een groter effect gevonden van transparantie op gepercipieerd vertrouwen, waarbij de conclusie wordt getrokken dat dit effect context-afhankelijk is (Grimmelikhuijsen, 2015; de Fine Licht, 2014).

Voorkennis en predispositie tot vertrouwen zijn desalniettemin factoren waarmee we rekening dienen te houden, en dit zal worden opgenomen in het theoretische model, zoals in figuur 4 afgebeeld.
Zoals in figuur 4 ook te zien is, zullen we het waargenomen vertrouwen onderverdelen in vertrouwen gebaseerd op cognitie en vertrouwen gebaseerd op emotie. Dit is een onderverdeling die veelvuldig gebruikt wordt in modellen met betrekking tot vertrouwen (McAllister, 1995; Madsen, 2000). Madsen en Gregor (2000) hebben onderzoek gedaan naar de factoren die vertrouwen bepalen in intelligente beslissingssystemen. Omdat vertrouwen een contextafhankelijk begrip is, en het onderzoek van Madsen zich specifiek richtte op intelligente systemen die ontworpen zijn om ondersteuning te bieden bij het maken van een beslissing (i.e. dezelfde context als die hier gehanteerd wordt), is deze studie erg bruikbaar. Ook in deze studie wordt een onderscheid gemaakt tussen *op emotie gebaseerd vertrouwen* en *op cognitie gebaseerd vertrouwen*. Dit onderscheid wordt meegenomen in het theoretisch model.



Figuur 4: Theoretisch model van de relatie tussen transparantie en vertrouwen

2.9 Een uitleg als mechanisme voor transparantie

Concluderend kunnen we vaststellen dat de definitie van transparantie geen eenduidige is; we hebben geconstateerd dat het openen van de black box die AI heet, niet afdoende is om transparantie te garanderen. Bovendien is het maar de vraag of het openen van de black box überhaupt wel de oplossing biedt waar we naar op zoek zijn. Transparantie is niet louter een technische aangelegenheid, maar eerder een socio-communico-technisch onderwerp dat afhankelijk van een bepaalde context vormgegeven dient te worden. Veel bestaand onderzoek was voornamelijk gefocust op het technische aspect van transparantie, maar dit heeft tot op heden geen resultaten opgeleverd die bruikbaar zijn voor alle stakeholders en in elke context. Omdat transparantie geen eenduidig begrip is, en vele facetten en stakeholders kent die diverse behoeften hebben, is het nodig om dit onderzoek vanuit verschillende invalshoeken te doen. In de huidige wetenschappelijke literatuur wordt er vooral onderzoek gedaan naar de mogelijkheden om systemen meer transparant te maken door middel van het geven van een wetenschappelijke uitleg van een algoritme (Mittelstadt, 2019; Guidotti, 2018)). Hoewel deze vorm van een uitleg zeker zijn toegevoegde waarde heeft, met name voor bepaalde stakeholders, zoals experts, is een dergelijke uitleg voor een leek, of degene die de beslissing aangaat, vaak onvoldoende, of ronduit onbegrijpelijk (Miller, 2019). Dat dit een probleem is, wordt ook duidelijk in de rechterlijke uitspraak rondom SyRi:

"...de menselijke gebruiker van zo'n zelflerend systeem begrijpt niet waarom het systeem concludeert dat er een verband is. Een bestuursorgaan dat zijn handelen (mede) baseert op zo'n systeem kan zijn optreden niet goed verantwoorden en zijn besluiten niet goed motiveren."

Een ander soort uitleg lijkt nodig om ook voor de leek een beslissing te kunnen uitleggen. In dit onderzoek willen we bekijken hoe en of verschillende soorten uitleg een verband hebben met de begrijpelijkheid van en het gepercipieerd vertrouwen in een beslissingssysteem dat zijn besluiten baseert op algoritmische constructen.
Consumentenvertrouwen speelt, zoals uiteengezet in de vorige paragraaf, een grote rol bij het gebruik van AI-systemen: als mensen geen vertrouwen hebben in de uitkomsten van een systeem, een voorspelling of een advies, dan zijn ze minder geneigd dit systeem te gebruiken (Rossi, 2019). Op welke manier kan transparantie wellicht een bijdrage leveren aan meer vertrouwen bij de gebruikers van algoritmische beslissingssystemen? Het geven van een uitleg is bij uitstek een manier om transparantie van een AI-systeem te verkrijgen (Tintarev, 2007). Als we het geven van een uitleg zien als een van de mechanismen waarmee transparantie verkregen kan worden, zullen we ons moeten verdiepen in de vraag wat een goede uitleg is. Uitlegbaarheid is een breed begrip; een expert heeft een ander soort uitleg nodig dan een leek om een uitleg te begrijpen, en een rechter die een oordeel moet vellen over of een beslissing die gemaakt is door een algoritmisch beslissingssysteem juridisch steekhoudt, vereist wellicht weer een andere uitleg. Voor dit onderzoek zullen we ons beperken tot het geven van een uitleg voor een gebruiker van een algoritmisch beslissingssysteem. Het geven van een uitleg aan een gebruiker kan meerdere doelen dienen: Een uitleg kan dienen ter verificatie of verbetering van het systeem, ter informatie of meer begrip, en niet geheel onbelangrijk, zoals we in de vorige paragraaf zagen, vertrouwen in een AI-systeem (Ribera, 2019). Ook empirisch onderzoek heeft aangetoond dat een uitleg van belang is om vertrouwen te wekken bij gebruikers (Herlocker, 2000). Het meeste onderzoek op het gebied van transparantie in de vorm van een uitleg heeft zich echter tot nu toe vooral beperkt tot de intuïtie van de onderzoekers met betrekking tot wat een ‘goede’ uitleg is: er is weinig oog voor de bestaande literatuur en modellen omtrent het begrip ‘uitleg’ vanuit de sociale wetenschappen (Miller, 2017). Miller (2017) citerend: ”Beware of the inmates running the asylum”, waarmee hij doelt op het fenomeen dat XAI-onderzoekers meer bezig zijn met een begrijpelijke uitleg voor zichzelf te verschaffen, dan voor de vermeende gebruikers. Toch ligt hier mijns inziens een kans om transparantie en diens relatie met vertrouwen beter te begrijpen en praktisch vorm te geven. Het openen van de black box is wellicht geen noodzaak: Menselijk denken is ook niet transparant, en toch vertrouwen wij in de meeste gevallen een ander persoon als die een ‘goede’ uitleg kan geven van zijn beweegredenen.
Grofweg kunnen we drie types van uitleg onderscheiden: de ‘wat’-uitleg, de ‘hoe’-uitleg en de ‘waarom’-uitleg (Miller, 2019). Een ‘wat’-uitleg onthult slechts het bestaan van een algoritme dat ten grondslag ligt aan een beslissing. Voor een leek is het vaak niet duidelijk hoe een beslissing gemaakt wordt, en in dit geval zal er onthuld worden dat de beslissing niet door een mens maar door een algoritmisch beslissingssysteem gemaakt is. Een ‘hoe’-uitleg veronderstelt dat de ‘black box’ geopend wordt: het is een exacte beschrijving van alle interne processen die een algoritme doorloopt om tot een bepaalde uitkomst te komen (Tintarev, 2007). Bij een ‘waarom’-uitleg blijft het beslissingssysteem een black box; de motivatie achter een bepaald besluit wordt uitgelegd, maar de onderliggende algoritmes en processen blijven buiten beschouwing. Onderzoek op het gebied van de sociale wetenschappen (Miller, 2017) laat daarnaast zien dat een goede uitleg contrastief is. Contrastief wil zeggen dat niet alleen uitgelegd wordt waarom er tot een bepaalde beslissing gekomen is, maar vooral waarom er niet

tot een tegengestelde beslissing is gekomen (‘waarom P, en niet Q’). Dit zullen we de ‘waarom niet’-uitleg noemen.
In dit onderzoek zal de hoeveelheid transparantie gemanipuleerd worden door het geven van verschillende soorten uitleg: wat, waarom en waarom niet. In tabel 3 staan deze variabelen nogmaals uitgelegd. De ‘hoe’-uitleg zullen we hier achterwege laten, omdat dit soort uitleg veronderstelt dat de ‘black box’ geopend wordt, en zoals we hebben gezien is dit voor een eindgebruiker een niet optimale uitleg (Wachter, 2017), in de zin dat de onderliggende algoritmische constructen van dusdanige complexe aard zijn dat de meeste eindgebruikers een dergelijke ‘uitleg’ niet zullen begrijpen. Bovendien hebben we gezien dat volledige transparantie ook problemen met zich meebrengt, die niet eenvoudig op te lossen zijn. Een andere reden om deze uitleg niet mee te nemen is het feit dat er uit andere onderzoeken is gebleken dat volledige transparantie juist een averechts effect kan hebben (Grimmelikhuijsen, 2012; de Fine Licht, 2014; Kizilcec, 2016)). Vanuit de hoek van de sociale wetenschappen onderschrijft Miller (2017) ook deze gedachte door aan te geven dat niet alles uitgelegd hoeft te worden (“een uitleg wordt geselecteerd op een bevooroordeelde manier”).

Uitleg	**Omschrijving**
Wat?	Een ‘wat’-uitleg onthult eigenlijk niets meer dan het bestaan van een algoritme dat de uitkomst van een systeem bepaalt.
Hoe?	Een ‘hoe’-uitleg verklaart hoe een systeem tot een conclusie komt. Het legt de stappen uit hoe er bij een input gekomen wordt tot een output. De redenering en interne processen van een systeem worden uitgelegd (het openen van de black box). Deze uitleg wordt niet meegenomen in het onderzoek.
Waarom?	Een ‘waarom’-uitleg zorgt voor een rechtvaardiging van een systeem en zijn uitkomsten, zonder de onderliggende structuren prijs te geven. In dit geval blijven de onderliggende algoritmes een ‘black box’.
Waarom niet?	Een ‘waarom niet’-uitleg, ofwel een contrastieve uitleg (waarom P, en niet Q), geeft inzicht in waarom een bepaalde beslissing is genomen, door de negatie ervan uit te leggen. Ook hier blijven de onderliggende structuren buiten beschouwing.

Tabel 3: De variabelen en hun omschrijving

2.10 Het onderzoeksmodel

Teneinde het begrip transparantie in relatie tot vertrouwen te kunnen onderzoeken, zullen we de volgende definitie van transparantie hanteren. De definitie is een combinatie van type 2 en type 4 van transparantie zoals Weller (2019) ze definieert in zijn artikel ‘Challenges for transparency’ (zie ook paragraaf 2.7).

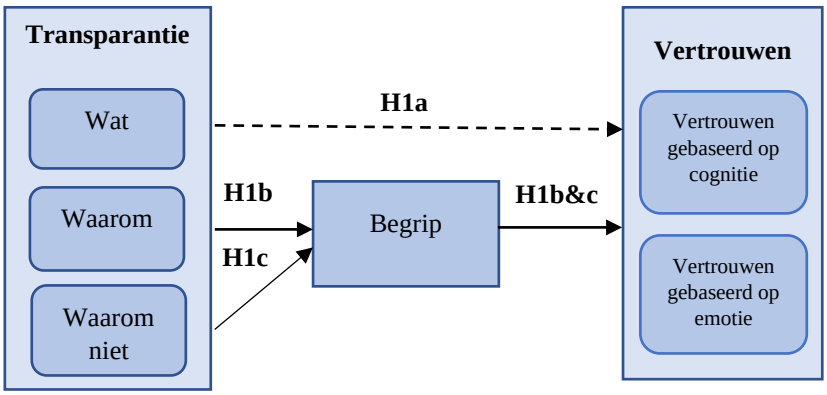
> *Transparantie betekent voor een gebruiker van een AI-systeem dat hij of zij meer begrijpt wat het systeem doet, zodanig dat een gebruiker een dergelijk systeem kan vertrouwen.*

In paragraaf 2.9 hebben we gezien dat transparantie verkregen kan worden door het geven van een uitleg (Tintarev, 2007). In paragraaf 2.8 is de relatie tussen transparantie en vertrouwen

besproken, waarbij we gezien hebben dat uit meerdere onderzoeken is gebleken dat er een verband bestaat tussen deze variabelen, en dat een hogere mate van transparantie in veel gevallen leidt tot meer begrip en gepercipieerd vertrouwen (Grimmelikhuijsen, 2014; Sinha, 2002). Hoewel de context in deze onderzoeken een andere is dan die wij zullen gebruiken, verwachten we een positief effect van transparantie op gepercipieerd vertrouwen. Hierbij stellen we het volgende:

Hypothese 1: Transparantie
Een hogere mate van transparantie met betrekking tot hoe een beslissing tot stand is gekomen met behulp van artificiële intelligentie (een geautomatiseerd beslissingssysteem dat gebruik maakt van algoritmes), leidt tot meer begrip en daarmee tot een hogere mate van gepercipieerd (cognitief) vertrouwen in (de uitkomst van) een dergelijk systeem.

Tevens zijn er in paragraaf 2.9 drie soorten uitleg onderscheiden die we willen gebruiken om de mate van transparantie in ons experiment te manipuleren. Verondersteld wordt dat een uitleg de begrijpelijkheid van het algoritmisch systeem verhoogt, wat vervolgens leidt tot meer vertrouwen. In figuur 5 wordt dit schematisch weergegeven. Een 'wat'-uitleg is niets meer dan een beschrijving van de situatie, met daarbij opgemerkt dat de beslissing die genomen is, gemaakt is door een algoritmisch beslissingssysteem. Omdat er geen uitleg gegeven wordt met betrekking tot de inhoud van de beslissing, verwachten we geen tot weinig invloed van deze uitleg op het begrip en vertrouwen dat iemand heeft in het algoritmisch systeem.



Figuur 5: Het conceptuele model.

Hypothese 1a: de 'wat'-uitleg
Een 'wat'-uitleg heeft geen tot weinig invloed op het begrip van en het gepercipieerde vertrouwen in een algoritmisch beslissingssysteem.

De 'waarom'-uitleg geeft een beschrijving van de situatie, met daarbij de redenen waarom er door het algoritmisch systeem tot een bepaald besluit overgegaan is, zonder de exacte werking van het algoritme prijs te geven. Dit soort uitleg wordt verondersteld door mensen goed te worden begrepen (Miller, 2017), en tevens wordt bij deze hogere mate van transparantie een positieve invloed op gepercipieerd vertrouwen verwacht (Grimmelikhuijsen, 2014; Sinha, 2002): een geïnformeerd persoon vertrouwt meer op de technologie die hem/haar voorgeschoteld wordt dan iemand die er geen informatie over krijgt.

Hypothese 1b: de 'waarom'-uitleg
Een 'waarom'-uitleg heeft een positieve invloed op het begrip van en het gepercipieerde vertrouwen in een algoritmisch beslissingssysteem.

Een contrastieve ('waarom niet') uitleg geeft een beschrijving van de situatie, met daarbij een contrasterende verklaring. Er wordt dus uitgelegd waarom er níet besloten is tot het tegengestelde geval. Dit soort uitleg blijkt uit onderzoek binnen de sociale wetenschappen in intermenselijke relaties veel gebruikt te worden (Miller, 2017). Daarnaast blijkt uit Miller's uitgebreide review (2017) dat mensen psychologisch gezien een voorkeur hebben voor een contrastieve uitleg. Dit leidt tot de volgende hypothese:

Hypothese 1c: de contrastieve (waarom niet) uitleg
Een 'waarom niet'-uitleg (contrastieve) uitleg heeft een positieve invloed op het begrip van en het gepercipieerde vertrouwen in een algoritmisch beslissingssysteem, meer nog dan een 'waarom'-uitleg.

3. Methodologie

In hoofdstuk 2 is de theorie besproken met betrekking tot de onderzoeksvragen. Er zijn een aantal vragen naar voren gekomen die onbeantwoord zijn gebleven en die het uitgangspunt vormen voor dit hoofdstuk. In dit hoofdstuk zal de methode van onderzoek worden toegelicht en zal het onderzoek nader worden uitgewerkt.

3.1 Onderzoeksmethode en gemaakte keuzes

Alvorens overgegaan wordt tot de bespreking van het ontwerp van het onderzoek, wordt kort stilgestaan bij de gemaakte keuzes, en hoe deze keuzes tot stand zijn gekomen. Grofweg kunnen we twee methoden onderscheiden bij het doen van onderzoek: kwalitatief onderzoek en kwantitatief onderzoek. Daarnaast wordt ook vaak een mix van deze methodes toegepast (Saunders, 2019). Kwalitatieve onderzoeksmethoden worden vaak toegepast om meningen te verzamelingen van participanten, en hier een gemene deler in te ontdekken. Het wordt vaak gebruikt om nieuwe theorieën te vormen. Kwantitatief onderzoek onderscheidt zich van kwalitatief onderzoek door de intentie om een relationeel verband tussen variabelen aan te willen tonen. Deze variabelen worden op een numerieke manier gemeten en geanalyseerd. Het grootste voordeel van kwantitatief onderzoek is dat de resultaten goed meetbaar zijn, omdat er gebruik wordt gemaakt van statistiek om de (numerieke) gegevens te analyseren. Kwantitatief onderzoek is in die zin ook meer objectief dan kwalitatief onderzoek, omdat de onderzoeker een minder grote rol speelt bij de interpretatie van de resultaten. Een nadeel van kwantitatief onderzoek is dat er niet doorgevraagd kan worden indien de resultaten van het onderzoek onverwacht zijn. Er wordt van tevoren een bepaald onderzoek uitgestippeld, en bij onverwachte resultaten kunnen de redenen hiervoor alleen achterhaald worden door een nieuw onderzoek te starten.
In dit onderzoek zal worden getracht om een verband aan te tonen tussen variabelen, waardoor het onderzoek in de basis kwantitatief van aard zal zijn. Daarnaast zal het onderzoek een exploratief en deels evaluerend of verklarend karakter hebben. Het specifieke verband tussen de variabelen die zijn gekozen is nog niet eerder bekeken, maar een aantal elementen uit deze studie zijn reeds onderzocht: er is bijvoorbeeld aangetoond dat transparantie in een bepaalde context een relatie heeft met vertrouwen (Grimmelikhuijsen, 2014; Kizilcec, 2016).

3.2 Onderzoeksopzet

Het onderzoek zal worden uitgevoerd door een (online) experiment op te zetten. Experimenteel onderzoek wordt vaak gezien als de ‘gouden standaard’ binnen de onderzoekswereld (Bhattacherjee, 2012). In deze opzet worden een of meer variabelen gemanipuleerd door de onderzoeker (als ‘behandeling’), de onderzochte subjecten worden willekeurig ingedeeld in verschillende ‘behandelgroepen’, en de resultaten van de ‘behandeling’ op de afhankelijke variabelen worden geobserveerd. Experimenteel onderzoek is zeer geschikt voor verklarend onderzoek, waarbij het doel is om relaties tussen variabelen te onderzoeken. In dit onderzoek willen we de relatie tussen een bepaalde uitleg en het begrip van en het vertrouwen in een AI-systeem onderzoeken. De gemanipuleerde variabele is hierbij het soort uitleg, en de afhankelijke variabelen zijn de begrijpelijkheid van de uitleg en het gepercipieerde vertrouwen in het AI-systeem waarover de uitleg gegeven wordt. In tabel 4 zijn de variabelen schematisch weergegeven.

Variabele	Soort variabele	Omschrijving
'Wat'-uitleg	onafhankelijk	Deze uitleg omschrijft dat er gebruik wordt gemaakt van een algoritme om tot een beslissing te komen, maar geeft geen nadere specificatie.
'Waarom'-uitleg	onafhankelijk	Deze uitleg beschrijft waarom er tot een bepaalde beslissing is gekomen; welke
'Waarom niet- uitleg	onafhankelijk	Deze uitleg wordt ook wel een contrastieve of 'counterfactual' uitleg genoemd
Waargenomen begrijpelijkheid	afhankelijk (bemiddelende variabele)	De begrijpelijkheid van de uitleg zoals waargenomen door de respondent. De verwachting is dat deze variabele een bemiddelende rol speelt tussen een bepaalde uitleg en het gepercipieerde vertrouwen, met andere woorden, het geven van een uitleg wordt verondersteld de begrijpelijkheid te verhogen, met als resultaat dat het waargenomen vertrouwen verhoogd wordt.
Waargenomen vertrouwen op basis van cognitie	afhankelijk	Het gepercipieerde vertrouwen van de respondent op basis van cognitie in de uitkomst van het algoritmisch beslissingssysteem
Waargenomen vertrouwen op basis van emotie	afhankelijk	Het gepercipieerde vertrouwen van de respondent op basis van emotie in de uitkomst van het algoritmisch beslissingssysteem

Tabel 4: overzicht van de variabelen

De procedure bestaat uit drie onderdelen: een kort welkomstwoord met de instructies van het experiment, de presentatie van de casus, en een vragenlijst na het aanbieden van de casus. Het experiment wordt afgenomen door de respondenten een (hypothetische) casus (scenario) voor te leggen en hierover een vragenlijst in te laten vullen. De antwoorden op de vragen worden vastgelegd in een 5-punts Likert schaal ("geheel mee oneens" tot "geheel mee eens"). Hypothetische scenario's worden binnen het onderzoek op het gebied van sociale psychologie en ethiek veelvuldig gebruikt om meningen, overtuigingen en de houding van mensen ten opzichte van een bepaald onderwerp in kaart te brengen, en studies hebben aangetoond dat het gedrag van mensen in dergelijke experimenten goed overeenkomt met hun gedrag in de realiteit (Woods, 2006). De respondenten worden willekeurig toegekend aan een scenario met één van de drie mogelijke transparantie-niveaus: een scenario met een 'wat'-uitleg (controlegroep), een scenario met een 'waarom'-uitleg en een scenario met een 'waarom niet'-uitleg. Per scenario is er een minimum van 15 respondenten vereist om van een voldoende betrouwbaar onderzoek te kunnen spreken, en om een onderbouwde statistische analyse te kunnen uitvoeren.

3.3 Participanten

De respondenten in dit onderzoek zijn inwoners van Nederland, die op het moment van het onderzoek (december 2020) 18 jaar of ouder waren. De deelnemers zijn geworven door de enquête te verspreiden via diverse sociale mediakanalen. Er is gekozen om een actueel onderwerp (door de overheid genomen maatregelen in de bestrijding van het COVID-19 virus) te kiezen voor het scenario in de enquête om de betrokkenheid van de respondenten bij de casus te vergroten. Omdat factoren als leeftijd, geslacht en opleidingsniveau een rol kunnen spelen bij het door respondenten gerapporteerde begrip en gepercipieerde vertrouwen, worden deze kenmerken uitgevraagd in de enquête.

3.4 Meetwaarden

In deze paragraaf wordt kort stilgestaan bij de te meten waarden. In bijlage 2 staan de vragen per onderdeel zoals die in de enquête zijn gebruikt.

Voorkennis: de algemene kennis die men voorafgaand aan het experiment heeft ten aanzien van artificiële intelligentie.

Begrip: het begrip ten aanzien van het AI-systeem uit de casus; dit wordt gemeten na het voorleggen van de casus.

Predispositie tot vertrouwen: het algemene vertrouwen dat men voorafgaand heeft in artificiële intelligentie en in de overheid.

Vertrouwen: het vertrouwen ten aanzien van het AI-systeem uit de voorgelegde casus. Zoals we in paragraaf 2.8 hebben gezien kan er een onderscheid gemaakt worden tussen *op emotie gebaseerd vertrouwen* en *op cognitie gebaseerd vertrouwen*. Dit onderscheid wordt meegenomen bij het opstellen van de vragenlijst. Voor de totstandkoming van de vragenlijst en de volledige enquête, zie bijlage 2.

In dit onderzoek wordt meer effect verwacht op het vertrouwen gebaseerd op cognitie, omdat we door middel van het geven van een uitleg het effect op de begrijpelijkheid en het daarmee veronderstelde samenhangende cognitieve vertrouwen willen onderzoeken.

3.5 Gegevensanalyse

Na afronding van de enquêtes zullen de resultaten geïnterpreteerd worden middels een aantal controlemiddelen en statistische analyses. Allereerst wordt er een factoranalyse gedaan (veel gebruikt bij onderzoeken in de vorm van een enquête) om te bepalen of de vragen waarvan vooraf is bepaald dat ze gezamenlijk een bepaald concept meten, ook daadwerkelijk dat concept meten. Met andere woorden: meten we ook werkelijk wat we willen meten? Daarnaast wordt de consistentie tussen de antwoorden op de gestelde vragen om een bepaald concept (in dit geval vertrouwen) gemeten, door gebruik te maken van Cronbach's alpha. Een waarde boven 0.7 duidt erop dat de vragen een voldoende interne consistentie hebben.
Qua statistische analyses wordt gebruik gemaakt van one-way analysis of variance (ANOVA), waarmee de variantie, ofwel de spreiding van de datawaarden gemeten wordt. Dit instrument wordt veelvuldig gebruikt voor het interpreteren van numerieke variabelen, om te verifiëren dat de verschillen tussen de variabelen significant zijn. Een ANOVA-toets veronderstelt een normale verdeling van de data, wat door middel van Levene's test aangetoond kan worden. Indien er geen normaalverdeling is, kan worden uitgeweken naar een Kruskal-Wallis H test: de non-parametrische variant van ANOVA. Vervolgens kan er met de T-test of de Mann-Whitney U Test (vooral gebruikt bij kleinere steekproefomvang of wanneer de verdeling van de data scheef is) gekeken worden waar de verschillen precies zitten.

3.6 Validiteit en betrouwbaarheid

In deze paragraaf worden de validiteit en de betrouwbaarheid van het onderzoek besproken. Validiteit of geldigheid van een onderzoek betekent in het algemeen de mate waarin de resultaten van het onderzoek daadwerkelijk overeenkomen met hetgeen men wilde meten en dat het onderzoek generaliseerbaar is. Er kan hierbij onderscheid worden gemaakt tussen interne en externe validiteit. Interne validiteit zegt iets over of de bevindingen daadwerkelijk

toegeschreven kunnen worden aan hetgeen is onderzocht, terwijl externe validiteit te maken heeft met de generaliseerbaarheid van het onderzoek. Onderzoek in de vorm van een (online) experiment zoals in deze studie het geval is, staat bekend om zijn hoge interne validiteit (ook wel causaliteit), omdat onderzoek in deze vorm de unieke mogelijkheid biedt om oorzaak en gevolg aan elkaar te koppelen door de manipulatie van de zogenaamde behandelgroep(en) en controlegroep, terwijl andere externe factoren zoveel mogelijk buiten beschouwing of gelijk worden gelaten (Bhattacherjee, 2012). De externe validiteit (generaliseerbaarheid) kan echter variëren, omdat men in werkelijkheid minder controle kan uitoefenen op externe factoren. Hier dient dus rekening mee gehouden te worden bij het interpreteren van de resultaten van het onderzoek. Betrouwbaarheid van een onderzoek zegt iets over de repliceerbaarheid en consistentie; als er in het onderzoek een eerder onderzoeksmodel gerepliceerd wordt, en dezelfde bevindingen gedaan worden, dan wordt dit onderzoek als betrouwbaar gezien. In tabel 6 wordt weergegeven welke maatregelen er zijn genomen om in het huidige onderzoek de validiteit en betrouwbaarheid zo veel als mogelijk te kunnen garanderen.

Maatregel	
Interne validiteit	De methode van het onderzoek is geschikt als meetinstrument. Een experiment in de vorm van een online enquête is in eerdere onderzoeken op dit gebied gebruikt, en geschikt bevonden als meetinstrument.
	De vragen uit de enquête zijn zoveel mogelijk overgenomen uit andere onderzoeken, zodanig dat de samenhang van de vragen reeds statistisch getoetst is.
	Door het opnemen van enkele vragen met betrekking tot de voorkennis van en de predispositie tot vertrouwen wordt de algemene houding van de respondenten ten opzichte van het onderzoeksonderwerp getoetst, omdat dit van invloed kan zijn op de resultaten.
	De enquête wordt online afgenomen, waardoor de onderzoeker de respondent niet kan beïnvloeden.
	De enquête wordt anoniem afgenomen waardoor sociaal wenselijke antwoorden worden voorkomen.
	De enquête wordt vooraf getest om eventuele onduidelijkheden of fouten op te sporen.
Betrouwbaarheid	Cronbach's alpha zegt iets over de betrouwbaarheid van de gebruikte schaal om iets te meten. In het onderzoek wordt er bijvoorbeeld gebruik gemaakt van verschillende vragen om iets te kunnen zeggen over het vertrouwen dat iemand heeft in een AI-systeem. Vertrouwen is een concept dat niet makkelijk meetbaar is, maar door ervoor te zorgen dat de vragen die men stelt met betrekking tot het concept 'vertrouwen' een hoge interne consistentie hebben, kan men er in het algemeen vanuit gaan dat de betrouwbaarheid van de meting hoog is, met andere woorden: je mag ervan uit gaan dat je meet wat je wilt meten, namelijk de mate van vertrouwen dat mensen ergens in hebben.
	De onderzoeker is niet bevooroordeeld ten opzichte van het onderzoeksonderwerp. Vanwege de aard van het onderzoek is er een minimale kans dat er bias optreedt.
	Het aantal respondenten dat toegekend wordt aan een bepaald scenario is voldoende om van een betrouwbare meting te kunnen spreken.
	De respondenten worden willekeurig toegekend aan een bepaald scenario, waardoor de kans op bias minimaal is.

Tabel 6: Maatregelen ten behoeve van interne validiteit en betrouwbaarheid.

4. Resultaten

In dit hoofdstuk worden de resultaten die uit het onderzoek naar voren zijn gekomen besproken.

4.1. Experimentele setting en context

Het experiment werd ontworpen om deelnemers verschillende maten van transparantie te laten ervaren door het lezen van een scenario waarin gesproken werd over een beslissing die gemaakt werd door een AI-systeem, met daarbij een uitleg hoe dat systeem tot de beslissing was gekomen. Nadat de respondenten het scenario hadden gelezen, werd ze gevraagd om een aantal vragen te beantwoorden over hun begrip van en de mate van vertrouwen in het AI-systeem dat in het scenario werd besproken. Het AI-systeem uit de casus betrof een systeem dat een beslissing maakt over bepaalde maatregelen die de overheid kan nemen om verspreiding van COVID-19 tegen te gaan. Het experiment vond plaats in Nederland, in de eerste helft van december 2020, nog voordat er in Nederland een tweede lockdown werd afgekondigd (deze situatie wordt gesuggereerd in de hypothetische scenario's).

De centrale vraag (i.e. Leidt een hogere mate van transparantie met betrekking tot hoe een beslissing tot stand is gekomen met behulp van artificiële intelligentie tot meer begrip en daarmee tot een hogere mate van gepercipieerd vertrouwen in (de uitkomst van) een dergelijk systeem?) refereert in dit geval aan de mate van begrip van en vertrouwen in het AI-systeem uit het scenario zoals dat gepercipieerd werd door Nederlandse burgers aan de vooravond van een tweede lockdown ter bestrijding van COVID-19.

4.2 Responsen

In de periode van 1-12-2020 tot en met 21-12-2020 is de data ten behoeve van het onderzoek verzameld. In totaal hebben 114 respondenten de enquête gestart. 70 respondenten hebben uiteindelijk de gehele enquête doorlopen, en alle vragen beantwoord. Het responspercentage komt hiermee op 61%. De gemiddelde tijd die men nodig had om alle vragen te beantwoorden was 8 minuten en 25 seconden (mediaan 7 minuten en 12s). De minimale tijd was iets minder dan 4 minuten, en de maximale tijd was iets meer dan 23 minuten. Het aantal deelnemers per conditie is weergegeven in tabel 7. Om er zeker van te zijn dat de antwoorden van de respondenten betrouwbaar zijn, wordt er eerst (handmatig) gekeken of de respondenten niet op elke vraag hetzelfde antwoord hebben gegeven, en of het antwoord op de controlevraag niet te veel afwijkt van de rest van de antwoorden in dezelfde categorie van de betreffende respondent. Tevens wordt er gekeken of er geen respondenten zijn die de enquête te snel hebben doorlopen, waardoor het de vraag is of de antwoorden wel betrouwbaar zijn.

CONDITIE	N
Wat	27
Waarom	18
Waarom niet	25

Tabel 7: aantal respondenten per conditie

4.3 Kenmerken van de respondenten

In totaal hebben 70 respondenten de enquête volledig doorlopen: 38 mannen (54%), 31 vrouwen (44%), en 1 wil niet zeggen (1%). Het merendeel van de respondenten valt in de leeftijdscategorie van 25 tot 45 jaar (56%). De leeftijdscategorie van 45 tot 65 jaar is ook goed vertegenwoordigd met 36% van de deelnemers. De categorieën tot 25 jaar, en ouder dan 65 waren het minst vertegenwoordigd met respectievelijk 6% en 3% van de deelnemers. Het

opleidingsniveau van de populatie ligt vrij hoog met 76% van de respondenten die een HBO of universitaire opleiding hebben genoten, 21% MBO, en 3% middelbare school.

In tabel 8 is de compositie per groep weergegeven, met daarbij ook de gemiddelde scores op de controlevariabelen voorkennis, predispositie tot vertrouwen in AI, en predispositie tot vertrouwen in de overheid. Deze achtergrondkenmerken hebben mogelijk een effect op het waargenomen begrip en vertrouwen (Grimmelikhuijsen, 2012), en daarom is het van belang om te kijken hoe de verdeling per groep op deze kenmerken is.

Groep	%man	%hoog-opgeleid	%jonger dan 45	gem. voorkennis	gem. predispositie vertrouwen in AI	gem. predispositie vertrouwen overheid
Wat (N=27)	48	74	63	3,3	3,3	3,8
Waarom (N=18)	33	72	61	3,3	3,4	3,7
Waarom niet (N=25)	48	80	60	3,7	3,4	3,9

Tabel 8: demografische kenmerken en gemiddelde scores op de controlevariabelen uit het experiment.

Om te controleren of de sample compositie per groep evenredig verdeeld was op de onafhankelijke (controle-)variabelen zijn er een aantal ANOVA tests uitgevoerd (voor de uitgebreide analyse zie bijlage 3, tabel A). Hieruit bleek dat er géén sprake was van een significant verschil tussen de groepen op deze variabelen. Hieruit kunnen we concluderen dat de groepen een evenredige verdeling laten zien op de demografische kenmerken, en dat de gemiddelde voorkennis, gemiddelde predispositie tot vertrouwen in AI en de overheid per groep eveneens evenredig verdeeld zijn. Er is geen reden om aan te nemen dat de groepen op deze vlakken van elkaar verschillen.

Experiment manipulatie controle

Om te controleren of de deelnemers de mate van transparantie ook daadwerkelijk zo hebben ervaren zoals beoogd met de manipulatie van het experiment, is de vraag ‘welk scenario denkt u te hebben gelezen?’ gesteld. In tabel 9 staan de uitkomsten van de antwoorden op deze vraag per scenario. Opvallend is dat het ‘waarom niet’-scenario slechts 4 keer (van de 25) als zodanig is herkend. Het merendeel van de respondenten in deze groep was in de veronderstelling het ‘waarom’-scenario te hebben gelezen. In de aanvullende analyse in paragraaf 4.6 zal gekeken worden of de indelingen in groepen op basis van het scenario dat de respondenten gedacht gelezen te hebben een afwijkend beeld in de analyse laat zien dan de analyses met betrekking tot de indeling van de groepen in het scenario dat ze daadwerkelijk gelezen hebben.

Groep/Scenario	Wat	Waarom	Waarom niet
Wat (N=27)	22	5	0
Waarom (N=18)	3	14	1
Waarom niet (N=25)	6	15	4

Tabel 9: experiment manipulatie controle

4.4 Betrouwbaarheid en validiteit van de enquête

Zoals in paragraaf 3.6 reeds besproken is het van belang voor de betrouwbaarheid en validiteit van het onderzoek dat de vragen in de enquête de constructen meten die we willen meten. Een factoranalyse laat zien of de vragen waarvan vooraf is bepaald dat ze een bepaald construct meten, ook daadwerkelijk bij dit construct horen. Met andere woorden: wordt er in het onderzoek gemeten wat we willen meten? In dit onderzoek heeft dit met name betrekking op de concepten begrip, vertrouwen gebaseerd op cognitie en vertrouwen gebaseerd op emotie. Als er middels een factoranalyse is aangetoond welke vragen samen een construct vormen, kan vervolgens de betrouwbaarheid van de vragenlijst worden vastgesteld middels het berekenen van de Cronbach's alpha. Met deze methode wordt gekeken of de verschillende vragen die samen een construct vormen (zoals vastgesteld met de factoranalyse) op een consistente manier worden beantwoord. Een hoge score op Cronbach's alpha, vanaf 0,700, indiceert een hoge interne consistentie van de antwoorden op de vragen die samen een construct vormen (Saunders, 2019), waardoor gesteld kan worden dat de vragenlijst voldoende betrouwbaar is.

Factoranalyse

Met behulp van een factoranalyse wordt gekeken welke vragen samen een construct vormen. Uit de factoranalyse komt naar voren dat de vragen SQ001 tot en met SQ008 (vertrouwen cognitief), met uitzondering van SQ006, tezamen met SQ001 tot en met SQ004 (vertrouwen emotie) een construct vormen (component 1). Het onderscheid dat volgens de literatuur (McAllister, 1995; Madsen, 2000) aanwezig zou zijn tussen vertrouwen gebaseerd op cognitie en vertrouwen gebaseerd op emotie zien we hier niet terug. In het vervolg van deze analyse zullen we dit onderscheid daarom ook niet maken, en zal er gesproken worden over het concept ‘vertrouwen’ als geheel. Vraag SQ006 zal hier niet in meegenomen worden, omdat deze vraag, zoals uit de factoranalyse blijkt, een eigen construct vormt (component 3). Naast ‘vertrouwen’ vormt ‘begrip’ ook een eigen construct (component 2). Voor de exacte resultaten van de factoranalyse wordt de lezer verwezen naar bijlage 3, tabel B.

Cronbach's Alpha

Vervolgens wordt de Cronbach’s alpha berekend voor de afzonderlijke constructen, in dit geval vertrouwen (als geheel) en begrip. Cronbach's alpha zegt iets over de betrouwbaarheid van de vragenlijst door de interne consistentie van de vragen die bij een bepaald construct horen te meten (Saunders, 2019). Van vragen die hoog scoren op Cronbach's alpha (>0.7 (Saunders,2019)) kan geconcludeerd worden dat deze kunnen worden gecombineerd tot een schaal waarmee in het verdere onderzoek gerekend kan worden om een bepaald construct te meten.

De vragen met betrekking tot het construct ‘vertrouwen’ scoren een Cronbach's alpha van 0,858. Dit is ruim boven de norm van 0,7. Door het weglaten van vraag SQ006 wordt de Cronbach's alpha verhoogd van 0,858 naar 0,878, en zoals we ook al bij de factoranalyse hadden geconcludeerd wordt deze vraag in de verdere analyse buiten beschouwing gelaten. De vragen met betrekking tot het construct ‘begrip’ scoren een Cronbach’s alpha van 0,745. Deze score is iets lager dan voor ‘vertrouwen’, maar nog steeds voldoende om te stellen dat we ook hier te maken hebben met een valide construct. Het weglaten van vragen is hier niet van toepassing omdat er slechts twee vragen waren.

Met een uiteindelijke score van 0,878 voor vertrouwen en 0,745 voor begrip kunnen we concluderen dat de vragenlijst ruim voldoende betrouwbaar is voor verdere analyse. Voor de exacte waarden uit de analyse wordt verwezen naar bijlage 3, tabel C.

4.5 Resultaten

In hoofdstuk 2 is er een hoofdhypothese geformuleerd:

Een hogere mate van transparantie met betrekking tot hoe een beslissing tot stand is gekomen met behulp van artificiële intelligentie (een geautomatiseerd beslissingssysteem dat gebruik maakt van algoritmes), leidt tot meer begrip en daarmee tot een hogere mate van gepercipieerd (cognitief) vertrouwen in (de uitkomst van) een dergelijk systeem.

Om te toetsen of het effect van transparantie in de vorm van een uitleg significant van invloed is op de mate van begrip en vertrouwen van mensen in een toepassing van artificiële intelligentie, zijn er een aantal analyses uitgevoerd. Om een eerste indruk te krijgen van de scores per groep, is er gekeken naar de gemiddelde scores en standaarddeviaties per scenario, zie tabel 11.

Vertrouwen			
Scenario	**Gemiddeld**	**N**	**Standaarddeviatie**
Wat	3,0471	27	0,67907
Waarom	2,9848	18	0,60242
Waarom niet	3,1236	25	0,57847

Begrip			
Scenario	**Gemiddeld**	**N**	**Standaarddeviatie**
Wat	3,3148	27	0,85652
Waarom	3,8611	18	1,17330
Waarom niet	3,9800	25	0,56789

Tabel 11: gemiddelde scores, respondenten en standaarddeviaties per groep op de afhankelijke variabelen.

In de vorige paragraaf is vastgesteld dat we te maken hebben met een numerieke variabele die te meten is op een schaal, dus we kunnen nu een statistische toets gebruiken om vast te stellen wat de waarschijnlijkheid is dat de groepen die we hebben vastgesteld van elkaar verschillen buiten slechts door kans. Omdat we te maken hebben met 3 groepen (wat, waarom, waarom niet) gebruiken we hiervoor *one-way analysis of variance* oftewel ANOVA. ANOVA analyseert de variantie (de spreiding van datawaarden) binnen en tussen groepen door het vergelijken van de gemiddelden. De waarde van ANOVA wordt weergegeven als F. Een parametrische test zoals ANOVA veronderstelt een normaalverdeling, met name als de groepen die met elkaar vergeleken worden niet even groot zijn, zoals hier het geval is (N=27, N=18, N=25). Om na te gaan of er aan de voorwaarde van

Levene's test		
	LEVENE STATISTIC	**SIGNIFICANTIE**
HOMOGENITEIT BINNEN GEMIDDELDE VERTROUWEN	0,044	0,957
HOMOGENITEIT BINNEN GEMIDDELDE BEGRIP	4,735	0,012

Tabel 12: Levene's test

een normaalverdeling voldaan is, wordt er eerst een Levene's test gedaan. Voor de variabele 'vertrouwen' is er volgens de Levene's test sprake van een normaalverdeling (F(2,67)=0,044 p=0,957), zie ook tabel 12. De nulhypothese van gelijke variantie wordt niet verworpen, omdat p > 0,05. Voor de variabele 'begrip' daarentegen lijkt er geen sprake van een normaalverdeling (F(2,67)=4,735 p=0,012) (nulhypothese wordt verworpen bij p<0,05). Dit betekent dat we voor de variabele 'vertrouwen' een ANOVA-test kunnen gebruiken om te kijken of de groepsgemiddelden significant van elkaar verschillen. Voor de variabele 'begrip' is een non-parametrische test zoals de Mann-Whitney test beter geschikt.

Variabele 'vertrouwen': One-way ANOVA

Om te toetsen of de gemiddelde scores op de variabele 'vertrouwen' significant van elkaar verschillen, wordt er een ANOVA-toets gedaan. In tabel 13 staan de uitkomsten van deze toets. Uit de ANOVA-toets komt naar voren dat er *geen* significant verschil wordt gevonden voor vertrouwen: F(2,67) = 0,759. Deze p-waarde ligt een stuk hoger dan 0,05 (de waarde die bij onderzoeken met een experimenteel karakter en wat kleinere datasets veelal gehanteerd wordt), dus hieruit kan worden geconcludeerd dat er geen verschil bestaat tussen de groepen op vertrouwen.

One-way ANOVA				
	SUM OF SQUARES	**MEAN SQUARE**	**F**	**SIG**
BETWEEN GROUPS	0,177	0,088	0,276	0,759
WITHIN GROUPS	21,427	0,32		
TOTAL	21,604			

Tabel 13: Anova toets

Variabele 'begrip': Kruskal-Wallis H test

Zoals uit Levene's test bleek, ontbreekt er een normaalverdeling op de variabele 'begrip'. Om deze reden is een parametrische test zoals ANOVA minder geschikt, en zullen we een non-parametrische test doen om te kijken of er een significant verschil is tussen de groepen op deze variabele. Hiervoor wordt het non-parametrische equivalent van de ANOVA-toets gebruikt: de Kruskal-Wallis H toets. Deze toets toont aan dat er een significant verschil is in waargenomen begrip tussen de verschillende groepen, $X^2(2) = 9{,}204$, $p = 0{,}010$ met een gemiddelde score op begrip van 26,78 in groep 1, 40,53 in groep 2 en 41,30 in groep 3 (zie tabel 14). In de Kruskal-Wallis H test zijn de drie groepen met elkaar vergeleken, en er is een significant verschil gevonden, maar het is nog onduidelijk waar dat verschil nu precies zit. Om hierachter te komen worden de verschillende groepen onderling met elkaar vergeleken in een volgende non-parametrische test, waarmee maximaal twee groepen met elkaar kunnen worden vergeleken: de Mann-Whitney U test. Voor groep 1 en 2 ('wat'-uitleg en 'waarom'-uitleg) wordt er een significant verschil tussen de groepen gevonden: U = 157, Z = -2,082, p = 0,037 (zie tabel 15). Voor groep 1 en 3 ('wat'-uitleg en 'waarom niet'-uitleg) wordt er ook een significant verschil gevonden: U = 188, Z = -2,995, p = 0,003. Voor groep 2 en 3 ('waarom'-uitleg en 'waarom niet'-uitleg) wordt er geen significant verschil gevonden: U = 220,5, Z = -0,119, p = 0,906. Groep 2 en 3 verschillen dus beide significant van groep 1, de controlegroep, maar niet van elkaar.

KRUSKAL-WALLIS H			
	GROEP	N	MEAN RANK
BEGRIP	Wat	27	26,78
	Waarom	18	40,53
	Waarom niet	25	41,3
	Totaal	70	

	BEGRIP
KRUSKAL-WALLIS H	9,204
DF	2
ASYMP. SIG.	0,01

Tabel 14: Kruskal-Wallis H toets

MANN-WHITNEY U				
	GROEP	N	MEAN RANK	SUM OF RANKS
BEGRIP	1	27	19,81	535
	2	18	27,78	500
	1	27	20,96	566
	3	25	32,48	812
	2	18	22,25	400,5
	3	25	21,82	545,5

Tabel 15: Mann-Whitney U test

STATISTIEKEN GROEP 1 & 2	
MANN-WHITNEY U	157
Z	-2,082
ASYMP. SIG.	0,037
STATISTIEKEN GROEP 1 & 3	
MANN-WHITNEY U	188
Z	-2,995
ASYMP. SIG.	0,003
STATISTIEKEN GROEP 2 & 3	
MANN-WHITNEY U	220,5
Z	-0,119
ASYMP. SIG.	0,906

Conclusie

De respondenten, die verdeeld waren over drie groepen, te weten een groep die een 'wat'-uitleg heeft gekregen, een groep die een 'waarom'-uitleg heeft gekregen, en een groep die een 'waarom niet'-uitleg heeft gekregen als onderdeel van het scenario, scoren *niet* significant verschillend op waargenomen vertrouwen. De groepen scoren echter *wel* significant verschillend op begrip. Dit verschil wordt gevonden tussen de groep met de 'wat'-uitleg en de 'waarom'-uitleg, en de 'wat'-uitleg en de 'waarom niet'-uitleg. Er is geen significant verschil gevonden tussen de groep met 'waarom'-uitleg en de groep met 'waarom niet'-uitleg.

4.6 Aanvullende analyses

In paragraaf 4.5 zagen we dat er geen significant resultaat werd gevonden van transparantie in de vorm van een uitleg op het gepercipieerde vertrouwen. Echter, veel respondenten die het hoogste niveau van transparantie ('waarom niet'-uitleg) hebben gezien, beoordelen dit niet als zodanig (zie ook tabel 4.3 in paragraaf 4.3). Om deze reden is er een aanvullende analyse gedaan om te kijken of het scenario dat de respondenten gelezen dachten te hebben van invloed was op gepercipieerd vertrouwen. Volgens Levene's test is er sprake van een normaalverdeling op de variabele 'vertrouwen', en kan er een ANOVA-toets gedaan worden. Ook hier wordt geen significant verschil gevonden: $F(2,67) = 0{,}532$; $p = 0{,}59$ (zie tabel 16).

LEVENE'S TEST		
	LEVENE STATISTIC	SIGNIFI-CANTIE
HOMOGENITEIT BINNEN GEMIDDELDE VERTROUWEN	0,212	0,81

One-way ANOVA				
	SUM OF SQUARES	MEAN SQUARE	F	SIGNIFI-CANCE
BETWEEN GROUPS	0,338	0,169	0,532	0,59
WITHIN GROUPS	21,266	0,317		
TOTAL	21,604			

Tabel 16: Levene's test en ANOVA toets

Voor de variabele 'begrip' is er een significant verschil tussen de groepen gevonden in paragraaf 4.5. Om te kijken of dit verschil ook aanwezig is tussen de groepen bij het door de respondenten waargenomen niveau van uitleg (m.a.w. welk scenario men dacht gelezen te hebben), wordt er nogmaals een Kruskal-Wallis H test gedaan met de nieuwe groepen. Ook hier wordt een significant verschil tussen de groepen gevonden: $X^2(2) = 10{,}683$, $p = 0{,}005$ met een gemiddelde score op begrip van 28,45 in groep 1, 43,16 in groep 2 en 27,10 in groep 3 (zie tabel 17).

KRUSKAL-WALLIS H			
	GROEP	N	MEAN RANK
BEGRIP	Wat	31	28,45
	Waarom	34	43,16
	Waarom niet	5	27,1
	Total	70	

	BEGRIP
KRUSKAL-WALLIS H	10,683
DF	2
ASYMP. SIG.	0,005

Tabel 17: Kruskal-Wallis H toets

Nu we weten dat er een significant verschil tussen de groepen is, wordt er een Mann-Whitney U test gedaan om te kijken tussen welke groepen dit verschil precies zit. Voor groep 1 en 2 wordt er een significant verschil tussen de groepen gevonden: U = 306, Z = -3,124, p = 0,002 (zie tabel 18). Voor groep 1 en 3 wordt er geen significant verschil gevonden: U = 75, Z = -0,12, p = 0,904. Tussen groep 2 en 3 ('waarom'-uitleg en 'waarom niet'-uitleg) wordt er een (licht) significant verschil gevonden: U = 45,5, Z = -1,794, p = 0,073. Dit laatste (licht) significante verschil (p < 0,10; dit significantieniveau mag gebruikt worden bij experimenten met kleinere groepsgrootte zoals hier het geval is) wordt mogelijk veroorzaakt door de geringe groepsgrootte (5 respondenten), en als gevolg hiervan kan er ten aanzien van dit verschil geen conclusie worden getrokken.

MANN-WHITNEY U				
	GROEP	N	MEAN RANK	SUM OF RANKS
BEGRIP	1. Wat	31	25,87	802
	2. Waarom	34	39,5	1343
	1. Wat	31	18,58	576
	3. Waarom niet	5	18	90
	2. Waarom	34	21,16	719,5
	3. Waarom niet	5	12,1	60,5

Tabel 18: Mann-Whitney U test

STATISTIEKEN GROEP 1 & 2	
MANN-WHITNEY U	306
Z	-3,124
ASYMP. SIG.	0,002
STATISTIEKEN GROEP 1 & 3	
MANN-WHITNEY U	75
Z	-0,12
ASYMP. SIG.	0,904
STATISTIEKEN GROEP 2 & 3	
MANN-WHITNEY U	45,5
Z	-1,794
ASYMP. SIG.	0,073

Aan de respondenten is daarnaast ook gevraagd naar het door hen ervaren niveau van transparantie van het AI-systeem in het scenario. In tabel 19 zijn de antwoorden op deze vraag per scenario weergegeven.

GERAPPORTEERDE TRANSPARANTIE					
	Niet transparant	Weinig transparant	Transparant	Volledig transparant	Totaal
Wat	12	7	8	0	27
Waarom	1	5	11	1	18
Waarom niet	1	14	10	0	25
Totaal	14	26	29	1	70

Tabel 19: Kruistabel voor gerapporteerde transparantie

Om te onderzoeken of het gepercipieerde niveau van transparantie invloed heeft op het waargenomen vertrouwen wordt er opnieuw een ANOVA-toets gedaan. De resultaten staan in tabel 20. Ook hier wordt geen significant effect op vertrouwen gevonden (F (3,66) = 1,087; p = 0,361)

LEVENE'S TEST		
	LEVENE STATISTIC	SIGNIFI-CANTIE
HOMOGENITEIT BINNEN GEMIDDELDE VERTROUWEN	0,273	0,762

One-way ANOVA				
	SUM OF SQUARES	MEAN SQUARE	F	SIGNIFI-CANCE
BETWEEN GROUPS	1,018	0,339	1,087	0,361
WITHIN GROUPS	20,587	0,312		
TOTAL	21,604			

Tabel 20: Levene's test en one-way ANOVA

Voor de variabele begrip wordt er wederom een significant effect gevonden door middel van een Kruskal-Wallis H toets: $X^2(2) = 19{,}934$, p = 0,000 met een gemiddelde score op begrip van 22,18 in de groep die geen transparantie rapporteerde; 29,12 in de groep die weinig transparantie rapporteerde en 46,47 in de groep die rapporteerde het AI-systeem uit de casus transparant te vinden (zie tabel 21). (De groep die volledige transparantie rapporteerde is uit de analyse gehaald, omdat die slechts 1 respondent bevatte)

	BEGRIP
KRUSKAL-WALLIS H	19,934
DF	2
ASYMP. SIG.	0,000

Tabel 21: Kruskal-Wallis H toets

KRUSKAL-WALLIS H			
	GROEP	N	MEAN RANK
BEGRIP	Niet transparant	14	22,18
	Weinig transparant	26	29,12
	Transparant	29	46,47
	Total	69	

Om te achterhalen tussen welke groepen een significant verschil aanwezig is, worden er een drietal Mann-Whitney U testen gedaan, waarvan het resultaat zichtbaar is in tabel 22.

MANN-WHITNEY U				
	GROEP	N	MEAN RANK	SUM OF RANKS
BEGRIP	1. Niet transparant	14	17,28	249,5
	2. Weinig transparant	26	21,94	570,5
	1. Niet transparant	14	11,86	166
	3. Transparant	29	26,90	780
	2. Weinig transparant	26	20,67	537,5
	3. Transparant	29	34,57	1002,5

STATISTIEKEN GROEP 1 & 2	
MANN-WHITNEY U	144,5
Z	-1,099
ASYMP. SIG.	0,272
STATISTIEKEN GROEP 1 & 3	
MANN-WHITNEY U	61
Z	-4,177
ASYMP. SIG.	0,000
STATISTIEKEN GROEP 2 & 3	
MANN-WHITNEY U	186,5
Z	-3,488
ASYMP. SIG.	0,000

Tabel 22: Mann-Whitney U test

Zoals uit tabel 22 is af te lezen, wordt er een significant verschil gevonden tussen zowel de groep die geen transparantie rapporteerde en de groep die wel transparantie rapporteerde (U = 61; Z = -4,117; p = 0,000), als tussen de groep die weinig transparantie rapporteerde en de groep die transparantie rapporteerde (U = 186,5; Z = -3,488; p = 0,000). Tussen de groep die geen transparantie rapporteerde en de groep die weinig transparantie rapporteerde wordt geen significant verschil gevonden (U = 144,5; Z = -1,099; p = 0,272). Uit deze resultaten kan geconcludeerd worden dat de manipulatie van het niveau van transparantie een gewenst en verwacht effect heeft gehad op het ervaren begrip dat de respondenten hebben gerapporteerd.

Multivariate analyse

In de vragenlijst waren naast de vragen over gepercipieerd begrip en vertrouwen ook enkele vragen opgenomen met betrekking tot demografische factoren, voorkennis, en predispositie tot vertrouwen in AI en de overheid, omdat deze zaken van invloed kunnen zijn op de beantwoording van de vragen over gepercipieerd begrip en vertrouwen (Grimmelikhuijsen, 2012). Met behulp van een multivariate (regressie-)analyse wordt de invloed van deze variabelen simultaan geobserveerd en geanalyseerd. De resultaten staan in tabel 23.

REGRESSIE (BEGRIP)				
	Beta	t	Significantie	VIF
Wat-scenario	-0,162	-1,342	0,184	1,459
Waarom-scenario	-0,052	-0,433	0,666	1,458
Voorkennis	0,255	2,278	0,026	1,244
Geslacht	-0,159	-1,416	0,162	1,258
Leeftijd	-0,05	-0,453	0,652	1,221
Opleidingsniveau	-0,012	-0,115	0,909	1,115
Ervaren transparantie	0,452	3,947	0,000	1,307

REGRESSIE (VERTROUWEN)				
	Beta	t	Significantie	VIF
Wat-scenario	0,029	0,234	0,816	1,507
Waarom-scenario	-0,037	-0,309	0,758	1,466
Voorkennis	0,128	1,101	0,275	1,360
Geslacht	0,011	0,100	0,921	1,305
Leeftijd	0,027	0,234	0,815	1,312
Opleidingsniveau	0,100	0,943	0,349	1,142
Ervaren transparantie	0,077	0,607	0,546	1,637
Begrip	0,056	0,422	0,675	1,797
Predispositie tot vertrouwen	0,569	5,079	0,000	1,261

Tabel 23: Multivariate analyse

Significante voorspellers voor begrip zijn voorkennis ($\beta = 0{,}255$, $p = 0{,}026$) en ervaren transparantie ($\beta = 0{,}452$, $p < 0{,}0005$). Het regressiemodel is significant en de verklaarde variantie is $R^2 = 0{,}308$; $F(7{,}62) = 5{,}38$, $p < 0{,}005$. De significante voorspeller voor vertrouwen is predispositie tot vertrouwen ($\beta = 0{,}569$, $p < 0{,}0005$). Het regressiemodel is significant en de verklaarde variantie is $R^2 = 0{,}314$; $F(9{,}60) = 4{,}515$, $p < 0{,}005$.

Zoals verwacht wordt er dus gevonden dat ervaren transparantie en voorkennis voorspellers zijn voor begrip. Predispositie tot vertrouwen blijkt de grootste voorspellende factor voor het uiteindelijk gerapporteerde vertrouwen.

5. Discussie, conclusies en aanbevelingen

In dit hoofdstuk worden de conclusies uit het onderzoek besproken. Er wordt gekeken naar wat de resultaten die er gevonden zijn betekenen, en in hoeverre ze aansluiten of vergelijkbaar zijn met resultaten van eerder onderzoek op dit gebied. Daarnaast is er ruimte voor het bespreken van aanbevelingen voor vervolgonderzoek, en zal er tenslotte worden ingegaan op de beperkingen van het huidige onderzoek.

5.1 Conclusies

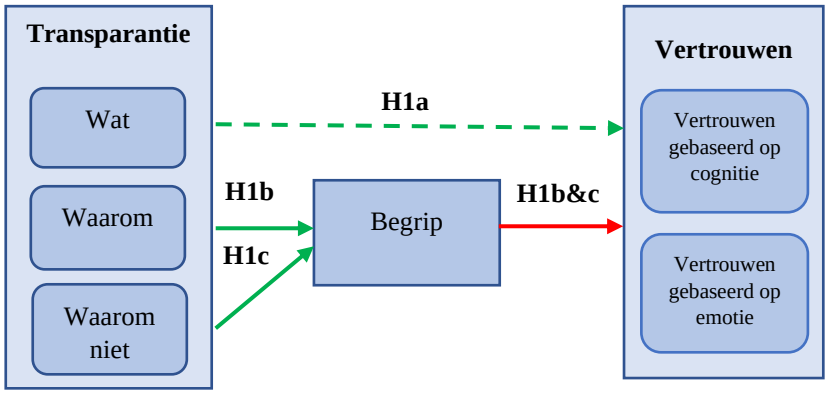
Er wordt door veel organisaties steeds meer geleund op beslissingen die gemaakt worden met behulp of met ondersteuning van AI-systemen. Deze besluiten kunnen potentieel een enorme impact hebben op de personen die ze betreffen. In dit onderzoek is onderzocht of transparantie in de vorm van het geven van een uitleg bij een dergelijke geautomatiseerde beslissing een invloed heeft op het door mensen waargenomen begrip en vertrouwen dat ze hebben in het AI-systeem dat de beslissing maakt. In het algemeen kan er geconcludeerd worden dat transparantie in de vorm van een uitleg een meetbaar effect heeft op het begrip van mensen van (een beslissing van) een AI-systeem, maar *niet* op het gepercipieerde vertrouwen in dat AI-systeem.

De volgende hypotheses zijn opgesteld in hoofdstuk 2. We zullen nu bekijken in hoeverre deze hypotheses aangenomen of verworpen kunnen worden.

> ***Hypothese 1: Transparantie***
> *Een hogere mate van transparantie met betrekking tot hoe een beslissing tot stand is gekomen met behulp van artificiële intelligentie (een geautomatiseerd beslissingssysteem dat gebruik maakt van algoritmes), leidt tot meer begrip en daarmee tot een hogere mate van gepercipieerd (cognitief) vertrouwen in (de uitkomst van) een dergelijk systeem.*
>
> ***Hypothese 1a: de 'wat'-uitleg***
> *Een 'wat'-uitleg heeft geen tot weinig invloed op het begrip van en het gepercipieerde vertrouwen in een algoritmisch beslissingssysteem.*
>
> ***Hypothese 1b: de 'waarom'-uitleg***
> *Een 'waarom'-uitleg heeft een positieve invloed op het begrip van en het gepercipieerde vertrouwen in een algoritmisch beslissingssysteem.*
>
> ***Hypothese 1c: de contrastieve ('waarom niet') uitleg***
> *Een 'waarom niet'-uitleg (contrastieve) uitleg heeft een positieve invloed op het begrip van en het gepercipieerde vertrouwen in een algoritmisch beslissingssysteem, meer nog dan een waarom-uitleg.*

Transparantie in de vorm van een 'waarom'- en 'waarom niet'-uitleg had een positief significant effect op het waargenomen begrip van (de uitkomst van) een algoritmisch beslissingssysteem, maar niet op gepercipieerd vertrouwen, waardoor hypothese 1b en 1c (deels) verworpen worden. Uit de resultaten blijkt daarnaast dat er geen reden is om hypothese 1a te verwerpen, en deze hypothese wordt aangenomen: een 'wat'-uitleg heeft geen tot weinig invloed op het begrip van en het gepercipieerde vertrouwen in een algoritmisch beslissingssysteem. Dit is niet verwonderlijk gezien het feit dat in dit onderzoek de 'wat'-uitleg in feite niets meer was dan de constatering dat er voor de gemaakte beslissing een AI-systeem gebruikt was. Deze groep kan dus ook als controlegroep gezien worden. Voor de

hoofdhypothese 1 betekent dit dat deze deels verworpen wordt: er is geen effect gevonden van transparantie op gepercipieerd vertrouwen. Wel werd er een positief significant effect waargenomen van transparantie op begrip. Het eerste deel van de hypothese - ‘een hogere mate van transparantie met betrekking tot hoe een beslissing tot stand is gekomen met behulp van artificiële intelligentie (een geautomatiseerd beslissingssysteem dat gebruik maakt van algoritmes), leidt tot meer begrip’ - houdt wel stand. Echter, er is geen reden om aan te nemen dat een ‘waarom niet’-uitleg oftewel een contrastieve uitleg in meer begrip resulteert dan een ‘waarom’-uitleg, zoals in het laatste deel van hypothese 1c wordt gesuggereerd.



Figuur 6: Schematische weergave van de conclusies.

Het in de literatuur beschreven onderscheid tussen vertrouwen gebaseerd op cognitie en vertrouwen gebaseerd op emotie wordt in dit onderzoek niet teruggevonden. Een en ander is schematisch weergegeven in figuur 6. Wat betekent dit voor de onderzoeksvraag die in hoofdstuk 1 geponeerd is?

In welke vorm kan transparantie bijdragen aan een beter begrip van en vertrouwen in (de uitkomsten van) AI-systemen bij de personen die deze uitkomsten aangaan?

Deze vraag kunnen we deels beantwoorden vanuit de resultaten van het onderzoek: transparantie in de vorm van een ‘waarom’- of ‘waarom niet’-uitleg kan bijdragen aan een beter begrip van de uitkomsten van AI-systemen bij de personen die deze uitkomsten aangaan.

Het in de literatuur veronderstelde verband tussen begrip en vertrouwen vinden we in dit onderzoek niet terug: ondanks dat de respondenten uit de groepen die een meer transparant scenario hadden gelezen meer transparantie en begrip rapporteren, leidde dit niet tot meer vertrouwen. Onze bevindingen lijken erop te wijzen dat niet meer kennis, maar vooral de predispositie tot vertrouwen in AI-systemen een groot aandeel heeft in het gepercipieerde vertrouwen in het gepresenteerde AI-systeem.

5.2 Discussie

Het doel van deze studie was het conceptualiseren van het begrip transparantie als mechanisme om verandering teweeg te brengen in gepercipieerd begrip en vertrouwen, door middel van het geven van meer informatie in de vorm van een uitleg. In het onderzoek lag de focus op het geven van drie verschillende soorten uitleg als middel van transparantie om begrip en vertrouwen in een AI-systeem te verhogen. Onze bevindingen dragen bij aan het bestaande onderzoek binnen XAI door aan te tonen dat transparantie in de vorm van een uitleg een positief effect heeft op begrip, en dat de ‘black box’ daarvoor niet noodzakelijkerwijs geopend hoeft te worden. Deelnemers aan het onderzoek gaven aan meer transparantie te ervaren bij een uitleg

waarom het AI-systeem tot een bepaalde beslissing was gekomen dan bij een uitleg waarin de beslissing niet uitgelegd werd. Ook gaven de deelnemers die een uitleg over het waarom (niet) hadden gekregen aan dat zij (de beslissing van) het AI-systeem beter begrepen. Om meer begrip te bewerkstelligen voor een leek is het dus niet nodig (en wellicht zelfs onwenselijk) om de 'black box' te openen, maar kan volstaan worden met het creëren van meer transparantie door middel van een uitleg.

De verwachting was dat indien mensen beter begrijpen waarom een beslissing wordt gemaakt, zij eerder geneigd zijn om deze beslissing (en het systeem dat de beslissing maakt) te vertrouwen (Grimmelikhuijsen, 2015). Uit ons onderzoek blijkt echter dat hier geen sprake van is: er wordt geen effect gevonden van meer transparantie op het gepercipieerde vertrouwen. Eerdere onderzoeken laten een ambigu beeld zien. In sommige onderzoeken wordt een positief effect van transparantie op vertrouwen gevonden (Grimmelikhuijsen, 2015), terwijl andere studies een neutraal of zelfs licht negatief effect vinden (De Fine Licht, 2011; Grimmelikhuijsen, 2012). Hierbij moet wel aangemerkt worden dat deze studies in allerlei verschillende contexten hebben plaatsgevonden, waardoor de resultaten wellicht niet goed met elkaar vergeleken kunnen worden. In een ander experiment dat qua context vergelijkbaar is met de context in ons onderzoek vinden Cramer et al. (2008) dat een 'waarom'-uitleg de acceptatie van een aanbevelingssysteem verhoogt, maar niet het vertrouwen in een dergelijk systeem. Dit sluit aan bij onze bevindingen waarin ook gevonden wordt dat hoewel een 'waarom'-uitleg een hogere mate van ervaren transparantie en begrip teweegbrengt, er geen effect wordt gevonden op vertrouwen. Om vertrouwen te wekken lijkt er meer nodig dan slechts meer transparantie van een AI-systeem.

Er zijn een aantal factoren die een rol kunnen spelen bij vertrouwen in technologie in het algemeen en AI-systemen in het bijzonder (Siau, 2018), die ook in dit onderzoek een rol kunnen hebben gespeeld. Ten eerste zijn er de menselijke karakteristieken, zoals bijvoorbeeld de predispositie tot vertrouwen dat een persoon in het algemeen heeft. Dit wordt in ons onderzoek ook teruggevonden: uit de resultaten blijkt dat predispositie tot vertrouwen (zowel in AI als in de overheid) een niet te onderschatten rol speelt bij het uiteindelijk gerapporteerde vertrouwen. Er zijn onderzoeken gedaan die deze bevinding ondersteunen, zoals dat door Etzioni (2016) die constateert dat mensen geen rationele besluitmakers zijn, en zich laten leiden door andere factoren dan kennis. Ten tweede noemen Siau en Wang (2018) de context waarin een AI-systeem gebruikt wordt. In dit geval wordt er een scenario geschetst waarin een AI-systeem in de vorm van een automatisch beslissingssysteem of aanbevelingssysteem gebruikt wordt om een voorspelling te doen met betrekking tot de meest geschikte maatregelen om COVID-19 te bestrijden op een bepaald moment in de tijd. Deze context kan een rol hebben gespeeld bij het gerapporteerde vertrouwen. In de scenario's die de deelnemers hebben gekregen, was er sprake van een AI-systeem dat een beslissing maakte over de te nemen maatregelen ten tijde van de COVID-19 crisis. Deze context kan ervoor gezorgd hebben dat het vertrouwen in de beslissing beïnvloed werd door externe factoren zoals de ervaren transparantie van de geldende maatregelen op dat moment of het vertrouwen dat men had in de maatregelen die op dat moment al genomen waren. Ten derde spelen de technologische kenmerken van het AI-systeem een rol bij vertrouwen. Onder deze technologische kenmerken vallen bijvoorbeeld de prestaties van het systeem en het doel dat het systeem dient, maar ook de transparantie van het systeem. In ons onderzoek wordt dit echter niet teruggevonden: meer transparantie leidde niet tot meer

vertrouwen. Dit kan mogelijk verklaard worden doordat predispositie tot vertrouwen een grotere rol speelt dan transparantie bij de mate van vertrouwen in een AI-systeem.

Een ander punt van discussie is de vraag wat een goede uitleg nu precies is. In dit onderzoek werd gevonden dat het eigenlijk niet uitmaakt of er een 'waarom'- of 'waarom niet'-uitleg werd gegeven: in beide groepen wordt eenzelfde ervaren niveau van transparantie gerapporteerd, en in beide groepen wordt er eenzelfde niveau van begrip gerapporteerd dat hoger was dan het gerapporteerde begrip in de controlegroep. De 'waarom niet'-uitleg was een uitgebreidere (en contrastieve) variant van de 'waarom'-uitleg, maar in dit onderzoek heeft deze extra informatie niet het effect gehad dat verwacht werd op basis van de literatuur (Miller, 2017). Meer transparantie hoeft ook niet altijd beter te zijn, zoals blijkt uit eerdere onderzoeken (Grimmelikhuijsen, 2015; Kizilcec, 2016), waarin gevonden werd dat een overmaat aan transparantie juist kan zorgen voor minder vertrouwen. Hoewel dit laatste waarschijnlijk geen rol heeft gespeeld in ons onderzoek (de 'hoe'-uitleg is in dit onderzoek buiten beschouwing gebleven), blijft de vraag hoeveel uitleg en transparantie nu eigenlijk optimaal is.

Het in de literatuur veel aangehaalde onderscheid tussen vertrouwen gebaseerd op cognitie en vertrouwen gebaseerd op emotie wordt in dit onderzoek niet teruggevonden. Vertrouwen wordt volgens velen (o.a. McAllister, 1995; Madsen, 2000) bepaald door een mix van kennis en gevoel/emotie. De resultaten van het onderzoek wijzen er echter op dat meer kennis niet zonder meer leidt tot meer vertrouwen. Het is mogelijk dat de respondenten zich meer hebben laten leiden door hun gevoel. Dit wordt ondersteund door het feit dat predispositie tot vertrouwen een grote rol blijkt te spelen in het uiteindelijk gepercipieerde vertrouwen. Er bestaat wel de mogelijkheid dat meer begrip op de langere termijn leidt tot meer vertrouwen, maar dit is in dit onderzoek niet onderzocht. Vertrouwen een complex concept, zeker als het gaat om vertrouwen in AI, omdat er hier, anders dan bij een interpersoonlijke relatie, sprake is van een derde speler, namelijk degene die een AI-systeem heeft geproduceerd of verkoopt. Voor ons onderzoek betekent dit dat het vertrouwen dat mensen hebben in de overheid of het RIVM (in de casus aangehaald als producent van het AI-systeem) een rol kan hebben gespeeld.

In het algemeen is het van belang om in het oog te houden dat een onderzoek naar het geven van een uitleg aan een leek niet het probleem van transparantie, en de daarbij horende ethische, verantwoordings- en vertrouwens-bezwaren oplost. Toch is het van belang dat een geautomatiseerde beslissing op een zodanige manier kan worden uitgelegd, dat een leek het kan begrijpen; Het is een noodzakelijke, maar niet voldoende voorwaarde.

5.3 Aanbevelingen voor verder onderzoek

Dit onderzoek levert een bijdrage aan het bestaande onderzoek op het gebied van transparantie door de relatie tussen transparantie in de vorm van een uitleg en het begrip van en vertrouwen in AI-systemen te onderzoeken. Er is gekeken naar het effect dat transparantie heeft op het begrip van en vertrouwen in een AI-systeem dat een beslissing maakt. Er wordt gevonden dat transparantie een positief effect heeft op begrip, maar er wordt geen effect gevonden voor vertrouwen. Voor met name de verantwoording van algoritmes is, naast begrip en vertrouwen, de ervaring van redelijkheid of eerlijkheid van een dergelijke beslissing door mensen van belang. Meer begrip is een mogelijke eerste stap tot de ervaring van meer inzicht en daarmee verantwoording van algoritmes. Het zou interessant zijn om te onderzoeken of transparantie in de vorm van een uitleg een bijdrage zou kunnen leveren aan een betere verantwoording van AI.

Zoals we in hoofdstuk 2 gezien hebben is transparantie geen eenduidig begrip. In dit onderzoek zijn er inzichten vanuit de sociale wetenschappen gebruikt om transparantie in de vorm van een uitleg te verschaffen. De resultaten suggereren dat een 'waarom' of 'waarom niet'-uitleg een positieve invloed heeft op de ervaren transparantie en op begrip, maar er wordt geen verschil gevonden tussen deze groepen. Uit sociaalwetenschappelijk onderzoek bleek eerder dat een 'waarom niet'- uitleg (oftewel een contrastieve uitleg) voor mensen een optimale vorm van een uitleg is (Miller, 2017). In ons onderzoek wordt gevonden dat mensen geen bewust onderscheid maken tussen de 'waarom'- en de 'waarom niet'-uitleg. Het definiëren en testen van de verschillende soorten uitleg is een aandachtspunt dat in verder onderzoek meegenomen dient te worden. Daarnaast kan een contrastieve uitleg op meerdere manieren vorm worden gegeven. Hier is ervoor gekozen om het contrastdeel zo dicht mogelijk bij het oorspronkelijke waarom-deel te houden, om een volgens Wachter et al. (2017) ideale '*contra-feitelijke*' uitleg te geven. Verder onderzoek naar het geven van een uitleg aan een leek zal moeten uitwijzen of een wat uitgebreidere contrastieve uitleg met meerdere contrasten meer effectief is.

Eerder hebben we gezien dat context een cruciale rol speelt bij vertrouwen. Ook in dit onderzoek moeten we ons realiseren dat de context een invloed heeft gehad op het door mensen ervaren vertrouwen. Niet alleen hebben we te maken met een technologische ontwikkeling die voor veel mensen tamelijk ongrijpbaar is, maar het geschetste scenario waarin de maatregelen die door de overheid worden genomen ten tijde van de Corona-crisis, is ook een niet te onderschatten contextuele factor. In Nederland is er ten tijde van het onderzoek veel ophef geweest over deze maatregelen, en dit kan ervoor gezorgd hebben dat respondenten met een gekleurde blik naar het (hypothetische) scenario hebben gekeken. Dit kan met name van invloed zijn geweest op het vertrouwen dat mensen hebben gerapporteerd. Aangezien we ook terugvinden dat predispositie tot vertrouwen een grote invloed heeft gehad op het uiteindelijk gerapporteerde vertrouwen, is het denkbaar dat dit met de context te maken heeft gehad. In een vervolgonderzoek zou het interessant kunnen zijn om te bekijken of er in een andere context wel een effect van transparantie gevonden kan worden op vertrouwen.

In een studie van Grimmelikhuijsen (2013) werden er culturele verschillen gevonden in het effect van transparantie op vertrouwen in een overheidsinstelling. Het is goed denkbaar dat culturele verschillen ook een rol spelen bij het vertrouwen en de adoptie van nieuwe technologieën zoals artificiële intelligentie (Siau, 2018). In Nederland en in Europa hebben mensen bijvoorbeeld een andere attitude jegens robots dan in sommige Aziatische landen. Het is mogelijk interessant om te onderzoeken of transparantie in het kader van een AI-systeem zoals onderzocht in dit onderzoek andere effecten teweegbrengt bij mensen met een verschillende culturele achtergrond.

Zoals in paragraaf 5.2 besproken is vertrouwen in (toegepaste) artificiële intelligentie een dynamisch concept. Vertrouwen in technologie wordt bepaald door menselijke, en omgevings- en technologische factoren. Voor het opbouwen van initieel vertrouwen onderscheiden Siau et al. (2018) binnen de technologische factoren twee dimensies: prestatie en proces, waarbij transparantie als een van de dimensies binnen het proces genoemd wordt. Andere technologische factoren die volgens Siau et al. (2018) een rol spelen zijn beproefbaarheid, representatie, imago en reviews van gebruikers. Transparantie maakt in dit model onderdeel uit van een groter geheel in het bewerkstelligen van initieel vertrouwen. Multidisciplinair vervolgonderzoek zou zich kunnen richten op de vraag of transparantie in combinatie met andere factoren een effect heeft op het vertrouwen dat men ervaart in AI.

5.4 Aanbevelingen voor de praktijk

Uit het onderzoek is gebleken dat meer transparantie in de vorm van het geven van een uitleg een positief effect had op het begrip dat mensen hadden voor (de beslissing) van een AI-systeem dat in een hypothetisch scenario door de overheid werd ingezet om te bepalen welke maatregelen er genomen moeten worden om de verspreiding van het COVID-19 virus tegen te gaan. Er werd geen effect gevonden op het gepercipieerde vertrouwen dat mensen hadden van het AI-systeem en de gemaakte beslissing. Meer begrip leidde dus niet automatisch tot meer vertrouwen. In de praktijk betekent dit dat het verschaffen van transparantie door (overheids)organisaties die werken met geautomatiseerde besluiten van belang is om het begrip bij de betreffende personen te verhogen. Dit kan uit overwegingen van verantwoording of rechtvaardiging, maar ook als mogelijk onderdeel van het proces tot meer vertrouwen van mensen in AI-systemen in het algemeen dienen.

In de vorige paragraaf werd besproken dat predispositie tot vertrouwen een belangrijke rol lijkt te spelen bij het uiteindelijk gepercipieerde vertrouwen in het AI-systeem, ook al hebben mensen meer informatie tot hun beschikking over het AI-systeem. Dit lijkt te impliceren dat het verschaffen van meer informatie met betrekking tot hoe een beslissing tot stand is gekomen weinig invloed heeft op het vertrouwen dat mensen ervaren. Om vertrouwen van mensen in (de beslissing van) een AI-systeem te winnen, lijkt een uitleg niet voldoende. De adoptie van een nieuwe technologie zoals AI heeft tijd nodig, en hangt voor een groot deel af van individuele attitudes en publieke opinie (Kizilcec, 2016).

Verder werd in dit onderzoek geen significant verschil gevonden tussen een ‘waarom’-uitleg en een ‘waarom niet’-uitleg, wat suggereert dat het vooral van belang is om een uitleg te verschaffen, maar om het begrip van mensen te verhogen lijkt het daarbij weinig uit te maken welk soort uitleg gebruikt wordt. Er is meer (multidisciplinair) onderzoek nodig om vast te stellen welk soort uitleg en transparantieniveau voor een consument optimaal is, en of hiermee het consumentenvertrouwen (op korte of lange termijn) gestimuleerd kan worden.

5.5 Beperkingen van het onderzoek

Ondanks de zorgvuldigheid waarmee het onderzoek is opgezet, zijn er een aantal beperkingen te noemen die inherent zijn aan deze manier van onderzoek, en een aantal zaken die mogelijk in een toekomstige opzet vermeden of verbeterd kunnen worden door de bevindingen van dit onderzoek in acht te nemen.

Allereerst is er een online experiment gedaan dat gebaseerd is op hypothetische scenario’s. Hoewel dit soort scenario’s vaker worden gebruikt in sociaalpsychologische onderzoeken om de perceptie van beslissingen te onderzoeken, zullen de bevindingen van deze studie moeten worden aangevuld met andere studies die de daadwerkelijke perceptie van mensen in een dergelijke situatie onderzoeken. De scenario’s die zijn gebruikt in het experiment zijn zoveel mogelijk bij de werkelijkheid gehouden, maar de gemaakte beslissingen raakten mensen niet direct, waardoor ze het eerste persoons-perspectief en de consequenties van de beslissing voor de werkelijke wereld misten. Dit kan gevolgen hebben gehad voor het door de respondenten gerapporteerde begrip en vertrouwen.

Het aantal respondenten per groep was voldoende maar wel minimaal om van een representatief onderzoek te kunnen spreken. Daarnaast kan er een mogelijke bias zijn ontstaan door de opbouw van de sample, omdat de werving van respondenten met name heeft

plaatsgevonden via sociale media. Dit heeft tot gevolg gehad dat een groot aandeel (56%) van de respondenten 25 tot 45 jaar oud is, en het merendeel een HBO of academische opleiding heeft. Deze sample is dus niet geheel representatief te noemen voor de Nederlandse bevolking.

De scenario's zijn zo zorgvuldig mogelijk opgesteld, en voorafgaand aan het onderzoek getest, maar uit de resultaten is gebleken dat het scenario met de 'waarom niet'-uitleg weinig effect heeft gesorteerd. Respondenten die dit scenario hebben gelezen, rapporteren niet meer ervaren transparantie dan de groep die een 'waarom'-uitleg heeft gehad. Hieruit valt af te leiden dat deze scenario's eenzelfde niveau van ervaren transparantie hadden, en als zodanig niet van elkaar verschilden. Hierdoor is een mogelijk effect verloren gegaan.

Vertrouwen is een breed begrip, en er is een veelheid van redenen te bedenken waarom mensen vertrouwen hebben in technologische ontwikkelingen in het algemeen, en in AI-systemen in het bijzonder. In ons onderzoek wordt aan de respondenten een aantal stellingen voorgelegd die zij beantwoorden op een 5-puntsschaal. Echter, de afwezigheid van kwalitatieve interviews om te achterhalen waarom mensen een bepaald niveau van vertrouwen rapporteren is een beperking die inherent is aan deze manier van onderzoek doen.

Ten laatste is er in ons onderzoek geen onderscheid gevonden tussen vertrouwen gebaseerd op cognitie en vertrouwen gebaseerd op emotie. Dit is een onderscheid dat in de literatuur veelvuldig gemaakt wordt. Ondanks dat er bij het opstellen van de vragenlijst expliciet rekening mee gehouden is, is het mogelijk dat dit onderscheid verloren is gegaan door het gebruik van vragen uit verschillende bestaande onderzoeken. Hierdoor is het beperkt mogelijk om conclusies te trekken over het soort vertrouwen dat nu precies gemeten wordt.

Referenties

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

AI HLEG. (2019). Ethische richtsnoeren voor betrouwbare Kunstmatige Intelligentie. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Albu, O. B., & Flyverbom, M. (2019). Organizational transparency: Conceptualizations, conditions, and consequences. *Business & Society*, *58*(2), 268-297.

Allam, Z. and Z. A. Dhunny (2019). "On big data, artificial intelligence and smart cities." Cities **89**: 80-91.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), 973-989.

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the association for information systems*, *6*(3), 4.

Bhattacherjee, A. (2012). Social science research: Principles, methods, and practices.

Biran, O., & Cotton, C. (2017, August). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (Vol. 8, No. 1, pp. 8-13).

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Anderson, H. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Buiten, M. C. (2019). "Towards Intelligent Regulation of Artificial Intelligence." European Journal of Risk Regulation **10**(1): 41-59.

Butcher, J., & Beridze, I. (2019). What is the State of Artificial Intelligence Governance Globally?. *The RUSI Journal*, *164*(5-6), 88-96.

Cihon, P. (2019). "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development." Future of Humanity Institute.

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., ... & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, *18*(5), 455.

Dafoe, A. (2018). AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*.

De Fine Licht, J., Naurin, D., Esaiasson, P., & Gilljam, M. (2014). When does transparency generate legitimacy? Experimenting on a context-bound relationship. *Governance*, *27*(1), 111-134.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56-62.

Etzioni, A. (2016). Is transparency the best disinfectant?. *Available at SSRN 2731880*.

Felzmann, H., et al. (2019). "Robots and transparency: The multiple dimensions of transparency in the context of robot technologies." IEEE Robotics & Automation Magazine **26**(2): 71-78.

Felzmann, H., et al. (2019). "Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns." Big Data & Society **6**(1): 2053951719860542.

Gasser, U. and V. A. Almeida (2017). "A layered model for AI governance." IEEE Internet Computing **21**(6): 58-62.

Grimmelikhuijsen, S. (2012). Linking transparency, knowledge and citizen trust in government: An experiment. International Review of Administrative Sciences, 78(1), 50-73.

Grimmelikhuijsen, S., Porumbescu, G., Hong, B., & Im, T. (2013). The effect of transparency on trust in government: A cross-national comparative experiment. *Public administration review*, *73*(4), 575-586.

Grimmelikhuijsen, S. G. and A. J. Meijer (2014). "Effects of transparency on the perceived trustworthiness of a government organization: Evidence from an online experiment." Journal of Public Administration Research and Theory **24**(1): 137-157.

Grimmelikhuijsen, S., & Klijn, A. (2015). The effects of judicial transparency on public trust: Evidence from a field experiment. Public Administration, 93(4), 995-1011.

Guidotti, R., et al. (2018). "A survey of methods for explaining black box models." ACM computing surveys (CSUR) 51(5): 1-42.

Gunning, D. (2017). Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web, 2, 2.

Heald, D. A. (2006). Varieties of transparency. In Transparency: The Key to Better Governance?: Proceedings of the British Academy 135 (pp. 25-43). Oxford University Press.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In Proceedings of the 2000 ACM conference on Computer supported cooperative work (pp. 241-250).

Hosseini, M., Shahri, A., Phalp, K., & Ali, R. (2018). Four reference models for transparency requirements in information systems. Requirements Engineering, 23(2), 251-275.

Kemper, J. and D. Kolkman (2019). "Transparent to whom? No algorithmic accountability without a critical audience." Information, Communication & Society 22(14): 2081-2096.

Kitchin, R. (2017). Thinking critically about and researching algorithms. Information, Communication & Society, 20(1), 14-29.

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). A governance framework for algorithmic accountability and transparency.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. U. Pa. L. Rev., 165, 633.

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In *11th australasian conference on information systems* (Vol. 53, pp. 6-8). Brisbane, Australia: Australasian Association for Information Systems.

Magrani, E. (2018). GOVERNANCE OF INTERNET OF THINGS AND ETHICS OF ARTIFICIAL INTELLIGENCE. *Revista Direitos Culturais*, *13*(31), 153-190.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, *38*(1), 24-59.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*.

Mittelstadt, B., et al. (2019). Explaining explanations in AI. Proceedings of the conference on fairness, accountability, and transparency.

Mittelstadt, B. D., et al. (2016). "The ethics of algorithms: Mapping the debate." Big Data & Society **3**(2): 2053951716679679.

Morley, J., et al. (2019). "From what to how. An overview of AI ethics tools, methods and research to translate principles into practices." arXiv preprint arXiv:1905.06876.

Pasquale, F. (2015). The black box society. Harvard University Press.

Ribera, M., & Lapedriza, A. (2019, March). Can we do better explanations? A proposal of user-centered explainable AI. In *IUI Workshops*.

Rossi, F. (2018). "Building trust in artificial intelligence." Journal of international affairs **72**(1): 127-134.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, *31*(2), 47-53.

Sinha, R., & Swearingen, K. (2002, April). The role of transparency in recommender systems. In CHI'02 extended abstracts on Human factors in computing systems (pp. 830-831).

Tintarev, N., & Masthoff, J. (2007, April). A survey of explanations in recommender systems. In 2007 IEEE 23rd international conference on data engineering workshop (pp. 801-810). IEEE.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. arXiv preprint arXiv:1806.07552.

Vedder, A. and L. Naudts (2017). "Accountability for the use of algorithms in a big data environment." International Review of Law, Computers & Technology **31**(2): 206-224.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2), 76-99.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

Wang, W. and K. Siau (2018). Artificial Intelligence: A Study on Governance, Policies, and Regulations. Thirteenth Annual Midwest Association for Information Systems Conference (MWAIS).

Weller, A. (2019). Transparency: motivations and challenges. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 23-40). Springer, Cham.

Woods, S., Walters, M., Koay, K. L., & Dautenhahn, K. (2006, March). Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In 9th IEEE International Workshop on Advanced Motion Control, 2006. (pp. 750-755). IEEE.

Zarsky, T. Z. (2013). Transparent predictions. U. Ill. L. Rev., 1503

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values, 41(1), 118-132.

Zednik, C. (2019). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology*, 1-24.

Zerilli, J., et al. (2019). "Transparency in algorithmic and human decision-making: Is there a double standard?" Philosophy & Technology **32**(4): 661-683.

Zhao, R., Benbasat, I., & Cavusoglu, H. (2019). Transparency in Advice-Giving Systems: A Framework and a Research Model for Transparency Provision. In *IUI Workshops*.

Bijlage 1 - Totstandkoming van de literatuurlijst

In deze bijlage wordt beschreven hoe de literatuurlijst die gebruikt is voor deze scriptie tot stand is gekomen. Aan de hand van de voorlopige onderzoeksvraag die in hoofdstuk 1 is geformuleerd, zijn er zoektermen bepaald die relevant kunnen zijn bij het beantwoorden van deze vraag. De meest belangrijke term is 'transparantie'. Omdat deze term erg breed is, en ook veel gebruikt wordt buiten de wereld van AI, is ervoor gekozen om deze zoekterm altijd in combinatie met de term 'artificiële intelligentie' of 'AI' te gebruiken. Om een beter beeld te krijgen van de governance van artificiële intelligentie en in het bijzonder van het begrip transparantie, en wat daarbij komt kijken, zijn de zoektermen 1 t/m 3 (en combinaties ervan) gebruikt om artikelen te vinden (*building blocks*). Daarnaast zijn zoektermen 4 en 5 gebruikt vanwege mijn interesse in de relatie tussen transparantie in de vorm van een uitleg en vertrouwen. De zoektermen zijn vertaald naar het Engels, omdat de meeste literatuur in het Engels geschreven is.

1. *Artificiële intelligentie/algoritmes (artificial intelligence, AI, algorithm(s))*
2. *Governance*
3. *Transparantie (transparency)*
4. *Vertrouwen (trust)*
5. *Uitleg (explain, explanation(s))*

Bij het zoeken naar relevante artikelen is er bovendien waar nodig geacht een filter gebruikt op het jaartal van het verschijnen van een artikel (afgelopen decennium) om het aantal zoekresultaten te beperken, en om ervoor te zorgen dat alleen de meest recente artikelen werden gevonden, zodanig dat er een zo actueel mogelijk beeld verkregen wordt. Met behulp van de *sneeuwbalmethode* zijn later ook oudere artikelen gevonden. Daarnaast is ook de *citatie-methode* gebruikt om nieuwe artikelen te vinden op basis van reeds gevonden, relevante literatuur. De volgende bronnen zijn hierbij geraadpleegd:

- EBSCO
- Google Scholar
- Google
- EU regelgeving

In onderstaande tabel wordt weergegeven welke zoekresultaten gevonden zijn. Deze lijst heeft de basis gevormd voor het theoretisch kader; In de loop van het schrijven zijn er nog andere artikelen gevonden, omdat er behoefte ontstond aan meer informatie of verdieping op een bepaald vlak, maar deze zijn niet hier opgenomen. Alle gebruikte artikelen zijn terug te vinden in de referentielijst.

Zoekterm (en)	Bron	Afbake-ning	Resulta-ten	Artikel	Relevantie
transparency, artificial intelligence	Google Scholar	geen	148.000	Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 2053951719860542.	Een van de eerste hits bij deze zoektermen. Mooi overzichtsartikel waarin transparantie vanuit verschillende invalshoeken bekeken wordt. Tevens worden er vele interessante artikelen in geciteerd.
nvt	sneeuwbal-methode	nvt	nvt	Weller, A. (2019). Transparency: motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 23-40). Springer, Cham.	Gevonden via het artikel van Felzmann. Vooral interessant omdat er meerdere types transparantie en de stakeholders onderscheiden worden. Tevens worden de risico's van transparantie helder in kaart gebracht.
transparency, governance, artificial intelligence	EBSCO	geen	4581	Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard?. *Philosophy & Technology*, *32*(4), 661-683.	Met name relevant vanwege de link met menselijk gedrag, en de introductie van de term 'explainable AI'.
transparency, governance, algorithms	EBSCO	geen	6997	Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), 973-989.	Veel geciteerd artikel. De auteurs proberen door het in kaart brengen van de beperkingen van transparantie een weg te vinden om het black box probleem te adresseren.
nvt	sneeuwbal-methode			Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56-62.	gevonden via het artikel van Ananny et al. Veel geciteerd artikel, levert een interessante bijdrage aan de discussie rondom transparantie
trust, artificial	EBSCO	2015-heden	22.783	Rossi, F. (2018). Building trust in	Toegankelijk artikel over vertrouwen in relatie tot

intelligence				artificial intelligence. *Journal of international affairs*, *72*(1), 127-134.	AI. F. Rossi is een lid van de HLEG, en via dit artikel kwam ik in aanraking met het voor de EU gepubliceerde stuk over AI.
nvt	sneeuwbal-methode	nvt	nvt	AI HLEG. (2019). Ethische richtsnoeren voor betrouwbare Kunstmatige Intelligentie. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence	Interessant stuk met richtlijnen mbt betrouwbare AI. In dit artikel wordt het belang van vertrouwen en wat daarvoor nodig is onderstreept. Tevens een goed overzicht van de huidige stand van zaken in Europa. Gevonden via artikel van Rossi.
explanation, artificial intelligence	EBSCO	geen	56.458	Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.	Veel geciteerd artikel (643 citaties), en het meest relevante artikel volgens EBSCO bij deze zoektermen.
nvt	citatie-methode	nvt	nvt	Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.	Gevonden via de citatie-methode door het artikel van Miller. Dit artikel geeft een goed overzicht van het vakgebied van XAI.
explanations, AI	Google Scholar	geen	712.000	Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288).	Eerste hit op Scholar bij deze zoektermen. Met name interessant vanwege hun kijk op het concept ‘uitleg’.
nvt	sneeuwbal-methode	nvt	nvt	Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.	gevonden via artikel van Mittelstadt. Zarski levert met dit artikel een interessante bijdrage aan de discussie rondom geautomatiseerde beslissingssystemen.

Tabel 1: Literatuuronderzoek

Bijlage 2 - Ontwerp van de enquête

Keuze van de casus

De keuze voor het soort scenario is niet geheel willekeurig tot stand gekomen. Er zijn verschillende overwegingen gemaakt om tot de huidige casus te komen. Zo is er in eerste instantie overwogen om een scenario te kiezen waarbij een beslissing over een persoon gemaakt wordt. Echter, uit diverse studies (Kizilcec, 2016) blijkt dat een negatieve uitkomst van een algoritmisch beslissingssysteem over een persoon een andere uitwerking heeft op het gepercipieerd vertrouwen dan een positieve uitkomst. Dit maakt het lastig om conclusies te verbinden aan de uitkomsten van het onderzoek. Ook is overwogen om een meer 'neutraal' scenario te gebruiken, bijvoorbeeld een verkeerssituatie, waarbij een algoritmisch beslissingssysteem een besluit maakt over het al dan niet aanleggen van een rotonde. Hier speelt echter de betrokkenheid van de respondenten een grote rol. Indien respondenten zich niet betrokken voelen bij het scenario, heeft dit mogelijk invloed op de respons, en mogelijk ook op de kwaliteit van de respons. De keuze voor het uiteindelijke scenario (maatregelen die door de overheid genomen worden om het Coronavirus te bestrijden) is gemaakt vanwege de actualiteit van het onderwerp, alsmede de betrokkenheid van alle respondenten bij het onderwerp. Een mogelijk nadeel van deze keuze is dat veel mensen vooraf al een mening hebben over hoe de overheid het land bestuurt in tijden van deze crisis, maar er wordt geprobeerd dit wordt zoveel mogelijk af te vangen door in de vragenlijst een aantal vragen op te nemen over het vertrouwen dat mensen in het algemeen hebben in het functioneren van de overheid. Dit scenario legt meteen ook een spanningsveld bloot: genomen maatregelen betekenen in de regel minder vrijheid voor een burger, maar wellicht bij meer begrip of meer vertrouwen in de maatregelen, zullen mensen eerder geneigd zijn zich eraan te houden. In elk scenario wordt gebruik gemaakt van een ander soort uitleg om de uitkomst van het AI-systeem te verklaren (tabel i).

Wat (controlegroep)	Ervan uitgaande dat er mensen zijn die niet weten dat er algoritmes gebruikt worden door de overheid om tot bepaalde besluiten te komen, zal in dit scenario het gebruik van een algoritme onthuld worden. In de casus zal worden uitgelegd dat de beslissing genomen wordt door een algoritmisch beslissingssysteem, maar er wordt verder niet ingegaan op welke manier deze beslissing tot stand is gekomen.
Waarom	In de casus zal worden uitgelegd welke beslissing er gemaakt is, en dat de beslissing genomen is door een algoritmisch beslissingssysteem, niet door een mens. Daarnaast zal worden uitgelegd waarom deze beslissing genomen is. Er wordt antwoord gegeven op de vraag: "Waarom heeft het systeem deze beslissing gemaakt?" Kijkend naar de beslisboom wordt er dus een tak van de beslisboom uitgelegd (namelijk de tak van de uiteindelijke beslissing). Een 'waarom' uitleg draagt bij aan de acceptatie van een aanbeveling van een algoritmisch systeem, zoals gebleken is uit een studie van Cramer et al. (2008)
Waarom niet	In de casus zal worden uitgelegd welke beslissing er gemaakt is, en dat deze beslissing gemaakt is door een algoritmisch beslissingssysteem, niet door een mens. Daarnaast zal worden uitgelegd waarom een tegengestelde beslissing NIET genomen is door het systeem. Er wordt antwoord gegeven op de vraag: "Waarom P, maar niet Q?", met andere woorden: op welke manier had de beslissing veranderd kunnen worden zodanig dat deze anders was uitgevallen. Een 'waarom niet' uitleg kan natuurlijk op veel manieren gegeven worden; je kunt alle mogelijke Q's beschrijven. Om de deelnemers hier niet mee te verwarren en bias te voorkomen, wordt er gekozen om een Q te kiezen die zo verwant mogelijk is met P.

Tabel i: kenmerken van de verschillende soorten uitleg die in de casus gebruikt worden

Scenario's

De volgende hypothetische scenario's zijn gebruikt in de enquête:

Wat

Stelt u zich het volgende scenario voor: Het is begin 2021, en momenteel bevinden we ons in een algehele lockdown ter voorkoming van een verdere verspreiding van het COVID-19 virus, dat ons land al sinds vorig jaar teistert. Het aantal infecties daalt langzaam, en er wordt gezocht naar een delicate balans tussen het herstellen van de economie en het verlagen van de sociale druk enerzijds, terwijl er anderzijds gestreefd wordt naar het voorkomen van een derde golf van besmettingen. Een eerste stap om de lockdown gedeeltelijk op te heffen, is het openen van de scholen. Scholen hebben een belangrijke maatschappelijke functie, en bovendien wordt zo de druk op individuele huishoudens verminderd. De Nederlandse overheid staat voor de moeilijke taak om te besluiten of de basisscholen en middelbare scholen weer geopend kunnen worden; een gedeeltelijke of gefaseerde opening is ook een mogelijkheid. Om tot een goed geïnformeerde beslissing te komen laat de regering zich informeren door het RIVM. Het RIVM richt zich op gezondheid en een veilige en gezonde leefomgeving. Dit doen zij door wetenschappelijk onderzoek en door het verzamelen en toepassen van kennis. Het RIVM heeft een centrale rol in de bestrijding van infectieziekten; zij doet onderzoek naar het Coronavirus, en adviseert de regering over de te nemen maatregelen. De signalering van de verspreiding van COVID-19 besmettingen loopt op dit moment achter de feiten aan. Zo gaan er snel enkele dagen verloren aan het verkrijgen van testuitslagen en het achterhalen van mogelijk nieuwe besmettingen via contactonderzoek. Om hier verandering in te brengen maakt het RIVM-voorspellingen over de verspreiding van het virus met behulp van een artificieel intelligent systeem (AI-systeem), dat zij zelf heeft ontwikkeld. Het systeem maakt gebruik van data afkomstig van de Nederlandse GGD's en de ziekenhuizen.

Op basis van de voorspellingen van het AI-systeem wordt besloten om de basisscholen te heropenen.

Waarom

Stelt u zich het volgende scenario voor: Het is begin 2021, en momenteel bevinden we ons in een algehele lockdown ter voorkoming van een verdere verspreiding van het COVID-19 virus, dat ons land al sinds vorig jaar teistert. Het aantal infecties daalt langzaam, en er wordt gezocht naar een delicate balans tussen het herstellen van de economie en het verlagen van de sociale druk enerzijds, terwijl er anderzijds gestreefd wordt naar het voorkomen van een derde golf van besmettingen. Een eerste stap om de lockdown gedeeltelijk op te heffen, is het openen van de scholen. Scholen hebben een belangrijke maatschappelijke functie, en bovendien wordt zo de druk op individuele huishoudens verminderd. De Nederlandse overheid staat voor de moeilijke taak om te besluiten of de basisscholen en middelbare scholen weer geopend kunnen worden; een gedeeltelijke of gefaseerde opening is ook een mogelijkheid. Om deze beslissing te kunnen maken wordt de hulp van een artificieel intelligent systeem (AI-systeem) ingeschakeld. Dit AI-systeem is ontwikkeld door het RIVM, en het maakt gebruik van data afkomstig van de Nederlandse GGD's en de ziekenhuizen. Het AI-systeem is een zelflerend netwerk dat getraind is om voorspellingen te kunnen doen over de verspreiding van het COVID-19 virus. Op dit netwerk kunnen simulaties worden gedaan van het heropenen van scholen, waarbij wordt gekeken hoe de verspreiding van het virus zich bij een bepaald scenario

in de komende maanden ontwikkelt. Om te bepalen of de scholen weer open kunnen, doet het RIVM een simulatie op het netwerk: Het AI-systeem voorspelt dat de verspreiding van het virus onder controle gehouden kan worden, en dat de capaciteit van IC-bedden voldoende is bij een opening van de basisscholen.

Het reproductiegetal staat bij deze voorspellingen centraal: Het reproductiegetal R meet hoe snel de toename of afname van het aantal besmette mensen gaat. Het getal staat voor het gemiddeld aantal mensen dat iemand met COVID-19 besmet: Een R<1 betekent een afname van de verspreiding, en een R>1 betekent een toename. Een andere belangrijke graadmeter van het AI-systeem is de bezetting van het aantal IC-bedden. Indien de bezettingsgraad ruim onder de 100% blijft (maximaal 80%), is een heropening van de scholen denkbaar. Een mogelijk scenario is bijvoorbeeld het openen van de basisscholen, terwijl de middelbare scholen gesloten blijven. Het heropenen van de basisscholen leidt volgens het AI-systeem tot een toename van het aantal COVID-19 gevallen in de komende 2 maanden, zelfs bij een lagere overdraagbaarheid bij kinderen. Het systeem voorspelt een reproductiegetal dat de komende 2 maanden tussen 0.74 en 1.05 ligt bij heropening van de basisscholen. De gevraagde IC-capaciteit zal volgens de voorspelling van het systeem onder de norm van 1500 beschikbare bedden blijven, namelijk maximaal 72% hiervan.

Op basis van de voorspellingen van het AI-systeem wordt besloten om de basisscholen te heropenen.

Waarom niet?

Stelt u zich het volgende scenario voor: Het is begin 2021, en momenteel bevinden we ons in een algehele lockdown ter voorkoming van een verdere verspreiding van het COVID-19 virus, dat ons land al sinds vorig jaar teistert. Het aantal infecties daalt langzaam, en er wordt gezocht naar een delicate balans tussen het herstellen van de economie en het verlagen van de sociale druk enerzijds, terwijl er anderzijds gestreefd wordt naar het voorkomen van een derde golf van besmettingen. Een eerste stap om de lockdown gedeeltelijk op te heffen, is het openen van de scholen. Scholen hebben een belangrijke maatschappelijke functie, en bovendien wordt zo de druk op individuele huishoudens verminderd. De Nederlandse overheid staat voor de moeilijke taak om te besluiten of de basisscholen en middelbare scholen weer geopend kunnen worden; een gedeeltelijke of gefaseerde opening is ook een mogelijkheid. Om deze beslissing te kunnen maken wordt de hulp van een artificieel intelligent systeem (AI-systeem) ingeschakeld. Dit AI-systeem is ontwikkeld door het RIVM, en het maakt gebruik van data afkomstig van de Nederlandse GGD's en de ziekenhuizen. Het AI-systeem is een zelflerend netwerk dat getraind is om voorspellingen te kunnen doen over de verspreiding van het COVID-19 virus. Op dit netwerk kunnen simulaties worden gedaan van het heropenen van scholen, waarbij wordt gekeken hoe de verspreiding van het virus zich bij een bepaald scenario in de komende maanden ontwikkelt. Om te bepalen of de scholen weer open kunnen, doet het RIVM een simulatie op het netwerk: Het AI-systeem voorspelt dat de verspreiding van het virus onder controle gehouden kan worden, en dat de capaciteit van IC-bedden voldoende is bij een opening van de basisscholen. De middelbare scholen daarentegen zullen gesloten moeten blijven.

Het reproductiegetal staat bij deze voorspellingen centraal: Het reproductiegetal R meet hoe snel de toename of afname van het aantal besmette mensen gaat. Het getal staat voor het gemiddeld aantal mensen dat iemand met COVID-19 besmet: Een R<1 betekent een afname

van de verspreiding, en een R>1 betekent een toename. Een andere belangrijke graadmeter is de bezetting van het aantal IC-bedden. Indien de bezettingsgraad ruim onder de 100% (maximaal 80%) blijft, is een heropening van de scholen denkbaar. Een mogelijk scenario is bijvoorbeeld het openen van de basisscholen, terwijl de middelbare scholen gesloten blijven. Het heropenen van de basisscholen leidt volgens het artificieel intelligente systeem tot een toename van het aantal COVID-19 gevallen in de komende 2 maanden, zelfs bij een lagere overdraagbaarheid bij kinderen. Het systeem voorspelt een reproductiegetal dat de komende 2 maanden tussen 0.74 en 1.05 ligt bij heropening van de basisscholen. De gevraagde IC-capaciteit volgens het systeem onder de norm van 1500 beschikbare bedden blijven, namelijk maximaal 72% hiervan. Het openen van de middelbare scholen en het voortgezet onderwijs zal volgens het systeem leiden tot een reproductiegetal tussen 0.80 en 1.21 en een IC-bezetting van tussen de 85% en 120%.

Op basis van de voorspellingen van het AI-systeem wordt besloten om de basisscholen te openen, maar de middelbare scholen blijven vooralsnog gesloten.

Vragenlijst

Na de scenario's kregen de deelnemers een aantal stellingen voorgelegd die zij op een 5-puntsschaal die liep van 'helemaal oneens' tot 'helemaal eens' konden beantwoorden. De vragen waren onderverdeeld in een aantal categorieën, zoals hieronder aangegeven. De vragen met betrekking tot vertrouwen (zowel gebaseerd op cognitie als emotie) zijn grotendeels overgenomen uit een onderzocht betrouwbaar instrument om vertrouwen te meten (Madsen, 2000). De vragen uit het meetinstrument van Madsen zijn naar het Nederlands vertaald, en door een tweede persoon terugvertaald naar het Engels, om ervoor te zorgen dat de oorspronkelijke strekking van een vraag behouden blijft (volledige vragenlijst staat onderin deze bijlage). Niet alle vragen uit het meetinstrument konden gebruikt worden, omdat enkele vragen niet van toepassing waren op onze casus. Hierom is ervoor gekozen om enkele andere vragen op te nemen uit een ander bestaand onderzoek naar transparantie en vertrouwen in AI-systemen (specifiek aanbevelingssystemen) (Cramer, 2008), De vragen met betrekking tot predispositie tot vertrouwen komen uit een studie van Rader (2018).

Begrip/gepercipieerde transparantie

SQ001 Ik begrijp waarom het AI-systeem tot deze beslissing komt.

SQ002 Ik begrijp waarop het AI-systeem haar beslissing baseert.

Vertrouwen (cognitief)

SQ001 Ik vertrouw erop dat het systeem de juiste beslissing maakt.

SQ002 Er is solide kennis van dit soort problemen in het AI-systeem ingebouwd.

SQ003 Het AI-systeem levert betrouwbare prestaties.

SQ004 Het advies dat het AI-systeem geeft is net zo goed of beter dan het advies dat een zeer competent persoon zou geven.

SQ005 Ik kan erop vertrouwen dat het AI-systeem naar behoren functioneert.

SQ006 Het gebruik van dit AI-systeem brengt risico's met zich mee.

SQ007 Ik heb er vertrouwen in dat het AI-systeem kan omgaan met alle mogelijke situaties.

SQ008 Ik vertrouw het AI-systeem.

Vertrouwen (emotie)

SQ001 Ik geloof in het advies dat het AI-systeem geeft, ook al weet ik niet zeker of het correct is.

SQ002 Ik vind het een prettig idee dat er een AI-systeem gebruikt wordt om dit soort beslissingen te maken.

SQ003 Ook als ik niet helemaal zeker ben van de juistheid van een beslissing van het AI-systeem, zou ik er toch op vertrouwen dat het de correcte beslissing is.

SQ004 Als ik onzeker ben over een beslissing, zou ik eerder het AI-systeem geloven dan mijzelf.

Voorkennis

SQ001 Ik ben bekend met artificiële intelligentie en de toepassingen ervan.

SQ002 Ik begrijp hoe AI-systemen werken en hoe dit soort systemen tot een beslissing komen.

Predispositie tot vertrouwen (overheid)

L001 Over het algemeen heb ik vertrouwen in de maatregelen die de overheid neemt in een crisissituatie.

L002 Ik vertrouw erop dat de Nederlandse overheid het beste voor heeft met haar burgers.

Predispositie tot vertrouwen (AI)

SQ001 Ik denk dat de meeste systemen die gebruik maken van artificiële intelligentie acteren in het beste belang van mensen.

SQ002 Ik denk dat artificiële intelligentie ons kan helpen bij het nemen van de juiste beslissingen.

SQ003 In het algemeen worden AI-systemen goed beheerd.

SQ004 De meeste AI-systemen kunnen voldoen aan de wensen van de gebruikers.

SQ005 Ik zou me op mijn gemak voelen om een AI-systeem te gebruiken omdat dit soort systemen meestal betrouwbaar zijn.

SQ006 Ik heb er vertrouwen in dat een AI-systeem doet wat ik zou willen dat het doet.

Testen van de enquête

Nadat de enquête is vormgegeven, wordt de enquête getest door deze aan 8 personen voor te leggen. Hierbij wordt gevraagd of deze personen de enquête eenmaal volledig willen doorlopen om te kijken of de enquête technisch goed functioneert, en daarnaast is gevraagd aan de respondenten om te letten op zaken als spelfouten, onduidelijkheden, onregelmatigheden, lengte en toegankelijkheid. Hieronder is een tabel weergegeven per respondent wat de feedback was en wat er met deze feedback is gedaan.

	opmerking/ oplossing	opmerking/ oplossing	opmerking/ oplossing	opmerking/ oplossing	opmerking/ oplossing
Respondent 1	Ziekte heet COVID-19, maar wordt veroorzaakt door het nieuwe Coronavirus/ aangepast in de tekst van de scenario's	Tekst van de laatste controlevraag is onduidelijk/tekst is geherformuleerd, en er zijn bullets geplaatst ter verduidelijking			

Respondent 2	Intro ": Gegeven" hoofdletter→ kleine letter/aangepast	Scenario "terwijl er" → dubbele spatie. "RIVM. Het" → veel spaties/ aangepast	Vragen "ik zou me comfortabel voelen" → anglicisme, "ik zou me op mijn gemak voelen" en "ik zou me zelfverzekerd voelen" → ik ben er zeker van" / aangepast	"Waarom een verwant besluit negatief uit is gevallen" → loopt niet lekker	
Respondent 3	Intro "10-20 minuten tijd in beslag nemen → pleonasme	In zowat de hele tekst noem je het een vragenlijst. 1 keer noem je het enquête. Consistentie.	Geanonimiseerd ipv anoniem?	Je geeft aan dat je na elke vraag eigenlijk toestemming geeft om de beantwoording te gebruiken. Ik vroeg mij af of dit ook echt zo is, bij ons werd er bij een afgebroken vragenlijst er vanuit gegaan dat men niet meer mee wil doen en dat dit dus een intrekking is van de toestemming. Ander is er voor de mensen geen mogelijkheid om de eerder gegeven vragen als niet ingevuld te beschouwen. Meer een ding waar er bij mijn studie heel erg over gevallen werd, dit omdat het daar vaak ook ging om gevoelige informatie.	Aangezien je maar 1 schaalsoort gebruikt zou ik alvast aangeven dat 1 helemaal mee oneens is en 5 helemaal mee eens (als ik de schaal nu goed voor ogen heb). Als je wel meerdere schalen gebruikt werd er bij mijn studie altijd van je verwacht dat je het benoemd voor je stelling.. Ook al staat het er duidelijk boven en moet je wel heel raar kijken wil je het verkeerd doen
Respondent 4	De eerste pagina is echt héééél veel tekst. Dat verwacht je niet als je meedoet aan een enquête. Ik zou dat stuk opknippen in drie of vier kortere stukjes, dus dat je er steeds een alinea bij krijgt door te klikken op "volgende".	Bijna aan het einde ga je pas het algemene kennis/vertrouwen in AI meten. Maar moet je hier niet mee beginnen, nog voordat je de specifieke casus bekend maakt? Lijkt mij methodologisch een betere manier, want de casus kan mij sturen.	"Ik ben er zelfverzekerd van dat een AI-systeem doet wat ik zou willen dat het doet. > De stelling loopt wat raar. Gaat het hier ook om een systeem dat ik inricht want dan snap ik de zelfverzekerdheid. Zo niet dat lijk mij vertrouwen		

			een betere benaming		
Respondent 5	Bij veel stellingen heb ik neutraal gekozen, maar het zou fijn zijn om aan te kunnen geven waarom	Soms staat er “systeem”, dan weer “AI-systeem”. Ik zou dit overal hetzelfde houden.			
Respondent 6	Sommige vragen lijken niet van toepassing (wat-scenario) tekst in introductie opnemen. / introductie-tekst is aangepast	AI-systeem handelt in het beste belang van mensen → beter is: AI-systeem acteert in het beste belang van mensen?			
Respondent 7	Het scenario was best een lap tekst. Merk dat het dan moeilijk is om aandacht vast te houden en betrokken te blijven op enquête. Later kom je erachter dat je maar één keer zoiets hoeft te lezen, dus misschien goed om dat te benadrukken dat het niet elke keer zo’n lap lezen is? / in de introductietekst verduidelijkt dat het scenario eenmalig is.	Bij de vraag over welk scenario lijkt het me handig als je de termen ‘wat-uitleg’, ‘waarom-uitleg’ en ‘waarom-niet uitleg’ even koppelt aan de drie scenario’s die je erboven beschrijft. Je hanteert in de beschrijving erboven net iets andere termen wat verwarrend werkt. / Wat/Waarom/Waarom niet toegevoegd aan de vraag.			
Respondent 8	De vragenlijst zag er goed uit en werkte technisch goed. Ik weet niet aan wie je 'm gaat aanbieden om in te vullen, maar ik dacht wel steeds: dit richt zich wel op 'de wat meer geschoolde mens', zeg maar, want er staan best moeilijke termen in.				

Meetinstrument van Madsen (2000) (vertaald naar het Nederlands)

Vertrouwen gebaseerd op cognitie

R1 Het systeem voorziet mij altijd van het advies dat ik nodig heb om mijn beslissing te kunnen maken.

R2 Het systeem levert betrouwbare prestaties.

R3 Het systeem reageert hetzelfde onder dezelfde condities op verschillende tijdstippen.

R4 Ik kan erop vertrouwen dat het systeem naar behoren functioneert.

R5 Het systeem is consistent bij het analyseren van problemen.

T1 Het systeem gebruikt geschikte methodes om tot een beslissing te komen.

T2 Er is solide kennis van dit soort problemen in het systeem ingebouwd.

T3 Het advies dat het systeem geeft is net zo goed als het advies dat een zeer competent persoon zou geven.

T4 Het systeem gebruikt de informatie die ik ingeef op een juiste manier.

T5 Het systeem gebruikt alle kennis en informatie die beschikbaar is om tot de oplossing voor het probleem te komen.

U1 Ik weet wat er gebeurt als ik het systeem de volgende keer gebruik, want ik begrijp hoe het zich gedraagt.

U2 Ik begrijp hoe het systeem mij ondersteunt bij beslissingen die ik moet maken.

U3 Hoewel ik niet precies begrijp hoe het systeem werkt, weet ik wel hoe ik het systeem moet gebruiken om een beslissing te maken over een bepaald probleem.

U4 Het is makkelijk te volgen wat het systeem doet.

U5 De volgende keer dat ik het systeem gebruik, herken ik wat ik moet doen om het advies van het systeem te krijgen dat ik nodig heb.

Vertrouwen gebaseerd op emotie

F1 Ik geloof in het advies van het systeem ook al weet ik niet zeker dat het correct is.

F2 Als ik onzeker ben over een beslissing, geloof ik het systeem eerder dan mezelf.

F3 Als ik niet zeker ben van een beslissing, heb ik er vertrouwen in dat het systeem met de beste oplossing komt.

F4 Wanneer het systeem een ongewoon advies geeft, ben ik er zeker van dat het advies correct is.

F5 Zelfs als ik geen reden heb om te verwachten dat het systeem een moeilijk probleem op kan lossen, ben ik er toch zeker van dat het systeem het zal oplossen.

P1 Ik zou een gevoel van verlies ervaren als het systeem niet beschikbaar was en ik het niet langer zou kunnen gebruiken.

P2 Ik voel een vorm van verbondenheid met het systeem.

P3 Ik vind dat het systeem past bij mijn stijl van het maken van beslissingen.

P4 Ik vind het fijn om het systeem te gebruiken voor het maken van beslissingen.

P5 Ik heb de persoonlijke voorkeur om beslissingen te maken met behulp van het systeem

Bijlage 3 – Aanvullende statistieken

In deze bijlage staan een aantal tabellen die ter ondersteuning van de statistische analyses uit hoofdstuk 4 dienen.

Anova toets van de onafhankelijke variabelen (Tabel A)

In onderstaande tabel is te zien dat er geen statistisch significant verschil bestaat tussen de groepen (wat, waarom, waarom niet) op de onafhankelijke variabelen (p > 0,10)

ANOVA						
		SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIFICANCE
Geslacht	**BETWEEN GROUPS**	0,894	2	0,447	1,355	0,265
	WITHIN GROUPS	22,092	67	0,330		
	TOTAL	22,986	69			
Leeftijd	**BETWEEN GROUPS**	0,213	2	0,107	0,405	0,668
	WITHIN GROUPS	17,630	67	0,263		
	TOTAL	17,843	69			
Opleidingsniveau	**BETWEEN GROUPS**	2,416	2	1,208	0,262	0,770
	WITHIN GROUPS	308,284	67	4,601		
	TOTAL	310,700	69			
Predispositie tot vertrouwen in AI	**BETWEEN GROUPS**	0,057	2	0,029	0,127	0,881
	WITHIN GROUPS	15,181	67	0,227		
	TOTAL	15,238	69			
Predispositie tot vertrouwen in de overheid	**BETWEEN GROUPS**	0,391	2	0,196	0,268	0,766
	WITHIN GROUPS	48,927	67	0,730		
	TOTAL	49,318	69			
Voorkennis	**BETWEEN GROUPS**	3,376	2	1,688	1,930	0,153
	WITHIN GROUPS	58,609	67	0,875		
	TOTAL	61,986	69			

Factoranalyse (Tabel B)

	Component Matrix		
	1	2	3
Vertrouwen cognitief (SQ001)	0,699	-0,05	-0,152
Vertrouwen cognitief (SQ002)	0,604	0,006	-0,036
Vertrouwen cognitief (SQ003)	0,645	0,389	-0,37
Vertrouwen cognitief (SQ004)	0,585	-0,291	-0,191
Vertrouwen cognitief (SQ005)	0,584	0,158	0,471
Vertrouwen cognitief (SQ007)	0,493	-0,402	-0,319
Vertrouwen cognitief (SQ008)	0,725	0,001	0,04
Vertrouwen emotie (SQ001)	0,784	-0,148	0,257
Vertrouwen emotie (SQ002)	0,819	-0,154	0,177
Vertrouwen emotie (SQ003)	0,718	-0,362	0,063
Vertrouwen emotie (SQ004)	0,697	-0,03	0,13
Vertrouwen cognitief (SQ006)	-0,258	-0,056	0,642
Begrip (SQ001)	0,379	0,686	0,172
Begrip (SQ002)	0,372	0,815	-0,136

Cronbach’s alpha (tabel C)

In onderstaande tabel is de score van de vragenlijst met betrekking tot het begrip ‘vertrouwen’ af te lezen. De oorspronkelijke score is 0,858, maar door het weglaten van vraag SQ006 wordt de score verhoogd naar 0,878

Cronbach’s Alpha Vertrouwen	
Cronbach's Alpha	**Aantal items**
0,858	12

	Cronbach's alpha als item verwijderd
Vertrouwen cognitief (SQ001)	0,843
Vertrouwen cognitief (SQ002)	0,848
Vertrouwen cognitief (SQ003)	0,849
Vertrouwen cognitief (SQ004)	0,848
Vertrouwen cognitief (SQ005)	0,851
Vertrouwen cognitief (SQ006)	**0,878**
Vertrouwen cognitief (SQ007)	0,855
Vertrouwen cognitief (SQ008)	0,842
Vertrouwen emotie (SQ001)	0,834
Vertrouwen emotie (SQ002)	0,831
Vertrouwen emotie (SQ003)	0,837
Vertrouwen emotie (SQ004)	0,842

Cronbach's Alpha Begrip	
Cronbach's Alpha	**Aantal items**
0,745	2