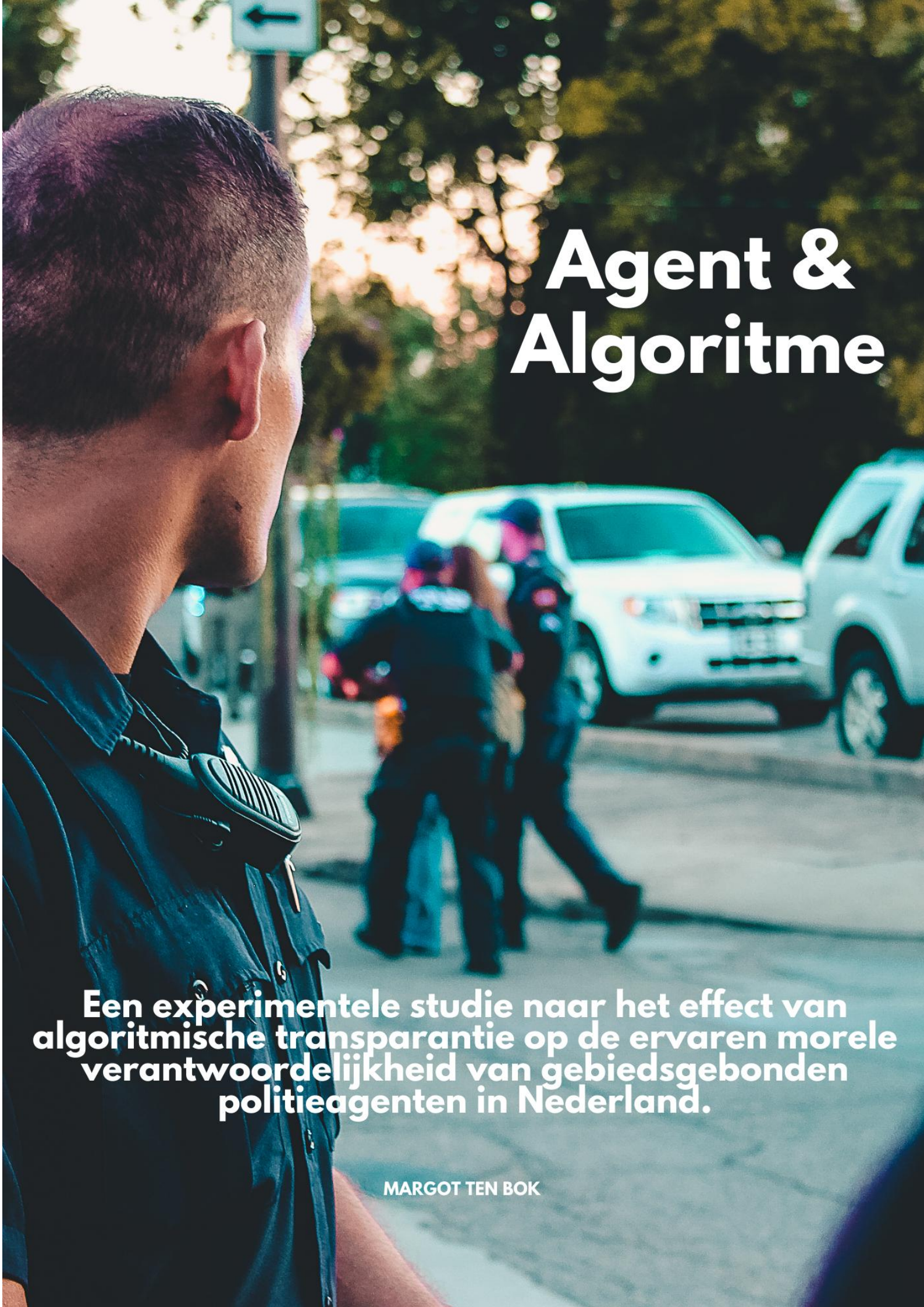


Agent & Algoritme

Een experimentele studie naar het effect van algoritmische transparantie op de ervaren morele verantwoordelijkheid van gebiedsgebonden politieagenten in Nederland.

MARGOT TEN BOK

Agent en algoritme

Een experimentele studie naar het effect van algoritmische transparantie op de ervaren morele verantwoordelijkheid van gebiedsgebonden politieagenten.

Masteropleiding Publiek Management
Margot ten Bok
6446663
Universiteit Utrecht
Departement Bestuurs- en organisatiewetenschap
3 juli 2022

Begeleiding
Eerste lezer: dr. Stephan Grimmelikhuijsen
Tweede lezer: Esther Nieuwenhuizen Msc
Politie: Carlos Soares

Voorwoord

Voor u ligt mijn masterscriptie als eindproduct van mijn vierjarige studententijd. Ik weet nog goed dat wij als voorbereiding op het schrijven van een scriptie tijdens het vak 'Kwalitatief Onderzoek' de opdracht kregen om een soortgelijk onderzoeksrapport te schrijven. Samen met Karst, Kennard, Ward en Jesse heb ik toen onderzoek gedaan naar het ethisch verantwoord inzetten van Artificial Intelligence bij het Ministerie van Infrastructuur en Waterstaat. Hoewel ik een technisch onderwerp in het begin best spannend en moeilijk vond, werd mijn interesse gewekt. Ik ging mij steeds meer verdiepen in de wereld van Artifical Intelligence. Tussen de onderwerpen die op de lijst stonden met de verschillende begeleiders erbij, trok dit onderwerp dan ook opnieuw mijn aandacht. De afgelopen paar maanden tot dit moment, juli 2022, heb ik dan ook besteed aan het schrijven van mijn masterscriptie.

Allereerst wil ik mijn scriptiebegeleider Stephan Grimmelickhuijsen bedanken voor alle concrete feedback en al het geduld. Aangezien ik nog nooit experimenteel onderzoek had gedaan zat ik met veel vragen. Tijdens het schrijfproces heb ik veel geleerd van de rol van experimenteel onderzoek in de bestuurs- en organisatiewetenschap. Ik waardeer het dat je zoveel tijd hebt gestoken in het begeleiden van mij en mijn medestudenten. Ook wil ik Esther Nieuwenhuizen bedanken voor het meedenken op zowel methodologisch als inhoudelijk vlak. Het vormgeven van een onderzoeksdesign was kort omschreven makkelijker gezegd dan gedaan. Het was fijn om onder andere over de vignetten te kunnen sparren. Verder wil ik Carlos Soares en Sandra de Wolf bedanken voor de begeleiding vanuit de Politie. Carlos wist mij met relevante actoren in contact te brengen en mee te denken vanuit praktisch oogpunt waar dat nodig was. Sandra heeft mij geholpen met het ontwikkelen van een beeld van de dagelijkse praktijk van de agent. Daarnaast wil ik natuurlijk mijn medestudenten Rosa Massop en Khadija Sanhaji bedanken voor de vele goede gesprekken die we samen hebben mogen voeren onder het genot van een cappuccino. Ook bedank ik alle agenten die hebben meegewerkt aan het onderzoek. Tot slot, bedank ik mijn moeder, vader †, broer en vrienden voor de constante motivatie en support die zij hebben gegeven gedurende mijn gehele studieperiode.

Deze masterscriptie draag ik op aan mijn vader, die ondanks zijn ziekte mij en mijn broer altijd gemotiveerd heeft om onze studie af te maken. Helaas heeft hij de uitreiking van mijn bachelor diploma al niet meer mee mogen maken. Door de motivatie die hij ons altijd heeft gegeven heb ik het afgelopen jaar de veerkracht gevonden om mijn master Publiek Management af te ronden.

Samenvatting

Artificial Intelligence (AI) is overal om ons heen. Ook de Politie ziet mogelijkheden om werk op straat veiliger, vollediger en efficiënter te maken met behulp van AI. Door deze focus op efficiëntie komen echter andere waarden zoals transparantie, rechtvaardigheid en responsiviteit onder druk te staan (Schiff, Schiff & Pierson, 2021). Door onder andere deze gebrekkige transparantie verdampt de verantwoordelijkheid in de complexe relaties rondom algoritmes (De Fine Licht & De Fine Licht, 2020). Dit is zorgelijk, want om een organisatie verantwoordelijk te kunnen stellen moeten de individuen die in die organisatie werken zich ook verantwoordelijk *voelen*.

Een mogelijk toekomstig hulpmiddel dat gebruik maakt van algoritmes is de bevragingsassistent. De bevragingsassistent is een digitaal hulpmiddel dat gebruik maakt van clusteralgoritmes, dat agenten ondersteunt doormiddel van het leveren van specifieke informatie over personen. Adequate informatievoorziening van agenten is belangrijk, omdat het kan voorkomen dat agenten of verdachten in gevaarlijke situaties belanden. Hoewel AI-systemen in staat zijn om objectieve ondersteuning te bieden, ze kunnen de beslissingen van de agent ook beïnvloeden (Zerilli et al., 2019). Dit doet de vraag rijzen waar precies tussen agent en algoritme de verantwoordelijkheid ligt. Om hier meer inzicht in te krijgen staat de volgende onderzoeksvraag in dit onderzoek centraal: *Wat is het effect van algoritmische transparantie op de ervaren morele verantwoordelijkheid van gebiedsgebonden Politieagenten?*

Deze vraag is onderzocht doormiddel van een surveyexperiment. In het experiment werden twee groepen gescheiden: 1) een groep die scenario's kregen te zien met een lage mate van algoritmische transparantie en 2) een groep die scenario's kregen te zien met een hoge mate van algoritmische transparantie. Algoritmische transparantie is in dit experiment gerealiseerd door een combinatie van een tekstuele visualisatie en een korte uitleg. Aan dit experiment hebben 153 agenten deelgenomen. Na het doorlopen van de scenario's moesten de agenten vragen beantwoorden die gingen over het toekennen van een juist of onjuist antwoord aan zichzelf of aan de bevragingsassistent. Deze schaal is gebruikt om het concept *ervaren morele verantwoordelijkheid* te operationaliseren.

Uit de resultaten blijkt het antwoord op de onderzoeksvraag drieledig te zijn:

1. Algoritmische transparantie heeft geen effect op de ervaren morele verantwoordelijkheid
2. Algoritmische transparantie zorgt wel voor meer begrijpbaarheid van het systeem en het zorgt ervoor dat fouten in een AI systeem niet onopgemerkt blijven
3. Agenten baseren hun oordeelsvorming op basis van verschillende informatiebronnen. De oordeelsvorming van agenten en de ervaren morele verantwoordelijkheid die daaruit volgt moet daarom in een informatie-ecologie gezien worden.

Op basis van deze bevindingen zijn inzichten opgedaan voor zowel de wetenschap als de praktijk. Zo schuilt er een risico voor toekomstig onderzoek dat onderzoekers in een 'tunnelvisie' alsmaar opzoek gaan naar een geschikt middel om algoritmische transparantie te realiseren, om ervaren morele verantwoordelijkheid te bewerkstelligen, terwijl het effect ervan er niet of nauwelijks lijkt te zijn. Daarnaast heerst er in de praktijk van de Politie een dominante focus op de technologie, waardoor de sociale context overschaduwd wordt. Het is belangrijk dat de bevragingsassistent uiteindelijk in de

informatie-ecologie van de agent past. Deze inzichten worden tot slot meegenomen in concrete aanbevelingen voor vervolgonderzoek en de praktijk van de Politie.

Inhoud

Hoofdstuk 1: Introductie

1.1 Aanleiding

Artificial Intelligence (AI) is overal om ons heen. Siri, Google Now of Cortana zijn voorbeelden van slimme assistenten die ons voorzien van informatie als we daarom vragen. Weten we even niet meer wat er op de agenda staat? Geen probleem, door het gebruik van Siri's slimme zoeksysteem weet je gelijk wat er bovenaan je to-do list staat. Er is geen eenduidige definitie van wat er precies verstaan wordt onder AI, maar vaak wordt de definitie van AI gekoppeld aan een systeem dat menselijke denkeigenschappen vertoont. Zo definieert Haugeland (1989, p.2) AI als '*the exciting new effort to make computers think'* en Russel en Norvig (2013, p. 3) omschrijven het als '*thinking humanly'.* Bellman koppelt AI aan activiteiten zoals het maken van beslissingen, het oplossen van problemen en het lerend vermogen van mensen (Russel & Norvig, 2013, p. 2). Ook publieke organisaties, zoals ziekenhuizen of de Politie, zien steeds vaker de kansen die AI toepassingen bieden. Zo ziet de Politie mogelijkheden om werk op straat veiliger, vollediger en efficiënter te maken met behulp van AI (interne bron).

Een manier waarop de Politie dat probeert te doen is door het ontwikkelen van Kantoor in Dienstvoertuig (K.I.D.). K.I.D. is een platform waarop verschillende applicaties kunnen worden geïnstalleerd. Een applicatie die mogelijk op K.I.D. kan worden geïnstalleerd en waar ideeën over zijn binnen de Politie is de 'bevragingsassistent'. De bevragingsassistent is een digitaal hulpmiddel voor de dagelijkse werkzaamheden van de gebiedsgebonden politieagent. In Nederland moeten gebiedsgebonden agenten een enorme hoeveelheid kennis paraat hebben. Uit gesprekken met agenten is dan ook gebleken dat ze het prettig zouden vinden om ondersteund te worden doormiddel van specifieke kennis (persoonlijke communicatie, 2022). De bevragingsassistent maakt gebruik van clusteralgoritmen met een zoekfunctie om agenten in deze specifieke kennis te kunnen ondersteunen. De bevragingsassistent zoekt woorden in registraties en koppelt deze woorden aan relevante gevarenclassificaties. Het is de ambitie dat de bevragingsassistent meldingen van personen kan opzoeken en op een logische manier kan structureren op basis van relevantie, zodat de agent altijd op de hoogte is van de meest belangrijke meldingen en hierop kan anticiperen wanneer dat nodig is.

Doordat publieke organisaties steeds vaker gebruikmaken van AI-toepassingen in het ondersteunen van dagelijkse werkzaamheden van ambtenaren, verandert een organisatie als de Politie van een street-level-bureaucracy in een screen-level of zelfs een system-level bureaucracy (Van Eck, Bovens, Zouridis, 2018). Dit betekent dat de ambtenaar die in direct contact staat met de burger, ook steeds vaker nauwer in contact komt te staan met digitale toepassingen (van Eck et al., 2018). Hoewel het gebruik van AI publieke organisaties een hoop voordelen kan opleveren bij het realiseren van publieke waarde, zoals productiviteit en efficiëntie (Shrum et al., 2019), komen door de focus op efficiëntie andere waarden zoals transparantie, rechtvaardigheid en responsiviteit onder druk te staan (Schiff, Schiff & Pierson, 2021).

Door onder andere de gebrekkige transparantie van algoritmes verdampt de verantwoordelijkheid in de complexe relaties rondom het gebruik van algoritmen (De Fine Licht & De Fine Licht, 2020). Verantwoordelijkheid is bij bureaucratische organisaties inherent ingewikkeld, omdat veel verschillende actoren betrokken zijn bij het oplossen van een probleem. Thompson (1980, p. 905)

noemde dit fenomeen ook wel *'the problem of many hands'*, waarbij er geen overzicht is van welke betrokkene verantwoordelijk is voor welk gedeelte in het proces. Het toepassen van AI maakt deze relaties nog complexer. De gebruiker aanwijzen als verantwoordelijke is te simplistisch, omdat ook andere actoren zoals de ontwerper of de beheerders van het AI-systeem een bepaalde verantwoordelijkheid bekleden (Wortham et al, 2017). Dit is problematisch, want verantwoordelijkheid is een belangrijk onderdeel van goed openbaar bestuur (Hood, 2006; Fox, 2007).

Om organisaties verantwoordelijk te kunnen stellen op macro-niveau, moeten gebruikers van AI-systemen op micro-niveau zich ook verantwoordelijk *voelen* (Overman et al., 2020). Dit is niet alleen belangrijk voor het bewerkstelligen van goed openbaar bestuur, maar uit onderzoek blijkt ook dat werknemers met een hoger verantwoordelijkheidsgevoel betere werkprestaties leveren (Schlenker & Weigold 1989). Wanneer werknemers daarentegen teveel geneigd zijn om anderen of iets anders dan henzelf verantwoordelijkheid toe te schuiven, kan dit resulteren in conflicten. Het ervaren van te veel persoonlijke verantwoordelijkheid voor een taak kan dan weer zorgen voor frustratie en inflexibiliteit (Roberts & Wargo, 1994). Hetgeen dat vooral belangrijk is, is dat verantwoordelijkheid wordt toegeschreven aan de juiste actor, zodat op het moment dat iets fout gaat bekend is waar het fout is gegaan (Thompson, 1980, p. 906). Hierdoor kan dezelfde fout voorkomen worden in de toekomst en kan betere publieke dienstverlening bewerkstelligd worden.

In human-AI interactie, de dynamiek tussen mens en systeem, is het niet altijd makkelijk te bepalen waar de verantwoordelijkheid ligt. Individuele ambtenaren begrijpen soms de werking van het systeem niet, waardoor ze niet in staat zijn om te reflecteren op hun eigen gedrag (Santoni de Sio & Mecacci, 2021). Daarnaast zijn makers en gebruikers van het AI systeem zich niet altijd voldoende bewust van het feit dat het systeem in overeenstemming moet zijn met hun moreel handelen (van Eck et al., 2018; Santoni de Sio & Mecacci, 2021). Hoewel AI-systemen namelijk in staat zijn om objectieve ondersteuning te bieden, ze kunnen de beslissingen van de agent ook beïnvloeden (Zerilli et al., 2019). Zo hebben de algoritmes die de Belastingdienst gebruikte voor controle van foutieve aanvragen met kinderopvangtoeslag en fraude tot etnisch profileren geleid (Amnesty International, 2021). Daarnaast bestaat altijd de mogelijkheid dat het AI-systeem een fout maakt. Zo kan de bevragingsassistent van de Politie bijvoorbeeld een melding plaatsen in de verkeerde categorie of een melding niet detecteren. Het is bij een dergelijke fout dan de vraag waar in de relatie en interactie tussen de gebruiker en het AI-systeem de verantwoordelijkheid ligt.

Sommigen interpreteren dit als een dilemma voor organisaties: of ze gebruiken lerende systemen en geven menselijke controle op, of ze prefereren menselijke controle en gebruiken lerende systemen niet optimaal (Santoni de Sio & Mecacci, 2021). In dit onderzoek wordt dit dilemma niet gezien als een zwart-wit keuze, maar meer als twee uitersten van een continuüm. Het roept de vraag op hoe organisaties van de voordelen van AI-systemen kunnen profiteren terwijl ze tegelijkertijd verschillende actoren zoals managers, AI ontwikkelaars en gebruikers verantwoordelijk kunnen houden (Doshi-Velez et al., 2017, p. 2). In dit onderzoek staat voornamelijk de verantwoordelijkheid van de gebruiker centraal. In hoeverre kent de agent verantwoordelijkheid toe aan het systeem en in hoeverre aan zichzelf? En als het algoritmische systeem transparanter is, kent de gebruiker dan meer verantwoordelijkheid toe aan zichzelf? Om het antwoord op deze vragen te vinden wordt gefocust op *ervaren morele verantwoordelijkheid*: de verantwoordelijkheid die iemand voelt over zijn beslissingen en acties (Oshana, 2004).

1.2 Onderzoeksvraag

Een logische reactie op de verantwoordelijkheidsproblematiek rondom AI betreft het vergroten van de transparantie van algoritmisch bestuur (Meijer & Grimmelikhuijsen, 2020). Transparantie wordt als tool gezien om verantwoordelijkheid te bewerkstelligen (EPRS, 2019). In het onderzoek van Kim & Hinds (2006) wordt ook veronderstelt dat transparantie een sleutelfactor is in het toekennen van succes of falen door gebruikers van AI-systemen. Het succes of falen heeft in die zin betrekking op of iets goed of slecht gaat in een situatie. Zo kennen mensen volgens de attributietheorie succes sneller toe aan zichzelf en falen sneller aan iets buiten henzelf, zoals een systeem (Miller & Ross, 1975; Ployhart & Ryan, 1997)**.** Een verhoogde algoritmische transparantie, zo analyseren Kim & Hinds (2006), is geassocieerd met verminderde toewijzing van succes of falen aan het AI-systeem en meer toewijzing aan de gebruikers zelf of aan andere collega's. Hierop gebaseerd is het de verwachting dat hoe hoger de algoritmische transparantie is, hoe meer mensen zich bewustzijn van hun eigen verantwoordelijkheden, beslissingen en acties. Deze verwachting wordt later vertaald naar een concrete hypothese en deze wordt verder toegelicht in het theoretisch kader.

In deze scriptie ligt de focus specifiek op *algoritmische transparantie.* Algoritmische transparantie kent drie belangrijke aspecten: traceerbaarheid, uitlegbaarheid en communicatie (AI HLEG, 2019). In dit onderzoek wordt vooral gefocust op het aspect 'uitlegbaarheid'. Dit concept wordt verder uitgediept in het theoretisch kader. Om het begrip algoritmische transparantie operationeel te maken is ervoor gekozen om gebruik te maken van visualisaties gecombineerd met een korte uitleg. Visualisaties als vorm van transparantie zouden bijdragen aan de begrijpbaarheid en de bruikbaarheid van het AI-systeem (Park & Gil-Garcia, 2022; Wortham et al. 2017). Dit is interessant, omdat dit een veelbelovende manier is om transparantie te vergroten; echter de effecten van deze vorm van algoritmische transparantie is beperkt onderzocht. Zo bestaan er buiten het onderzoekspaper van Kim & Hinds (2006) weinig onderzoeken die het effect van algoritmische transparantie op ervaren morele verantwoordelijkheid onderzoeken. Zodoende staat de volgende onderzoeksvraag in deze thesis centraal:

Wat is het effect van algoritmische transparantie op de ervaren morele verantwoordelijkheid van gebiedsgebonden Politieagenten?

Om tot een antwoord op de hoofdvraag te komen zijn een aantal theoretische deelvragen geformuleerd.

- *Hoe wordt algoritmische transparantie gedefinieerd?*
- *Hoe wordt ervaren morele verantwoordelijkheid gedefinieerd?*
- *Wat is er in de wetenschappelijke literatuur bekend over de relatie tussen algoritmische transparantie en ervaren morele verantwoordelijkheid?*

Deze deelvragen zijn beantwoord doormiddel van literatuuronderzoek. Vervolgens is de onderzoeksvraag empirisch getoetst doormiddel van een experiment. De kwantitatieve bevindingen zijn tot slot in context geplaatst doormiddel van post-experimentele gesprekken.

1.3 Wetenschappelijke relevantie

De 'High-Level Expert Group on Artificial Intelligence' (AI HLEG), ingesteld door de Europese Commissie, stelt dat transparantie nauw verband houdt met het beginsel van verantwoording (Europese Commissie, 2019, p. 21), maar is dit eigenlijk wel zo? De resultaten van onderzoeken naar het verband tussen transparantie en verantwoordelijkheid lopen uiteen (Cucciniello, Porumbescu, & Grimmelikhuijsen, 2016). Sommige onderzoeken wijzen op een positief verband tussen transparantie en verantwoordelijkheid, terwijl andere onderzoeken helemaal geen verband uitwijzen (Cucciniello, et al., 2016). Ook zijn er onderzoeken die stellen dat de relatie tussen de twee concepten complexer is en dat elke vorm van transparantie alleen maar een gedeelte van verantwoordelijkheid kan bevorderen (Fox, 2007; Hansen & Flyverbom, 2015; Heimstädt, 2017; Whittington and Yakis-Douglas, 2020). Zo beargumenteert Fox (2007, p. 669) dat transparantie alleen overlapt met 'soft accountability', dat betrekking heeft op de uitlegbaarheid van beslissingen.

Naar het effect van transparantie op *ervaren morele verantwoordelijkheid* in een AI-context is nog weinig onderzoek gedaan. Veel onderzoek dat gedaan is over ervaren morele verantwoordelijkheid heeft met name betrekking op de attributietheorie (Miller & Ross, 1975; Ployhart & Ryan, 1997). Het eventuele effect van transparantie wordt in die onderzoeken buiten beschouwing gelaten. Kim & Hinds (2006) doen in hun onderzoek, naast het effect van autonomie, wel specifiek onderzoek naar het effect van transparantie op ervaren morele verantwoordelijkheid. In dit onderzoek gaat het over het toekennen van 'credit' of 'blame' aan zichzelf, de robot of aan anderen. Een beperking in het onderzoek is echter dat het middel dat Kim & Hinds (2006) gebruiken om transparantie te bewerkstelligen niet zorgt voor meer begrijpbaarheid en uitlegbaarheid van het AI-systeem.

Recentelijk onderzoek van Park & Gil-Garcia onderstreept het belang van het gebruik van visualisaties als instrument om begrijpbaarheid te bewerkstelligen. De huidige empirische kennis over het gebruik van visualisatie-tools om de kloof tussen transparantie en verantwoordelijkheid te overbruggen is echter nog beperkt (Park & Gil-Garcia, 2022). Dit is opvallend, want visualisaties als vorm van transparantie zouden bijdragen aan de begrijpbaarheid en daarmee de bruikbaarheid ervan (Park & Gil-Garcia, 2022; Wortham et al. 2017). Zo blijkt uit experimenteel onderzoek vanuit de tak van de robotica (Wortham et al, 2017) dat zowel een video van de werking van de robot, als ook een directe observatie van de werking van de robot, bijdraagt aan de begrijpbaarheid van het systeem door de gebruiker.

Dit onderzoek bouwt voort op kwantitatief onderzoek van Kim & Hinds (2006) en onderzoekt het verband tussen transparantie en ervaren morele verantwoordelijkheid doormiddel van een experiment. Met de aanname dat visualisaties bijdragen aan de begrijpbaarheid van AI-systemen (Park & Gil-Garcia, 2022; Wortham et al. 2017) worden de mogelijkheden van visualisaties als vorm van transparantie verder verkend. Dit wordt gedaan door een tekstuele visualisatie te combineren met een korte uitleg. Daarmee is dit onderzoek een aanvulling op eerder onderzoek.

1.4 Maatschappelijke relevantie

Dit onderzoek is daarnaast ook maatschappelijk relevant. Goede informatie en voldoende inzicht is namelijk essentieel voor effectieve en efficiënte uitvoering van de politietaak (den Hengst, ten Brink & ter Mors, 2017). Er zijn uit de politiepraktijk talloze situaties bekend waarbij het achterwege blijven

van adequate informatie heeft geleid tot grote missers (den Hengst, ten Brink & ter Mors, 2017). Een voorbeeld van zo'n misser is vuurwapengebruik tegen slecht geïnformeerde collega's door het onterecht ontbreken van de gevarenclassificatie 'vuurwapengevaarlijk' op een aan te houden verdachte (den Hengst et al., 2017, p. 26). Andersom zijn er ook voorbeelden uit de praktijk van mensen die onterecht vastzitten aan een gevarenclassificatie (Nationale Ombudsman, 2010). Incomplete informatievoorziening kan dus leiden tot gevaarlijke of ongewenste situaties voor zowel de agent als de verdachte.

Een gevarenclassificatie wordt sinds 2015 in afstemming met een Hulpofficier van Justitie toegekend (Ministerie van Justitie en Veiligheid, 2019). Een gevarenclassificatie is een Politie-interne kwalificatie ten behoeve van het beschermen van politiemedewerkers, collega's of de verdachte zelf (Ministerie van Justitie en Veiligheid, 2019). Het projectteam AI bij de Politie is geïnteresseerd in het ontdekken van de mogelijkheden van artificial intelligence met betrekking tot deze gevarenclassificaties. Zij willen verkennen of het ook mogelijk is om terug te gaan naar de bron, namelijk de registraties. Daarnaast zijn ze geïnteresseerd in het logischer structureren van registraties. Om dit te bewerkstelligen is onderzoek nodig naar het ethisch verantwoord omgaan met de applicatie.

Het projectteam AI heeft zich, om zo verantwoord mogelijk om te gaan met AI, opgesplitst in zes verschillende perspectieven: het technisch perspectief, planmatig perspectief, juridisch/ethisch perspectief, gebruikersperspectief, kosten/baten perspectief en strategisch perspectief (interne bron). Elk perspectief is een stukje van de puzzel. Vanuit de literatuur blijkt dat onderzoek naar transparantie belangrijk is voor o.a. het gebruikersperspectief. Zo blijkt uit onderzoek van Sinha et al. (2002) dat gebruikers van een aanbevelingssysteem liever werken met een transparant dan een niet-transparant systeem. Een transparant systeem zorgt voor meer vertrouwen in het samenwerken met het AI-systeem en gebruikers zijn vaak benieuwd naar de manier waarop het systeem tot een aanbeveling is gekomen. Wanneer een gebruiker niet weet hoe het systeem beslissingen heeft genomen kan dit zorgen voor ambiguïteit over de ervaren verantwoordelijkheid (Heckman et al., 1998). De gebruiker weet niet goed hoe en wanneer het systeem te gebruiken.

Dit onderzoek draagt daarnaast ook bij aan het ethisch perspectief. Vragen waar de Politie mee worstelt vanuit dit perspectief zijn bijvoorbeeld: *'Voldoet de gewenste toepassing van AI aan de eisen voor uitlegbaarheid en transparantie (met name m.b.t. de te gebruiken algoritmen?)* of *'Waar ligt de verantwoordelijkheid voor de handelingen die volgen uit het gebruik van de toepassing?'* (Interne bron, 2022). Tot slot draagt het onderzoek bij aan het technisch perspectief, omdat ingegaan wordt op het middel om transparantie te realiseren en het design van de bevragingsassistent.

Dit onderzoek verkent deze vragen en geeft concrete aanbevelingen over de mogelijkheden van algoritmische transparantie en het vervullen van de verantwoordingsplicht en draagt daarmee bij aan het op een verantwoorde wijze inzetten van AI bij de Politie. Hoewel dit onderzoek specifiek is toegespitst op de Politie kan het daarnaast als voorbeeld dienen voor onderzoek in andere vergelijkbare sectoren waarin gewerkt wordt met street-level-bureaucrats, zoals uitvoerende organisaties van de overheid.

1.5 Leeswijzer

In deze paragraaf wordt kort besproken hoe de volgende hoofdstukken eruit zien. In hoofdstuk 2 worden relevante begrippen verkend en gedefinieerd doormiddel van de literatuur. De paragraven in hoofdstuk 2 vormen samen het theoretisch kader. Het onderzoeksdesign, een beschrijving van de casus, een realisatie van de experimentele vignetten, de operationalisatie, de methoden en de ethische verantwoording staan in hoofdstuk 3 de methode. In hoofdstuk 4 worden de resultaten van het experiment geanalyseerd en geïnterpreteerd. Dan worden op basis van de bevindingen in hoofdstuk 5 een conclusie en discussie besproken en worden aanbevelingen gedaan. Hoofdstuk 6 en 7 bestaan uit de literatuurlijst en de bijlagen.

Hoofdstuk 2: Theoretisch kader

In het theoretisch kader worden de volgende deelvragen beantwoord:

- *Hoe wordt algoritmische transparantie gedefinieerd?*
- *Hoe wordt ervaren morele verantwoordelijkheid gedefinieerd?*
- *Wat is er in de wetenschappelijke literatuur bekend over de relatie tussen algoritmische transparantie en ervaren morele verantwoordelijkheid?*

Allereerst worden verschillende begrippen zoals algoritmische transparantie en ervaren morele verantwoordelijkheid in dit hoofdstuk gedefinieerd. Dan worden twee hypotheses geformuleerd op basis van theorieën over het causaliteitsverband tussen algoritmische transparantie en ervaren morele verantwoordelijkheid. Hoe de begrippen worden toegepast en gemeten binnen dit onderzoek komt later aan bod in het methodehoofdstuk.

2.2 Transparantie

In deze paragraaf wordt aan de hand van bestaande literatuur de volgende deelvraag beantwoord: *Hoe wordt algoritmische transparantie gedefinieerd?* Allereerst wordt een beschrijving gegeven van algoritmische transparantie en hoe het begrip geïnterpreteerd wordt in dit onderzoek. Vervolgens worden de mogelijkheden van algoritmische transparantie in de vorm van visualisaties verder verkend. Dit samen vormt de basis voor een concreet en bruikbaar instrument van algoritmische transparantie dat ingezet kan worden in het experiment van dit onderzoek.

2.1.1 De definitie van algoritmische transparantie

Transparantie wordt door de 'High Level Expert Group on Artificial Intelligence' (AI HLEG) gezien als één van de zeven richtlijnen om ethisch verantwoord om te gaan met AI (AI HLEG, 2019). Het algemene concept van transparantie bij overheidsorganisaties heeft in de meeste definities betrekking op de beschikbaarheid van informatie over eigen besluitvormingsprocessen, procedures, functioneren en prestaties (Hood & Heald, 2006; Curtin and Meijer 2006; Welch and Wong 2001). In dit onderzoek gaat het specifiek over de transparantie van het AI-systeem: de algoritmische transparantie.

Algoritmische transparantie omvat de transparantie van elementen zoals de gegevens, het systeem en bedrijfsmodellen (AI HLEG, 2019). Grimmelikhuijsen (2022) beredeneert dat algoritmische transparantie bereikt is *"when external actors can access the underlying data and code of an algorithm and the outcomes produced by it are explainable in a way a human being can understand."* Deze definitie bevat twee belangrijke elementen: toegankelijkheid en verklaarbaarheid. Deze twee elementen komen ook terug in de beschrijving van transparantie van de AI HLEG (2019, p. 22). Zij duiden drie aspecten van transparantie: 1) traceerbaarheid, 2) uitlegbaarheid en 3) communicatie (AI HLEG, 2019, p. 22). In de volgende alinea's zullen deze drie aspecten van transparantie worden uitgelegd en vervolgens wordt beredeneerd dat het in dit onderzoek vooral relevant is om te focussen op het aspect 'uitlegbaarheid'.

Traceerbaarheid heeft betrekking op de documentatie van gegevenssets en de processen waaruit de beslissing van het AI-systeem voortkomt (AI HLEG, 2019, p. 22). Door goede documentatie is het

mogelijk om eventuele fouten van het AI-systeem te traceren en deze te voorkomen in de toekomst. Traceerbaarheid maakt controleerbaarheid en verklaarbaarheid mogelijk (AI HLEG, 2019, p. 22; Mora-Cantallops et al., 2021). Goede documentatie en het traceren van eventuele fouten is lastig als het algoritme niet *toegankelijk* is (Grimmelikhuijsen, 2022). Een algoritme kan om verschillende redenen niet toegankelijk zijn, bijvoorbeeld omdat het bedrijf dat het algoritme heeft gemaakt commerciële doeleinden heeft en de werking van het systeem binnen het bedrijf wil houden. Dit vormt een risico, omdat het niet mogelijk is om bij een ontoegankelijk algoritme te doorgronden of deze biased, discriminerend of onnauwkeurig is (Grimmelikhuijsen, 2022).

Uitlegbaarheid heeft betrekking op het verklaren van zowel de technische processen van een AI-systeem als de daaraan gerelateerde menselijke beslissingen. Het is belangrijk dat de resultaten gerepresenteerd door het AI-systeem voor de gebruikers ervan te begrijpen zijn. Algoritmische beslissingen kennen een zogenoemde *black box*, wanneer ze niet meer door mensen te begrijpen zijn en daarmee ook niet uit te leggen zijn aan burgers (Burrell, 2016). Een black box kan enigszins acceptabel zijn, als de gevolgen van de applicatie niet groot zijn. Wanneer een AI-systeem echter (significante) gevolgen kan hebben voor het leven van individuen, moet een geschikte verklaring van het besluitvormingsproces van zowel het AI-systeem als de mens mogelijk zijn (AI HLEG, 2019, p. 22). Dit is vooral zo bij organisaties met street-level-bureaucrats, omdat hier algoritmes steeds vaker de discretionaire ruimte van de street-level-bureaucrat beïnvloeden (Young et al. 2019). Street-level-bureaucrats staan in direct contact met de burger en hebben daarmee een directe impact op het leven van die burger, denk bijvoorbeeld aan een agent die iemand staande houdt. Tot slot, draagt de uitlegbaarheid van algoritmische besluitvormingsprocessen positief bij aan het vertrouwen dat burgers hebben in algoritmes en de street-level-bureaucrats die ermee werken (Grimmelikhuijsen, 2022).

Het derde aspect van *communicatie* omvat dat AI-systemen zich tegenover gebruikers niet als mensen mogen voordoen (AI HLEG, 2019). Het moet altijd duidelijk zijn dat het om een AI-systeem gaat. Dit aspect heeft vooral betrekking op chat-bots of dialoog systemen, waarbij een mens praat met het AI-systeem. Holmquist (2017, p. 33) beargumenteert dat goede communicatie tussen het AI-systeem en de gebruiker een van de hoofdfactoren is van het succesvol implementeren van AI. De gebruiker moet begrijpen hoe het systeem de interactie veranderd (Holmquist (2017, p. 32). Het is daarbij belangrijk dat eindgebruikers op de hoogte zijn van zowel de capaciteiten als ook de beperkingen van het AI-systeem (AI HLEG, 2019).

In dit onderzoek richt ik mij vooral op het aspect van ‘uitlegbaarheid’. Om morele verantwoordelijkheid af te kunnen leggen, moeten agenten zowel hun eigen acties als die van het systeem goed kunnen begrijpen en verklaren. Het aspect van ‘uitlegbaarheid’ is daarom het meest relevant. Het aspect ‘communicatie’ is deels relevant. Het gebruiken van een robot als symbool komt met name voor in dialoog-systemen. Dit onderzoek richt zich op de casus van de bevragingsassistent , dat geen dialoog-systeem is. Om deze reden wordt het aspect van communicatie niet meegenomen in dit onderzoek. Voor nader onderzoek is het wel interessant om te kijken of er een samenhang is tussen het begrijpen van de capaciteiten en beperkingen van het AI-systeem en het uitleggen van de werking van het AI-systeem. Elk systeem, ongeacht of het een dialoogsysteem is of niet, kent namelijk beperkingen. Het aspect ‘traceerbaarheid’ is zeker belangrijk, maar is meer van belang voor experts en is voor leken ingewikkeld. De documentatie van gegevenssets en de processen hebben meer betrekking op de verantwoordelijkheid van de ontwerper van het systeem, dan op de

verantwoordelijkheid van de eindgebruiker. In dit onderzoek ligt de focus echter op de eindgebruiker, namelijk de agent. Om deze reden is ervoor gekozen om niet te focussen op het aspect 'traceerbaarheid'. Het aspect 'uitlegbaarheid' staat in dit onderzoek centraal.

2.1.2 De realisatie van algoritmische transparantie

Om algoritmische transparantie te realiseren moeten ontwerpers twee zaken overwegen: 1) wat de getoonde informatie moet zijn en 2) de manier waarop de informatie getoond wordt (Theodorou et al., 2017). Hetgeen *waarover* men transparant is wordt ook wel het *subject* van transparantie genoemd (Grimmelikhuijsen & Hong, 2013, p. 576). De manier *waarop* informatie getoond wordt is het *middel* om die transparantie te realiseren. In de voorgaande paragraaf is beredeneerd waarom vooral gefocust wordt op het aspect 'uitlegbaarheid' van algoritmische transparantie. Om deze reden wordt in het volgende gedeelte niet het subject en middel van transparantie toegelicht, maar het subject en middel van uitlegbaarheid.

Subject

De uitlegbaarheid van algoritmische transparantie kan zich focussen op verschillende aspecten van het AI-systeem. Kim en Routledge (2018) onderscheiden twee soorten uitleg: uitleg over de systeemfunctionaliteit en uitleg over de specifieke beslissing. Nieuwenhuizen (2020, p. 28) noemt dit onderscheid ook wel *de procedurele uitleg* en *de inhoudelijke uitleg*. De procedurele uitleg gaat over de procedure waarop de resultaten tot stand komen. Dit betreft een uitleg over de logica, betekenis, gevolgen en functionaliteit van een geautomatiseerd besluitvormingssysteem (Kim en Routledge, 2018). De inhoudelijke uitleg gaat vaak over een specifieke beslissing of aanbeveling. In de casus van de bevragingsassistent maakt het systeem geen specifieke beslissing of aanbeveling. Het is aan de agent om uiteindelijk te anticiperen op de informatie die de BA geeft. De BA is er vooral voor om de meldingen te categoriseren en dient ter ondersteuning van de agent.

Een algoritme dat informatie categoriseert wordt ook wel een *clusteralgoritme* genoemd. Een clusteralgoritme is een voorbeeld van een zwakke variant van AI (Fjelland, 2020). Zwakke varianten van AI zijn intelligente systemen die één specifieke intellectuele functie hebben. Hoewel het systeem dus geen aanbevelingen doet, maakt het wel gebruik van algoritmes. In dit onderzoek zijn de inhoudelijke en procedurele uitleg niet zo duidelijk te scheiden als in de theorie geschetst wordt. Enerzijds wordt inzicht gegeven in de procedure, doordat de bevragingsassistent de meldingen die horen bij de categorie laat zien en de gevonden woorden dikgedrukt zijn. Anderzijds wordt na de gemaakte keuze een inhoudelijke uitleg gegeven over waarom het systeem deze gegevens laat zien. Dit wordt nader toegelicht in hoofdstuk over de methodiek.

De procedurele uitleg bij een AI-systeem kan betrekking hebben op verschillende aspecten die tot ondoorzichtigheid kunnen leiden. Koene et al. (2019, p. 2) noemen twee concrete aspecten die betrekking hebben op de procedurele uitlegbaarheid: 1) uitleg over het gehele systeem of 2) uitleg over een specifieke uitkomst. Uitleg over het systeem kan te maken hebben met codes, de input van data analyse, statistische analyses van resultaten of de analyse van de gevoeligheid van de input. Dit soort zaken zijn nauwelijks te begrijpen zonder dat een systeemontwerper het kan toelichten en het is ook van weinig waarde voor de eindgebruiker. Het is dan relevanter om transparant te zijn over een bepaalde uitkomst. In de casus van de bevragingsassistent kan bijvoorbeeld transparantie worden gegeven over de manier waarop het systeem informatie gecategoriseerd heeft en toont aan

de gebruiker. Het *subject* van uitlegbaarheid heeft in dit onderzoek dus betrekking op de transparantie van de manier van categoriseren van de bevragingsassistent. Het gaat erom dat het AI-systeem uitleg geeft over de keuze om een melding in een bepaalde categorie te plaatsen, zodat de agent de werking van het AI-systeem kan begrijpen en uitleggen. Transparantie in dit onderzoek sluit aan op de definitie van transparantie van Kim & Hinds (2006) die ook onderzoek doen naar het effect van transparantie op morele verantwoordelijkheid. Zij interpreteren transparantie simpelweg als een systeem dat uitleg geeft over zijn acties (Kim & Hinds, 2006, p. 2).

Middel

Welke informatie relevant is om te laten zien is in elke context verschillend (Theodorou et al., 2017). Een manier om informatie te tonen is doormiddel van visualisaties. De huidige empirische kennis over het gebruik van visualisatie-tools als vorm van algoritmische transparantie is nog beperkt (Park & Gil-Garcia, 2022). Dit is opvallend, want visualisaties als vorm van transparantie zouden bijdragen aan de begrijpbaarheid en daarmee de bruikbaarheid van de informatie van het AI-systeem (Park & Gil-Garcia, 2022; Wortham et al. 2017). Binnen literatuur van de robotica, de tak van wetenschap die zich bezighoudt met het ontwikkelen en bestuderen van robots, speelt de vraag hoe het design van een robot eruit moet zien, zodat het voor mensen begrijpbaar is. Wortham et al. (2017) hebben vanuit dit vraagstuk experimenteel onderzoek gedaan en zij demonstreren dat zelfs een simpele visualisatie van de werking van het AI-systeem van de robot kan bijdragen aan de transparantie en begrijpbaarheid van het AI-systeem. Uit hun onderzoek blijkt dat zowel een video van de werking van de robot, als ook een directe observatie van de werking van de robot, significant bijdraagt aan de begrijpbaarheid en uitlegbaarheid van het systeem door de gebruiker (Wortham et al. 2017).

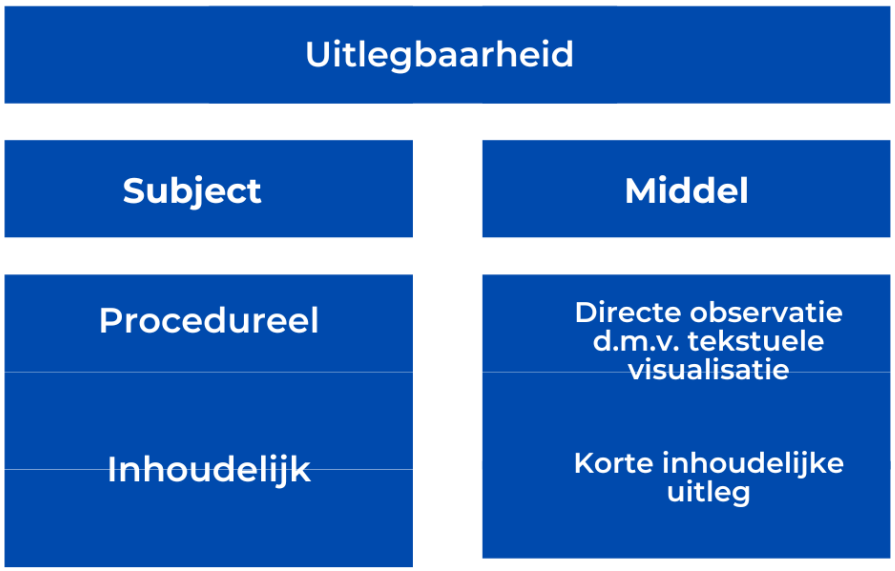
Om het effect van algoritmische transparantie te meten is het belangrijk dat het middel van transparantie ook daadwerkelijk bijdraagt aan de begrijpbaarheid en uitlegbaarheid van het AI-systeem. Het is mogelijk om een video te maken over de werking van het systeem, maar een uitleg over de werking van het gehele systeem blijk niet altijd nodig (Koene et al., 2019). Om deze reden wordt gefocust op de transparantie van de manier van categoriseren van de bevragingsassistent. De algoritmische transparantie wordt gerealiseerd doormiddel van een tekstuele visualisatie. Dit is een vorm van een directe observatie van de werking van het systeem en zou dus moeten bijdragen aan de begrijpbaarheid en uitlegbaarheid (Wortham et al., 2017). Dit houdt in dat het voor de eindgebruiker zichtbaar is waarom het algoritme ervoor heeft gekozen om een melding in een bepaalde categorie te plaatsen. Een gesimplificeerd voorbeeld hiervan is te zien in onderstaande figuur 2.1.

Figuur 2.1

Naam: Robin Jansen
Categorie: Bedreiging met wapen
Meldingen:
02-05-2019 Robin loopt met een **mes** op straat…"
04-06-2020 Robin heeft een psychose en dreigt te steken met een **schroevendraaier**…"

In de figuur is te zien dat het AI-systeem de twee meldingen onder de categorie ‘bedreiging met wapen’ heeft geplaatst, doordat hij het woord ‘mes’ en ‘schroevendraaier’ heeft gedetecteerd. Om een duidelijk overzicht te geven van het gekozen subject en middel van uitlegbaarheid in dit onderzoek is een schematische weergave gemaakt die hieronder te zien is in figuur 2.2.

Figuur 2.2



Samengevat kan algoritmische transparantie gerealiseerd worden doormiddel van uitlegbaarheid. Het *subject* van uitlegbaarheid kan zich uitten in een procedurele uitleg en een inhoudelijke uitleg. Hoewel in theorie de twee concepten van elkaar gescheiden worden, overlappen deze elkaar in de praktijk. De uitleg heeft betrekking op de manier van categoriseren van de bevragingsassistent. De tekstuele visualisatie en tevens directe observatie dient als *middel* om een procedurele uitleg te realiseren. De uitleg die na de keuze wordt gegeven dient meer als een inhoudelijke uitleg.

2.2 Ervaren morele verantwoordelijkheid

In deze paragraaf wordt aan de hand van bestaande literatuur de volgende deelvraag beantwoord: *Hoe wordt ervaren morele verantwoordelijkheid gedefinieerd?* Allereerst wordt het verschil tussen formele en informele verantwoordelijkheid verkend. Vervolgens wordt het verschil tussen *relationele ervaren verantwoordelijkheid* en *ervaren morele verantwoordelijkheid* nader toegelicht en wordt de keuze voor ervaren morele verantwoordelijkheid beargumenteerd. Tot slot wordt ingegaan op de definitie van ervaren morele verantwoordelijkheid.

2.2.1 Formele en informele verantwoordelijkheid

Naast transparantie is ook *verantwoordelijkheid* een vereiste voor het ethisch verantwoord werken met AI (AI HLEG, 2019). Op grond van deze vereiste moet het mogelijk zijn om zowel voor als na de toepassing verantwoordelijkheid voor AI-systemen en de resultaten daarvan te garanderen. Het Kenniscentrum Data & Maatschappij (2020) stelt dat de eis impliceert dat de potentiële risico’s van AI-systemen op een transparante manier moeten worden geïdentificeerd en gelimiteerd. Waar het op neer komt is dat iemand verantwoordelijk moet kunnen worden gesteld indien AI-systemen

schade veroorzaken en dat er in een adequate schadeloosstelling moet kunnen worden voorzien (Data & Maatschappij, 2020).

In de definitie van verantwoordelijkheid kan onderscheid gemaakt worden tussen de *formele* en *informele* variant. De formele verantwoordelijkheid omvat bijvoorbeeld de verkiezingen of prestatiecontroles van organisaties door externe actoren (Peters, 2014). Dit wordt ook wel *verantwoording* genoemd. Echter, zo stellen Overman et al. (2020): *"Formal accountability mechanisms remain utterly symbolic if they do not lead to real accountability processes, in which actors actually explain their behaviour."* Om formele verantwoordelijkheid te kunnen bereiken is *informele* verantwoordelijkheid nodig: de acties en gedragingen van individuen waaruit verantwoordelijkheidsgevoel blijkt (Romzek, LeRoux, and Blackmar 2012, p. 443). Verantwoordelijkheid op meso-niveau; het niveau van de organisatie, vereist verantwoordelijkheidsgevoel op micro-niveau; het niveau van het individu (Coleman 1990, Grimmelikhuijsen et al., 2017). Om de Politie als organisatie verantwoordelijk te kunnen houden moeten agenten zich dus ook verantwoordelijk voelen. Een manier waarop informele verantwoordelijkheid in kaart kan worden gebracht is door te kijken naar *ervaren morele verantwoordelijkheid.*

Hellström (2012) beschouwt morele verantwoordelijkheid als een kwaliteit die we toekennen aan anderen en mogelijk ook aan geautomatiseerde systemen. Deze visie vervaagt onder andere de grenzen tussen morele verantwoordelijkheid en formele verantwoordelijkheid. Stahl (2004, p. 105) beargumenteert dat deze aspecten elkaar overlappen. Zo wordt formele verantwoordelijkheid beïnvloed door morele opvattingen uit de samenleving. Hoewel de focus in dit onderzoek ligt op de ervaren morele verantwoordelijkheid, is het belangrijk om te benadrukken dat de discussie ook andere aspecten van verantwoordelijkheid kan raken. In de volgende paragraaf wordt ingegaan op het verschil met ervaren relationele verantwoordelijkheid en wordt dieper ingegaan op het concept ervaren morele verantwoordelijkheid.

2.2.2 Ervaren relationele verantwoordelijkheid

Een veel gebruikte definitie van verantwoordelijkheid in de bestuurs- en organisatiewetenschap is die van Bovens (2007). Bovens (2007) stelt dat verantwoordelijkheid gekenmerkt is door een actor en een forum, waarbij de actor degene is die verduidelijking moet geven over zijn gedrag en het forum degene is die vragen kan stellen, een oordeel kan vellen en eventuele vervolgstappen kan ondernemen via het gebruik van sancties. Dit sluit aan op de beredenering van Fox (2007) die stelt dat begrippen als 'transparantie' en 'verantwoordelijkheid' inherent relationeel zijn: wat is transparant voor wie en wie legt verantwoordelijkheid af aan wie? Het relationele aspect van verantwoordelijkheid komt terug in het meetinstrument van Overman et al. (2020), die een schaal hebben ontwikkeld voor de *ervaren relationele verantwoordelijkheid.* Dit meetinstrument is gefocust op legitimiteit en expertise en lijkt vooral betrekking te hebben op de relatie tussen de actor en het forum, zoals beschreven door Bovens (2007). Een ander veelgebruikt meetinstrument voor ervaren verantwoordelijkheid is van Hochwarter et al. (2007). Deze schaal lijkt dan weer meer betrekking te hebben op de relatie tussen het management en de werknemers of werknemers en werkgevers.

In dit onderzoek gaat de ervaren verantwoordelijkheid niet zozeer over de relatie tussen actor en forum, of manager en operationeel werknemer, maar meer over de relatie en interactie tussen mens

en het AI-systeem. In hoeverre kent de gebruiker verantwoordelijkheid toe aan het systeem en in hoeverre aan zichzelf? En als de werking van het systeem transparanter is, kent de gebruiker dan meer verantwoordelijkheid toe aan zichzelf? Om het antwoord op deze vragen te vinden wordt gefocust op *ervaren morele verantwoordelijkheid*: de verantwoordelijkheid die iemand voelt over zijn beslissingen en acties (Oshana, 2004).

2.2.3 Ervaren morele verantwoordelijkheid: het toekennen van succes of falen

Het begrip 'morele verantwoordelijkheid' kent zijn oorsprong in de geschreven werken van de filosoof Aristoteles (384-323 v.C.). Aristoteles beschreef in zijn werk Nicomachean Ethics III.1–5 dat *"an agent is described as morally responsible for an action, if it is worthy of praise or blame for having performed the action"* (Aristoteles, 1985, uit Hellström, 2012, p. 6). Het werk van Aristoteles heeft door de eeuwen heen discussies op gang gebracht over intentie, de vrije wil, bewustzijnsvermogen of determinisme (zie voor een overzicht Eshleman, 2009). Volgens Aristoteles moet de actor voldoen aan twee voorwaarden om moreel verantwoordelijk te kunnen zijn: 1) De actor moet de capaciteit hebben om een beslissing te maken en 2) de actie moet vrijwillig zijn. Dit roept de vraag op of een AI-systeem überhaupt wel moreel verantwoordelijk kan worden gehouden. Over de stelling of geautomatiseerde systemen een morele verantwoordelijkheid kunnen hebben is dan ook discussie binnen de filosofie (Friedman, 1990; Johnson 2006, Hong et al., 2020). Hoewel dit niet de vraag is die centraal staat in dit onderzoek, raakt deze discussie het onderzoek wel, omdat onder andere gekeken wordt naar de toegeschreven verantwoordelijkheid aan de bevragingsassistent.

Friedman (1990) stelt dat geautomatiseerde systemen geen morele verantwoordelijkheid kunnen bekleden, omdat een systeem geen intenties kent. Johnson (2006) haalt ook dit standpunt aan en beargumenteert dat bepaalde mentale staten zoals intentie altijd ontbreken in machines, waardoor een robot nooit moreel verantwoordelijk kan zijn. Echter, uit een experimentele studie van Friedman et al. (1995) blijkt dat mensen wel daadwerkelijk verantwoordelijkheid *toeschrijven* aan systemen. Uit dit onderzoek blijkt dat 79% van de deelnemers beoordeelt dat systemen besluitvormingscapaciteiten hebben en 45% van de deelnemers dat systemen ook intenties hebben. Het is een interessante bevinding dat mensen morele verantwoordelijkheid kunnen toeschrijven aan een systeem.

Experimenteel onderzoek van Hong et al. (2020) wijst uit dat mensen zelfs geneigd zijn om meer schuld toe te kennen aan AI-systemen dan aan andere mensen. De reden hiervoor is dat mensen geneigd zijn te denken dat systemen meer accuraat zijn dan mensen (Sundar & Kim, 2019), een fout van het systeem schaadt hun vertrouwen en zorgt daarmee voor een grotere 'blame-factor' aan systemen dan aan mensen (Hong et al., 2020, p. 1771). Andere onderzoekers stellen dat morele verantwoordelijkheid geen vast gegeven is maar een continuüm tussen geen verantwoordelijkheid en tot volle verantwoordelijkheid (Asaro, 2006). Dit blijkt bijvoorbeeld uit het gegeven dat we kinderen in sommige gevallen minder verantwoordelijk stellen voor hun acties dan volwassenen. Op basis van dit inzicht zou je kunnen zeggen dat systemen tot op een zekere hoogte verantwoordelijk kunnen worden gehouden, maar misschien niet in dezelfde mate als dat we verantwoordelijkheid kunnen toeschrijven aan mensen. Het is dan de vraag waar precies de grens ligt.

Het toekennen van succes of falen in human-AI interactie is dus nauw gerelateerd aan het meten van ervaren morele verantwoordelijkheid. Ervaren morele verantwoordelijkheid wordt in dit onderzoek geïnterpreteerd zoals in het onderzoek van Eshleman (2009, p. 1) die stelt dat: *to be morally responsible for something, say an action, is to be worthy of a particular kind of reaction – praise, blame, or something akin to these – for having performed it.* Dit wordt verder uitgewerkt in de operationalisatie in het methodehoofdstuk. Eerst worden in de volgende paragraaf op basis van de theorie hypotheses opgesteld over het causaliteitsverband tussen algoritmische transparantie en ervaren morele verantwoordelijkheid in de praktijk.

2.3 De relatie tussen algoritmische transparantie en ervaren morele verantwoordelijkheid

In deze paragraaf wordt de laatste theoretische deelvraag onderzocht: *Wat is er in de wetenschappelijke literatuur bekend over de relatie tussen algoritmische transparantie en ervaren morele verantwoordelijkheid?* Om deze vraag te beantwoorden wordt eerst de theorie van self-serving bias toegelicht. De oorzaken van self-serving bias vormen de basis voor de hypotheses over het causaliteitsverband tussen algoritmische transparantie en het toekennen van schuld of succes aan systemen of aan mensen zelf. Op basis van de theorie worden twee hypothesen geformuleerd die getoetst worden in de praktijk.

2.3.1 De theorie van 'self-serving bias'

Zoals in het vorige hoofdstuk al kort is aangehaald blijkt uit experimenteel onderzoek van Hong et al. (2020) dat mensen in het geval van een gemaakte fout geneigd zijn om meer schuld toe te kennen aan AI-systemen dan aan andere mensen. De reden hiervoor is dat mensen geneigd zijn te denken dat systemen meer accuraat zijn dan mensen (Sundar & Kim, 2019). Een fout van het systeem schaadt hun vertrouwen en zorgt daarmee voor een grotere 'blame-factor' aan systemen dan aan mensen (Hong et al., 2020, p. 1771). Deze bevinding wordt ondersteund door de attributietheorie die wijst op 'self-serving bias' (Miller & Ross, 1975; Ployhart & Ryan, 1997). Self-serving bias houdt in dat mensen gewenste uitkomsten vaak toeschrijven aan interne, stabiele en controleerbare factoren, terwijl ongewenste uitkomsten eerder worden toegeschreven aan externe, onstabiele en oncontroleerbare factoren (Shepperd et al., 2008).

Onderzoeken over self-serving bias zijn gedaan binnen allerlei verschillende contexten. Zo blijkt uit onderzoek van Michele et al. (1998) dat atleten eerder verantwoordelijkheid op zich nemen over goede sportprestaties dan over slechte sportprestaties. Hetzelfde geld voor autobestuurders, die een ongeluk toeschrijven aan externe factoren zoals het weer of de conditie van de auto, terwijl ze het voorkomen van een ongeluk toeschrijven aan interne factoren zoals hun eigen alertheid of rijvaardigheden (Stewart, 2005). Mensen hebben bijna nooit invloed op externe factoren, wat impliceert dat zij zelf 'niks aan de mislukking konden doen' en dus geen verantwoordelijkheid ervaren voor de fout. Weiner (1986) beredeneert dan ook dat wanneer mensen bij een mislukking refereren naar interne factoren, dus oorzaken binnen in zichzelf zoals persoonlijkheid, keuzes of gedrag, mensen ook meer gemotiveerd zijn voor het voorkomen van fouten in de toekomst, omdat ze denken dat de oorzaken beïnvloedbaar zijn. Tegelijkertijd wil je ook niet dat fouten *onterecht* worden toegekend aan interne factoren, terwijl het duidelijk komt door iets externs. Wanneer een auto bijvoorbeeld slipt door gladheid op de weg, wil je niet dat de automobilist verwijst naar zijn

eigen rijgedrag, maar naar de regen, omdat dit ook de daadwerkelijke oorzaak is van het slippen van de auto.

Gebaseerd op eerdere onderzoeken en de theorieën van self-serving bias is de verwachting dat agenten die het juiste antwoord geven in de survey dit meer toekennen aan zichzelf en agenten die het onjuiste antwoord geven dit toeschrijven aan de bevragingsassistent. Echter, omdat niet gedetermineerd kan worden of agenten het juiste of het onjuiste antwoord zullen geven, is ervoor gekozen om dit geen hypothese te maken. Het kan daardoor namelijk niet worden getoetst. Daarnaast wordt in dit onderzoek het effect van transparantie onderzocht en niet de attributietheorie. Wel worden de oorzaken van self-serving bias en het mogelijke effect van algoritmische transparantie hierop gebruikt als basis voor het vormen van de hypotheses.

2.3.2 De oorzaken van self-serving bias

Shepperd et al. (2008, p. 895-896) onderzoeken de verschillende oorzaken voor self-serving bias en beredeneren dat het toeschrijven van falen of succes aan interne of externe factoren onder andere gerelateerd is aan de mate van controleerbaarheid of de mate waarin je iets kunt voorzien. Als mensen een ongewenst resultaat kunnen controleren of tenminste kunnen zien aankomen, dan kunnen ze acties en maatregelen ondernemen om de fout te voorkomen. Op basis van deze redenering is mijn verwachting dat algoritmische transparantie bijdraagt aan de voorzienbaarheid van de agent, waardoor de agent bij een transparanter systeem dus meer verantwoordelijkheid toeschrijft aan zichzelf. In de casus van de bevragingsassistent heeft het systeem dan tenminste aangegeven op basis waarvan een melding is geplaatst onder een bepaalde categorie. De agent kan de keuze van de bevragingsassistent vervolgens zelf interpreteren en hier zelf op anticiperen, waardoor de verwachting is dat de agent in mindere mate morele verantwoordelijkheid toeschrijft aan het systeem.

Een andere oorzaak voor self-serving bias heeft te maken met de verwachtingen die mensen hebben van uitkomsten (Shepperd et al., 2008, p. 899-900). Mensen zijn geneigd om een bepaalde uitkomst te verwachten op basis van eerdere ervaringen, plannen en intenties (Shepperd et al., 2008). De uiteindelijke resultaten bevestigen of ontkrachten de verwachting. Wanneer een uitkomst in lijn ligt met de verwachting, zoeken mensen er geen verdere reden achter. Een voorbeeld hiervan is een student die geleerd heeft voor een bepaald cijfer en dit cijfer ook behaald. Pas wanneer de uitkomst niet in lijn ligt met een verwachting (een student verwacht een 8 en ontvangt een 6), zal die gaan zoeken naar een oorzaak voor de ‘fout’. Aangezien eigen bekwaamheid en inzet een reden waren voor een positieve verwachting, zijn mensen niet geneigd om het gebrek of de fout aan diezelfde bekwaamheid toe te schrijven. Ze gaan de fout buiten zichzelf zoeken. Algoritmische transparantie zorgt ervoor dat mensen meer toegang hebben tot de ‘gedachtegang’ van het AI-systeem, wat leidt tot meer realistische verwachtingen. Het is de verwachting dat mensen dan ‘fouten’ minder vaak toeschrijven aan het AI-systeem, omdat hun verwachting in lijn ligt met de uitkomst. In de scenario’s over de bevragingsassistent is dan ook de verwachting dat mensen bij een transparant scenario vaker uitkomen op een *juist* antwoord dan een *onjuist* antwoord en lager scoren op stellingen die gaan over de bevragingsassistent dan respondenten die een scenario krijgen met een lage mate van transparantie.

Echter, niet uit elk onderzoek blijkt een positief effect van transparantie op de morele ervaren verantwoordelijkheid. Uit het onderzoek van Kim & Hinds (2006) blijkt dat meer algoritmische transparantie *niet* zorgt voor *minder* toekenning van succes of falen aan het robot systeem of aan zichzelf. Echter, wordt in de discussie een belangrijke beperking van het onderzoek besproken (Kim & Hinds, 2006, p. 5-6). Het *middel* dat zij gebruikt hebben om transparantie te realiseren zorgt significant voor *minder* begrijpbaarheid in plaats van *meer* begrijpbaarheid. Hoe transparanter het systeem, hoe minder deelnemers het systeem dus begrepen. Het middel dat zij gebruikten voor transparantie maakte onderscheid in twee groepen: een groep met een laag level van transparantie en een groep met een hoog level van transparantie. Bij beide groepen maakte de robot een onverwachte beweging. De robot begon opeens rondjes te draaien. De groep die behoorde tot de hoge mate van transparantie, kreeg een uitleg over de onverwachte beweging van de robot: 'I have recalibrated my sensors' en de groep behorend tot de lage mate van transparantie kreeg geen uitleg over de onverwachte beweging van de robot.

De auteurs beredeneren daarom dat het effect van transparantie in hoge mate afhangt van het gebruikte middel om algoritmische transparantie en uitlegbaarheid te realiseren (Kim & Hinds, 2006). Het middel om transparantie te realiseren komt uitgebreid aan bod in het methodehoofdstuk, maar in het kort komt het neer op een nabootsing van het experiment van Kim & Hinds (2006), alleen dan is het middel een korte procedurele en inhoudelijke uitleg gecombineerd met een tekstuele visualisatie, omdat blijkt dat visualisaties significant zorgen voor juist meer begrijpbaarheid en uitlegbaarheid (Wortham et al. 2017). Het systeem legt bij de transparante scenario's uit waarom een gegeven antwoord juist of onjuist is en de respondent kan de fout ook inzien door de tekstuele visualisatie in het scherm. De respondent kan namelijk de meldingen die behoren bij de weergegeven categorieën inzien, terwijl de niet-transparante scenario's geen uitleg krijgen en ook geen meldingen kunnen zien.

Het middel om transparantie te realiseren wordt getoetst door een vraag over begrijpbaarheid als manipulatiecheck te gebruiken. Met de aanname dat het middel om transparantie te realiseren in dit onderzoek zorgt voor meer begrijpbaarheid en uitlegbaarheid en op basis van de theorieën over de oorzaken van self-serving bias zijn de volgende hypotheses opgesteld:

H1: Een hoge mate van algoritmische transparantie zorgt ervoor dat succes of falen meer toegekend wordt aan de agent zelf dan aan de bevragingsassistent

H2: Een lage mate van algoritmische transparantie zorgt ervoor dat succes of falen meer toegekend wordt aan de bevragingsassistent dan aan de agent zelf

Als deze hypotheses bevestigd kunnen worden, zou geconcludeerd kunnen worden dat bij een hoge mate van transparantie de gebruiker meer morele verantwoordelijkheid ervaart dan bij een lage mate van transparantie.

Hoofdstuk 3: Methode

3.1 Onderzoeksmethoden

In deze paragraaf wordt allereerst ingegaan op het onderzoeksdesign. In dit onderzoek is gebruikgemaakt van survey experiment. Het experiment bestaat uit twee geschakelde experimenten. Het eerste experiment gaat over scenario 1 en het tweede experiment gaat over scenario 2. Daarnaast is gebruik gemaakt van post experimentele gesprekken. De onderzoeksvraag wordt hiermee empirisch getoetst. Ten tweede wordt de casus van de bevragingsassistent verder toegelicht. Tot slot wordt uitgelegd hoe de casus is gebruikt in de getoetste vignetten.

3.1.1 Surveyexperiment als onderzoeksdesign

De vraag die in dit onderzoek centraal staat onderzoekt een mogelijk causaal verband tussen algoritmische transparantie en ervaren morele verantwoordelijkheid. De onderzoeksvraag luidt:

Wat is het effect van algoritmische transparantie op de ervaren morele verantwoordelijkheid van gebiedsgebonden Politieagenten?

De onafhankelijke variabele in deze vraag is de *algoritmische transparantie* en de afhankelijke variabele is de *ervaren morele verantwoordelijkheid.* Het vaststellen van causaliteit is ingewikkeld, omdat vaak meerdere variabelen van invloed zijn. Toch wordt een experiment als een zeer geschikte onderzoeksmethode gezien om een causale relatie vast te stellen (Bryman, 2016, p. 44). Bij het doen van een experiment is het belangrijk dat de omstandigheden hetzelfde blijven, behalve hetgeen dat onderzocht wordt.

Het onderzoeksdesign dat in dit onderzoek wordt gehanteerd is een *survey experiment*. Hierin wordt onderscheid gemaakt tussen een control group: de scenario's met een lage mate van transparantie en de experimental group: de scenario's met een hoge mate van transparantie. De antwoorden die de twee groepen geven op de stellingen worden met elkaar vergeleken. Dit experiment wordt twee keer uitgevoerd. Respondenten krijgen namelijk allemaal twee scenario's te zien: één scenario waarin het AI-systeem werkt en één scenario waarin het AI-systeem een fout maakt. In tabel 3.1 is het experiment schematisch weergegeven.

Tabel 3.1 Schematisch overzicht scenario 1

	Scenario 1: AI-systeem werkt	**Scenario 2: AI-systeem maakt een fout**
Lage mate van transparantie (controlgroup)	Scenario waarin AI-systeem werkt en lage mate transparantie	Scenario AI-systeem maakt een fout en lage mate transparantie
Hoge mate van transparant (experimental group)	Scenario waarin AI-systeem werkt en hoge mate transparantie	Scenario AI-systeem maakt een fout en hoge mate transparantie

Door gebruik te maken van randomisatie in het programma Qualtrics is ervoor gezorgd dat respondenten willekeurig scenario 1 en scenario 2 krijgen te zien. Dit betekent dat een respondent zowel scenario 1 als scenario 2 krijgt te zien en een hoge *en/of* lage mate van transparantie. Voordat de respondent het verdere verloop van een scenario weet wordt gebruik gemaakt van een voormeting met de vraag: *'In hoeverre is je antwoord gebaseerd op de bevragingsassistent of op je eigen inschatting?'*(waarbij 1= eigen inschatting en 100= bevragingsassistent). Deze vraag is met name interessant om de attributietheorie verder te exploreren. Het geeft een indicatie van hoe consistent agenten zijn in het geven van hun antwoorden. Als ze bijvoorbeeld eerst zeggen dat hun antwoord 100 procent gebaseerd is op de bevragingsassistent en vervolgens geen succes of schuld toeschuiven aan de bevragingsassistent dan is dit niet consistent. Na het lezen van de scenario's volgt een vragenlijst over de ervaren morele verantwoordelijkheid en de uitlegbaarheid. Aangezien de manipulatiecheck het experiment niet mag beïnvloeden wordt de vraag over uitlegbaarheid pas gevraagd na de vragen over de ervaren morele verantwoordelijkheid. De manipulatiecheck luidt: *'Het is mij duidelijk hoe de bevragingsassistent tot de weergegeven categorieën is gekomen'* (helemaal eens- helemaal oneens op een schaal van 1-7).

Dit experiment wordt ook wel een surveyexperiment genoemd. Bij een surveyexperiment wordt een survey uitgezet met verschillende beschreven scenario's (ook wel vignetten) en een vragenlijst. De scenario's bestaan uit een korte omschrijving van een hypothetische situatie. Een surveyexperiment kent voor- en nadelen. Het is bijvoorbeeld mogelijk om grote groepen respondenten te bereiken. In tegenstelling tot een labexperiment is een surveyexperiment goedkoop en makkelijk te implementeren (Blom-Hansen, Morton & Serritzlew, 2015, p. 161). Echter kent een surveyexperiment ook nadelen. Zo is de manipulatie of interventie van een surveyexperiment vaak minder sterk vergeleken met bijvoorbeeld een labexperiment (Blom-Hansen et al., 2015, p. 161). Dit heeft als gevolg dat je vaak een relatief hoog aantal respondenten nodig hebt om een effect te vinden.

In dit experiment is uitgegaan van een medium sized effect (d=0.5) van transparantie. Hoewel in eerdere vergelijkbare studies (zie Nieuwenhuizen, 2020) een effectsize van 0.15 is gebruikt, wordt het effect in dit experiment iets hoger verwacht. In het experiment van Nieuwenhuizen (2020) is namelijk enkel een uitleg gebruikt als middel van transparantie, maar in dit experiment wordt een uitleg gecombineerd met een tekstuele visualisatie. Uitgaande van een medium effectsize (d=0.5) en een power van 0.8 en een 'between two independent means' toets, betekent dit dat er een steekproef van minimaal 102 participanten nodig is om een effect aan te tonen. Dit is berekend met het programma *Gpower 3.1*. Het totaal aantal respondenten in dit experiment was *N*= 153. Dit draagt bij aan de statistische conclusievaliditeit en dat betekent dat het experiment zo is opgezet dat een kwantitatieve conclusie mogelijk is, omdat het aantal benodigde respondenten behaald is.

Een ander nadeel van een surveyexperiment is dat respondenten minder gefocust zijn op het doen van het experiment dan wanneer zij zich bijvoorbeeld in een laboratoriumomgeving bevinden (Blom-Hansen et al., 2015, p 161). Het kan zijn dat respondenten daardoor afhaken midden in het experiment of het experiment tegelijkertijd doen met een andere activiteit. Dit nadeel kan worden opgevangen doormiddel van een *attention check* (Abbey & Meloy, 2017). In dit onderzoek is ook een attention check toegevoegd aan de vragenlijst door de stelling: "Let op: vul hier in 'helemaal eens'" (schaal: 1-7) of "Let op: vul hier in 'helemaal oneens'" (schaal: 1-7). Echter, zijn deze attention checks niet gebruikt, omdat uit de gesprekken bleek dat sommige respondenten dachten dat het een test

was of ze het systeem zouden opvolgen. Om toch te weten te komen of de respondenten hun aandacht bij het experiment hadden is ook een timer toegevoegd. Deze timer heeft betrekking op hoelang respondenten kijken naar het scherm van de bevragingsassistent. In scenario 1 kijken respondenten gemiddeld *M*=82,28 seconden naar het scherm en in scenario 2 kijken respondenten gemiddeld *M*=74,14 seconden naar het scherm. In scenario 1 zijn er twee opvallende afwijkingen: een respondent die er 6,49 seconden over heeft gedaan en een respondent die er 1616,82 seconden over heeft gedaan. De respondent die er 6,49 seconden over heeft gedaan heeft wellicht het scenario niet goed bekeken en de respondent die er 1616,82 seconden over heeft gedaan is naar alle waarschijnlijkheid iets anders gaan doen tussendoor. Het weglaten van deze twee respondenten levert geen verschuivingen op in de resultaten. In scenario 2 zijn er geen opvallende afwijkingen gevonden. In de volgende paragraaf wordt ingegaan op de casus van de bevragingsassistent. De vignetten zijn gebaseerd op deze casus.

3. 1.2 Casus: de bevragingsassistent

Op dit moment is de Politie bezig met experiment K.I.D.: Kantoor In Dienstvoertuig. K.I.D. is een digitale omgeving in de auto waarop agenten alle relevante informatie op één scherm kunnen zien. Een van de doelen van experiment K.I.D. is het mogelijk maken van de implementatie van AI-toepassingen. Een AI-toepassing die in de toekomst mogelijk wordt geïnstalleerd op K.I.D. is de bevragingsassistent. De Politie is aan het exploreren hoe de bevragingsassistent een rol kan krijgen in de dagelijkse werkzaamheden van gebiedsgebonden agenten en wat die rol dan moet zijn. Maar wat zijn de huidige ideeën over hetgeen wat de bevragingsassistent behoort te gaan doen? En wat zijn praktische problemen met het huidige BVI-IB (Basisvoorziening Informatie-Integrale Bevraging) systeem?

De bevragingsassistent is een toekomstig systeem dat registraties kan categoriseren en kan laten zien op basis van relevantie. Dit doet die door het gebruiken van zoeksystemen en clusteralgoritmen. Dit is iets nieuws, want op dit moment laat het BVI-IB systeem, een systeem waarin personen kunnen worden opgezocht, registraties zien op volgorde van tijd. Zo staat een registratie uit 2019 onder een registratie uit 2022. Voorbeelden van registraties kunnen gaan over verschillende soorten zaken, zoals alcoholgebruik, overlast, verzet of wapenbezit. Wanneer een persoon weinig of geen meldingen op diens naam heeft staan hoeft dit geen probleem te vormen. Echter, wanneer een persoon meerdere meldingen kent, kan het zijn dat een relevante melding soms niet opvalt, terwijl deze wel belangrijke informatie kan bevatten. Het BVI-IB systeem wordt daardoor niet optimaal gebruikt.

Goede informatie en voldoende inzicht is essentieel voor effectieve en efficiënte uitvoering van de Politietaak (den Hengst, ten Brink & ter Mors, 2017). Er zijn uit de Politiepraktijk talloze situaties bekend waarbij het achterwege blijven van adequate informatie heeft geleid tot grote missers (den Hengst, ten Brink & ter Mors, 2017). Een voorbeeld van zo'n misser is vuurwapengebruik tegen slecht geïnformeerde collega's door het onterecht ontbreken van de gevarenclassificatie 'vuurwapengevaarlijk' op een aan te houden verdachte (den Hengst et al., 2017, p. 26). Incomplete informatievoorziening kan dus leiden tot gevaarlijke situaties voor zowel de Politie als de verdachte.

De bevragingsassistent zou hierin verandering moeten brengen. De bevragingsassistent voegt meldingen toe aan een gevaarclassificatie die bestaat uit de volgende categorieën: alcoholist,

harddruggebruik, medische indicatie, vuurwapengevaarlijk, wapengevaarlijk, geweld/verzetpleger, vluchtgevaarlijk en zelfmoordneiging (Ministerie van Justitie en Veiligheid, 2019, p. 17). Vervolgens is het de bedoeling dat de bevragingsassistent de categorie die agenten doorgaans als meest relevant ervaren als eerst zien. Zo staat bijvoorbeeld 'vuurwapengevaarlijk' of 'wapengevaarlijk' bovenaan, omdat de verwachting is dat agenten het belangrijk achten om hierop te anticiperen (persoonlijke communicatie, 2022). Op het moment dat een persoon geen meldingen heeft, dan staat er 'niet relevant'.

Een risico van het gebruik van de BA kan zijn dat het een melding verkeerd categoriseert. Zo kan het voorkomen dat het systeem het woord 'schroevendraaier' herkent als wapen en daarmee de persoon categoriseert als wapengevaarlijk. Op basis van deze informatie kan de Politie overreageren met gevolgen voor de verdachte, vooral wanneer de verdachte helemaal niet wapengevaarlijk blijkt te zijn. Andersom kan ook voorkomen: het AI-systeem kan een bepaald woord ook *niet* als wapen herkennen, terwijl het wel een wapen is. Dit kan de agent in een gevaarlijke situatie brengen.

Wanneer het systeem transparant is over de gemaakte categorieën, kan de agent beter anticiperen op de informatie van het systeem. Zo kan het systeem bijvoorbeeld de melding beschrijven als: *'Man doet melding van agressieve vrouw met **schroevendraaier**'.* Door te laten zien dat het systeem het woord 'schroevendraaier' had geplaatst in de categorie 'wapengevaarlijk', kan de agent de keuze van het systeem beter begrijpen en hierop anticiperen. Zo kan de agent door het lezen van de context besluiten om niet de uiterste middelen in te zetten, omdat te lezen is in de context dat de man de melder is. Wanneer de agent alsnog verkeerd anticipeert op de melding, kan die beter de werking van het systeem als ook zijn eigen acties uitleggen en de eventuele schade beperken.

3.1.3 Realisatie van experimentele vignetten

In deze paragraaf wordt besproken hoe de experimentele vignetten tot stand zijn gekomen. Zo wordt allereerst besproken met welke aspecten rekening is gehouden. Ten tweede, komt aan bod hoe deze aspecten zijn gewaarborgd en bij welke verschillende actoren input is verzameld. Tot slot, worden de keuzes die gemaakt zijn voor de eindversie van de vignetten beargumenteerd.

Bij het toetsen van experimentele vignetten zijn twee aspecten belangrijk: 1) de respondent moet zich kunnen inleven in de voorgelegde scenario's en 2) de scenario's moeten realistisch zijn en overeenkomen met de werkelijkheid (Bryman, 2016, p. 260). Deze twee aspecten zijn vooral belangrijk, omdat het gaat om een hypothetische applicatie die de Politie mogelijk wil inzetten in de toekomst. De applicatie is nog niet ontwikkeld en moet dus lijken op iets dat mogelijk is binnen de Politie. Hoe realistischer de scenario's, hoe relevanter de resultaten voor de praktijk. In het begin waren scenario's ontwikkeld waarin de keuze van de agent was gedetermineerd. De vervolgactie van de agent was bij dit scenario al bepaald. Na feedback van mijn scriptiebegeleider heb ik ervoor gekozen om de keuze meer bij de agent te leggen. In de scenario's die volgden lag het aan de keuze van de agent of die uit zal komen op een 'juist' of 'onjuist' antwoord. Na deze beslissing te hebben genomen, is bij verschillende actoren input verzameld.

Om te beginnen zijn een aantal casussen besproken met een operationeel expert. Uit gesprekken met de operationeel expert bleek dat het lastig was om (binnen de tijd van het onderzoek) realistische scenario's te schrijven die goed pasten bij de strekking van de onderzoeksvraag. Feedback luidde bijvoorbeeld: *"Het lijkt erop dat scenario's zich vooral richten op geweldsgebruik en*

bevoegdheden van dienders. Daar worden zij al regelmatig op getoetst en is daarnaast vrij ingewikkeld met betrekking tot wetteksten en bevoegdheden." Daarbij kwam ook dat niet elke agent op dezelfde manier reageert op een situatie, wat kon afdoen aan het inlevingsvermogen. Op basis van deze en aanvullende opmerkingen van de operationeel expert en met inachtneming van de reikwijdte van dit onderzoek is daarom gekozen om voor dit onderzoek geen uitgebreide situaties te schrijven. Wel is enige vorm van context gegeven door bijvoorbeeld te beschrijven: *'je krijgt een melding van een incident waarbij Robin Jansen betrokken is'.* De precieze inhoud van die melding en de situatie is dus weggelaten.

Verder zijn de scenario's ontwikkeld met een medewerker van de afdeling Team Wetenschappelijke Ontwikkeling (TWO). De input van deze medewerker had vooral betrekking op het design en de werking van de bevragingsassistent. Zo is uitgelegd dat de bevragingsassistent gebruik maakt van zoekalgoritmes om gegevens te categoriseren. Ook is verteld dat het systeem op twee manieren een fout zou kunnen maken: 1) het systeem kan een woord niet vinden en laat daarom een categorie niet zien, terwijl dit wel moest *of* 2) het systeem kan wel een woord vinden en laat een categorie zien, maar het woord hoort niet bij de betreffende categorie. In de vignetten is gekozen voor de tweede beschreven fout. Hoewel de eerste fout ook interessant is om te onderzoeken voor de praktijk, bleek deze minder interessant voor het beantwoorden van de wetenschappelijke onderzoeksvraag. Om deze reden is gekozen om dit scenario wel te ontwikkelen voor de Politie, maar niet te gebruiken in dit onderzoek.

De scenario's zijn tot slot voorgelegd aan drie gebiedsgebonden agenten. Een van de dingen die agenten benoemden was dat ze te weinig informatie hadden om, alleen op basis van de bevragingsassistent, een keuze te maken. Om deze reden is het volgende toegevoegd: *'In hoeverre is je antwoord gebaseerd op de bevragingsassistent en in hoeverre op je eigen inschatting?'* en '*Op basis van de informatie van de bevragingsassistent'.* De agent gaf aan dat door het toevoegen van een dergelijke vraag en opmerking hij zich beter kon inleven in het scenario. Een ander belangrijk punt dat naar voren kwam uit de gesprekken met de twee agenten was dat de vragen die gingen over ervaren morele verantwoordelijkheid als onduidelijk werden ervaren. De agenten vonden de vragen vaag en gaven aan niet goed te weten wat er werd bedoeld. Deze feedback ging over de schaal die gebaseerd was op de schaal van onderzoek van Hong (2020). Dit gaf reden om deze schaal niet te gebruiken en de uiteindelijke schaal te baseren op onderzoek van Kim & Hinds (2006). De uiteindelijke scenario's en de schaal gebaseerd op onderzoek van Kim & Hinds (2006) werd door een derde agent als goed begrijpbaar en realistisch ervaren en deze zijn dan ook gebruikt in dit onderzoek.

3.2 Operationalisatie

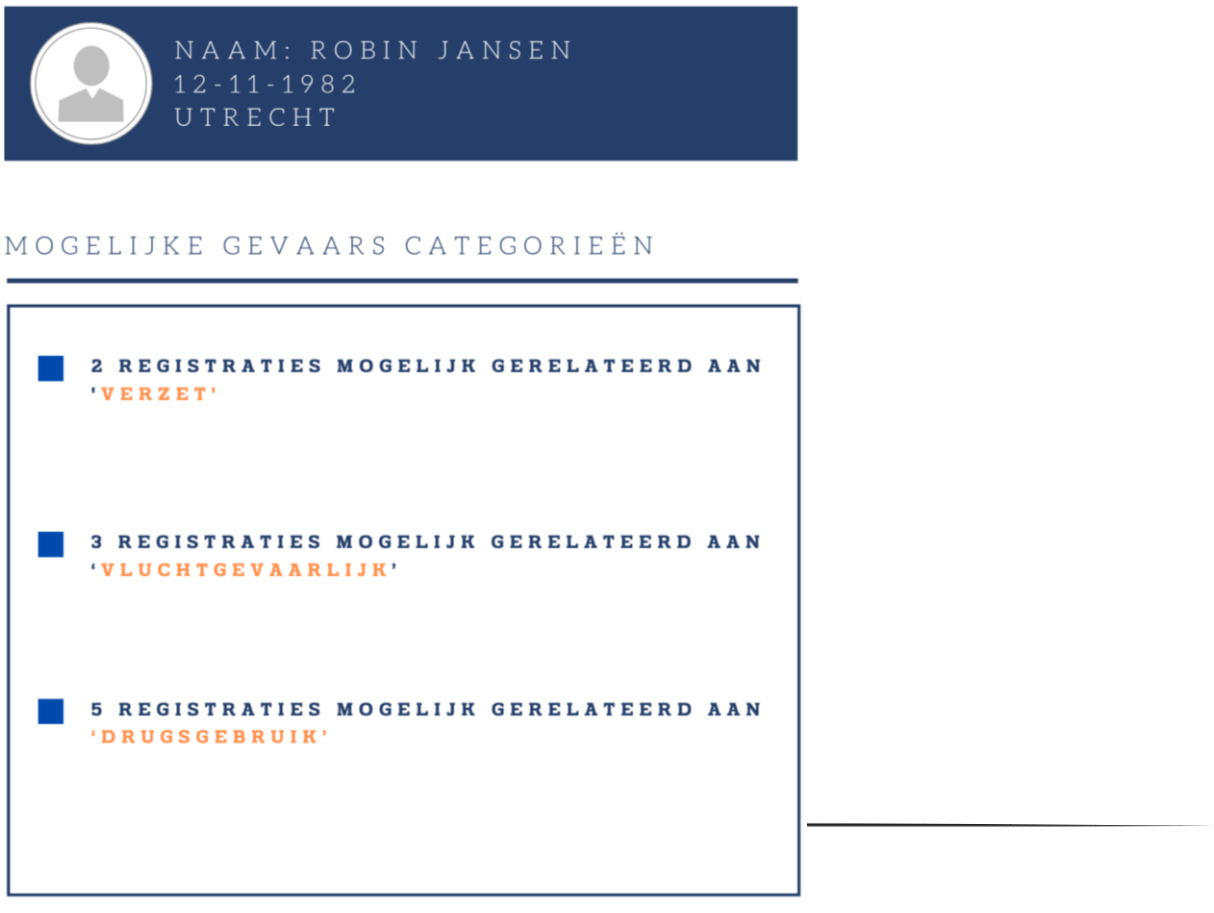
In deze paragraaf wordt ingegaan op de operationalisatie van algoritmische transparantie en de operationalisatie van ervaren morele verantwoordelijkheid. Allereerst zal worden toegelicht wat in de literatuur verstaan wordt onder een lage en hoge mate van transparantie. Vervolgens wordt uitgelegd hoe dit wordt toegepast in het experiment. In de laatste paragraaf zal een schaal worden gegeven die gebruikt wordt voor het meten van ervaren morele verantwoordelijkheid.

3.2.1 Operationalisatie algoritmische transparantie

Om algoritmes te ontwikkelen die gebruikers vertrouwen en verantwoordelijk kunnen stellen is het nodig om de 'black box' transparant te maken. In de literatuur wordt dit fenomeen ook wel 'opening up the black box' genoemd (De Fine Licht & De Fine Licht, 2020, p. 924). Dit betekent dat het algoritme een uitleg geeft voor zijn beslissingen of zijn werking. Het 'openen' van de black box maakt het beter mogelijk voor gebruikers om de werking van het AI-systeem te begrijpen en interpreteren, als ook het uitleggen ervan (e.g., Zarsky 2016; Lepri et al. 2017; Zerilli et al. 2018). In dit onderzoek is bij het scenario met een hoge mate van transparantie ook meer informatie weergegeven over de werking van het systeem. Daarbij is ook een korte uitleg gegeven. Het scenario met een hoge mate van transparantie past dus het fenomeen 'opening up the blackbox' meer toe dan het scenario met een lage mate van transparantie, omdat in het scenario met een lage mate van transparantie maar weinig informatie beschikbaar is.

In het theoretisch kader is beargumenteerd waarom visuele algoritmische transparantie geschikt wordt geacht om uitlegbaarheid van het AI-systeem te realiseren. Het systeem legt bij de scenario's met een hoge mate van transparantie uit waarom een gegeven antwoord juist of onjuist is en daarnaast kan de respondent de werking van het algoritme inzien door de tekstuele visualisatie in het scherm. De respondent kan namelijk de meldingen die behoren bij de weergegeven categorieën inzien, terwijl de scenario's met een lage mate van transparantie geen uitleg krijgen en ook geen meldingen kunnen zien. In alle scenario's zijn gender-neutrale namen gebruikt zoals 'Robin' en 'Renee', om het eventuele effect van gender uit te sluiten. In figuur 3.1 is een voorbeeld te zien van een scenario met een lage mate van transparantie.

Figuur 3.1



Op de bevragingsassistent met een lage mate van transparantie zijn de categorieën weergegeven. Je weet dat de persoon meldingen kan hebben die te maken hebben met deze categorieën, maar je weet niet op basis waarvan het AI-systeem deze categorieën laat zien. Op het moment dat het AI-systeem een fout maakt, krijgt de respondent hier geen procedurele of inhoudelijke uitleg over. De respondent krijgt alleen het volgende te zien:

> Uw inschatting blijkt onjuist. Je verwachtte dat Renee Visser wapengevaarlijk is, maar eenmaal op locatie aangekomen bleek die dat niet te zijn.
>
> Of
>
> Je inschatting blijkt juist te zijn. Renee Visser was inderdaad niet wapengevaarlijk.

Bij het systeem met een hoge mate van transparantie kun je een gedeelte van de melding zien. Daarnaast zie je welke woorden het systeem heeft gevonden en geassocieerd aan de categorie. Deze woorden zijn dikgedrukt. Het systeem met een hoge transparantie ziet eruit zoals in figuur 4.2.

Figuur 3.2

Op het moment dat de bevragingsassistent een fout maakt, wat in dit scenario het geval is, dan kan de agent dit zien door de zichtbare melding. De agent wordt daarnaast voorzien van een korte inhoudelijke uitleg. De respondent krijgt het volgende te zien:

> Je inschatting blijkt onjuist. Je verwachtte dat Renee Visser wapengevaarlijk is, maar eenmaal op locatie aangekomen bleek die dat niet te zijn. De bevragingsassistent had de categorie 'wapengevaarlijk' laten zien doordat het systeem het woord 'schroevendraaier' in een melding had gevonden, maar Renee Visser bleek toen zelf de melder te zijn.
>
> Of
>
> Je inschatting blijkt juist te zijn. Renee Visser was inderdaad niet wapengevaarlijk. De bevragingsassistent had de categorie 'wapengevaarlijk' laten zien doordat het systeem het woord 'schroevendraaier' in een melding had gevonden, maar Renee Visser bleek toen zelf de melder te zijn.

Het verschil tussen het systeem met een lage mate van transparantie en een hoge transparantie is dus dat de meldingen bij de scenario's met een lage mate van transparantie niet zichtbaar zijn en bij de scenario's met een hoge transparantie wel. Daarnaast wordt in de scenario's met een hoge mate van transparantie een korte uitleg gegeven over de werking van de bevragingsassistent en in de scenario's met een lage mate van transparantie wordt dat niet gegeven.

Over of de uitleg van de scenario's procedureel of inhoudelijk is valt zoals eerder genoemd te discussiëren. Enerzijds kan worden beargumenteert dat de uitleg procedureel is, omdat uitgelegd wordt hoe het systeem erbij gekomen is om de gekozen categorieën te laten zien. Namelijk doordat het woorden detecteert en plaatst onder een categorie. Aan de andere kant wordt ook uitgelegd *waarom* de bevragingsassistent de gekozen categorieën laat zien, namelijk omdat het een woord heeft gevonden dat bij die categorie past. Hieruit blijkt dat een procedurele en inhoudelijke uitleg elkaar soms kunnen overlappen. Een volledige weergave van de vignetten is te vinden in bijlage 7.1. In bijlage 7.2 is een weergave van de survey flow te zien. In de volgende paragraaf wordt uitgelegd hoe de afhankelijke variabele *ervaren morele verantwoordelijkheid* is geoperationaliseerd.

3.2.2 Operationalisatie ervaren morele verantwoordelijkheid

In deze paragraaf wordt uitgelegd hoe het concept 'ervaren morele verantwoordelijkheid' geoperationaliseerd is. De ervaren morele verantwoordelijkheid wordt in dit onderzoek gedefinieerd als de verantwoordelijkheid die iemand voelt over zijn beslissingen en acties (Oshana, 2004). Ervaren morele verantwoordelijkheid wordt gemeten aan het toekennen van succes en falen aan zichzelf of aan een systeem. De schaal die gebruikt wordt om dit te meten is gebaseerd op de schaal van Kim & Hinds (2006). Deze schaal wordt geschikt geacht, omdat zij ook het effect van transparantie op het toekennen van schuld of succes onderzoeken. Daarnaast is Cronbach's Alpha in het onderzoek van Kim & Hinds (2006) voor elke schaal hoger dan 0.8, op 'attribution of credit to robot' (twee vragen; α = .678) na en daarmee wordt het als betrouwbaar geacht.

In de schaal van Kim & Hinds (2006) worden drie concepten uitgewerkt: het toekennen van schuld of succes aan de robot, het toekennen van schuld of succes aan zichzelf en het toekennen van schuld of succes aan andere personen. Onder andere omwille van de reikwijdte van dit onderzoek is ervoor gekozen om het toekennen van schuld of succes aan andere personen niet te onderzoeken. Het toekennen van schuld of succes aan andere personen wordt in dit onderzoek daarnaast als minder relevant geacht, omdat van een agent wordt verwacht dat die morele verantwoordelijkheid voelt over zijn *eigen* keuzes en acties.

De stellingen worden vertaald naar het Nederlands en aangepast zodat ze goed aansluiten bij de casus. De stellingen worden gemeten aan de hand van een 7 punts Likert-schaal (1-7). Hierbij betekent 1 helemaal oneens en 7 helemaal eens. Aangezien er twee scenario's zijn afgespeeld wordt gekeken naar Cronbach's Alpha van beide scenario's. In tabel 3.2 staat een overzicht van de stellingen die de concepten meten.

Tabel 3.2: Concepten en stellingen

Het concept dat gemeten wordt	Stellingen
Toekennen juist/onjuist antwoord aan zelf	Het (on)juist inschatten van de situatie is voornamelijk te danken aan mijn eigen interpretatie van de informatie van de bevragings-asisstent
	Een compliment/klacht over het (on)juist inschatten van de situatie zou moeten worden toegeschreven aan mijzelf
Toekennen juist /onjuist antwoord aan bevragingsassistent	Het (on)juist inschatten van de situatie komt voornamelijk door de informatie die de bevragingsassistent heeft weergegeven
	Een compliment/klacht over het (on)juist inschatten van de situatie zou moeten worden toegeschreven aan de bevragingsassistent
Begrijpbaarheid	Het is mij duidelijk hoe de bevragingsassistent tot de weergegeven categorieën is gekomen

Zoals te zien is in tabel 3.2 zijn de vragen voor het juist en onjuist inschatten van de situatie samengevoegd. Dit is gedaan, omdat de hypotheses niet verschillen voor het geven van een juist of onjuist antwoord. Om de betrouwbaarheid van de schalen te meten is gekeken naar Cronbach's Alpha.

Scenario 1

De schaal voor het toekennen van een juist/onjuist antwoord aan zelf is niet betrouwbaar (twee vragen; α = .495). Dit betekent dat de twee vragen niet hetzelfde concept meten. De schaal voor het toekennen van een juist/onjuist antwoord aan de bevragingsassistent is wel redelijk betrouwbaar (twee vragen; α = .622).

Scenario 2

Voor scenario 2 geldt dat de schaal voor het toekennen van een juist/onjuist antwoord aan zelf niet betrouwbaar tot twijfelachtig is (twee vragen; α = .567). De schaal voor het toekennen van een

juist/onjuist antwoord aan de bevragingsassistent is wel redelijk betrouwbaar (twee vragen; α = .626).

Hieruit kan geconcludeerd worden dat het twijfelachtig is of de vragen hetzelfde concept meten, vooral als het gaat om de vragen die gaan over de agenten zelf. Om deze reden is ervoor gekozen om alle vragen afzonderlijk van elkaar te toetsen en analyseren.

3.2.3 Manipulatiecheck

Bij scenario 1 is het verschil in gemiddelde voor een lage mate van transparantie (*M*=3.70; *SD* = 2.060) en een hoge mate van transparantie (*M*=4.85 ; *SD* =1.646) wel significant (*t*(151)= -3.837 ; p = <0.001). De respondenten die een lage mate van transparantie te zien kregen reageren dus significant minder hoog op de vraag: '*Het is mij duidelijk hoe de bevragingsassistent tot de weergegeven categorieën is gekomen'* (schaal eens-oneens 1-7) dan de respondenten die een hoge mate van transparantie te zien kregen. De manipulatiecheck is voor scenario 1 dus geslaagd.

Bij scenario 2 is het verschil in gemiddelde voor een lage mate van transparantie (M=3.40 ; SD = 1.935) en een hoge mate van transparantie (M=4.90 ; SD = 1.481) ook significant (t (151)= -5.422; p= <.001). De respondenten die een lage mate van transparantie te zien kregen reageren dus significant minder hoog op de vraag: '*Het is mij duidelijk hoe de bevragingsassistent tot de weergegeven categorieën is gekomen'* (schaal eens-oneens 1-7). De manipulatiecheck is voor het tweede scenario dus ook significant.

Als de vraag '*Het is mij duidelijk hoe de bevragingsassistent tot de weergegeven categorieën is gekomen'* (schaal eens-oneens 1-7) uit scenario 1 en scenario 2 samen worden gevoegd, dan is de Cronbach's Alpha (α =0.763) betrouwbaar. De vraag heeft dus in scenario 1 hetzelfde gemeten als in scenario 2. Hieruit kan geconcludeerd worden dat zowel voor scenario 1 als scenario 2 geldt dat respondenten die een lage mate van transparantie te zien kregen significant minder hoog scoren op de vraag: '*Het is mij duidelijk hoe de bevragingsassistent tot de weergegeven categorieën is gekomen'* (schaal eens-oneens 1-7), dan respondenten die een hoge mate van transparantie te zien kregen. Het middel om transparantie te bereiken in dit onderzoek zorgt dus significant voor meer begrijpbaarheid en daarmee is de manipulatie in dit onderzoek geslaagd.

3.2.4. Overige vragen

Naast de vragen over ervaren morele verantwoordelijkheid en uitlegbaarheid zijn nog een aantal demografische vragen toegevoegd. Deze vragen zijn toegevoegd om te kijken of de steekproef een representatieve groep agenten omvat. Dit zijn vragen over geslacht, leeftijdscategorie, ervaring en hoogst genoten opleiding. De demografische vragen zijn te vinden in bijlage 7.3.

3.3 Selectie respondenten

Bij een surveyexperiment is het wenselijk dat de steekproef representatief is voor de populatie. In dit geval is de populatie het geheel aan GGP agenten in Nederland. Respondenten zijn benaderd via leidinggevenden van verschillende basisteams in Nederland. Dit zijn basisteams uit zowel het Noorden, Oosten, Westen, Zuiden als ook het Midden van Nederland. Met de regio is rekening

gehouden, omdat de steekproef representatief moet zijn voor de agenten in Nederland. Voor het selecteren van deze basisteams is rekening gehouden met hun deelname aan experiment K.I.D. Bewust zijn zoveel mogelijk teams benaderd die niet deelnemen aan experiment K.I.D., om overbelasting te voorkomen. Verder is nog één klas benaderd bij de Politieacademie met agenten in opleiding. Aangezien deze agenten in opleiding al vaker hebben meegelopen op straat werden zij geschikt geacht als respondenten. Tabel 3.3 geeft een kort overzicht van de totale steekproef en de populatie. Een compleet overzicht van de beschrijvende statistieken staat in bijlage 7.3.

Tabel 3.3: Overzicht steekproef

Variabele	Verdeling steekproef	Verdeling populatie*
Geslacht	72,4% Man	61%
	27% Vrouw	39%
Gemiddelde leeftijdscategorie	Tussen de 30-40	45,2
Gemiddeld aantal jaar ervaring	Meer dan 10 jaar	Onbekend
Gemiddeld opleidingsniveau	MBO	Onbekend

*Bron: Ministerie van Algemene Zaken (2022) en Nationale Politie (2020)

Allereerst is in de tabel te zien dat er in de steekproef iets meer mannen en iets minder vrouwen zitten dan in de populatie. Zo is in de steekproef bijvoorbeeld 72,4% man, terwijl dit in de populatie 61% is. De verhouding tussen mannen en vrouwen is wel redelijk representatief, namelijk ongeveer 3 kwart man en 1 kwart vrouw. Ten tweede is te zien in de tabel dat de gemiddelde leeftijd in de steekproef representatief is voor de populatie. De gemiddelde leeftijd in de populatie is 45,2 jaar en de gemiddelde leeftijd in de steekproef is de categorie tussen 30 en 40 jaar. Over het gemiddelde opleidingsniveau en gemiddeld aantal jaar ervaring van de populatie is niets bekend. Wel is bekend dat veel medewerkers loyaal zijn en lang blijven werken bij de Politie (Nationale Politie, 2020). Dit is een verklaring voor het feit dat meer dan de helft van de agenten in de steekproef meer dan 10 jaar ervaring hebben opgedaan (zie Tabel 7.3 in bijlage 7.3). Hieruit kan geconcludeerd worden dat de steekproef representatief is voor de populatie.

3.3.1 Beschrijving van de experimentele groepen

De respondenten zijn door gebruik te maken van een *random assignment* opgedeeld in twee willekeurige groepen (Bryman, 2016, p. 45). Dit geldt voor beide scenario's. Dit wordt gedaan zodat de onderzoeker ervan uit kan gaan dat het enige verschil tussen de groepen de manipulatie is. In deze paragraaf worden de experimentele groepen beschreven. Allereerst staan in de onderstaande tabellen de grootte van de steekproef en de verdeling tussen de controle groep en de experimentele groep.

Tabel 3.4: Steekproef scenario 1

	Steekproefgrootte (*n*)	Percentage (%)
Lage transparantie (controle groep)	*n* = 69	45,1
Hoge transparantie (experimentele groep)	*n* = 84	54,9
	n totaal = 153	100

Tabel 3.5: Steekproef scenario 2:

	Steekproefgrootte *(n)*	
Lage transparantie (controle groep)	*n* = 73	47,7
Hoge transparantie (experimentele groep)	*n* = 80	52,3
	n totaal= 153	100

De groepen zijn nagenoeg gelijk en dat betekent dat de randomisatie gelukt is. Ongeveer 50% van de respondenten heeft het scenario met een lage mate van transparantie gezien en ongeveer 50% heeft het scenario met een hoge mate van transparantie gezien. Doormiddel van een randomisatiecheck is gebleken dat er ook geen verschillen zijn in demografische gegevens tussen de twee groepen. De randomisatiecheck is terug te vinden in bijlage 7.4.

3.4 Data-analyse

Om het effect van algoritmische transparantie op ervaren morele verantwoordelijkheid te onderzoeken is gebruik gemaakt van een independent sample t-toets in het programma *SPSS Statistics 28*. Voor elke vraag is afzonderlijk een t-toets uitgevoerd, omdat bleek dat de Cronbach's Alpha voor de schalen van de concepten niet betrouwbaar of twijfelachtig waren. Voor het toetsen van de manipulatiecheck is ook een independent t-toets gebruikt. Daarnaast is gebruikgemaakt van een Chi Kwadraat toets om te kijken of er significante verschillen zaten tussen de groepen bij het beantwoorden van de vraag of Robin Jansen vluchtgevaarlijk is en of Renee Visser wapengevaarlijk is.

3.5 Ethische verantwoording

In deze paragraaf wordt beschreven op welke manier rekening is gehouden met ethische wetenschappelijke standaarden in dit onderzoek. Dit wordt gedaan aan de hand van vier ethische principes van Rosenberg (2015).

Om te beginnen zijn respondenten voorafgaand aan het onderzoek geïnformeerd over het onderzoek. Hierbij is gewezen op vrijwilligheid en vrijblijvendheid van het onderzoek. Daarnaast is benoemd dat het onderzoek gebruik wordt voor een masterscriptie, eventueel verder onderzoek van de Politie en/of mogelijk wetenschappelijke publicatie. Respondenten hebben na het lezen van deze informatie expliciet toestemming moeten geven voor hun deelname. Hiermee is het principe van 'informed consent' zo goed mogelijk gewaarborgd (Rosenberg, 2015, p. 263).

Dit onderzoek voldoet daarnaast ook aan het principe van 'making positive improvements' (Rosenberg, 2015, p. 264). Het onderzoek is erop gericht om het dagelijkse werk van gebiedsgebonden Politieagenten te ondersteunen. Daarnaast draagt het bij aan onderzoek over de

ethiek van artificial intelligence. De voordelen van dit onderzoek zijn naar alle waarschijnlijkheid voor respondenten niet direct voelbaar, maar indirect wel aanwezig.

Het volgende principe is het principe van ‘doing no harm’ (Rosenberg, 2015, p. 264). In dit onderzoek is het de verwachting dat er geen situaties zijn ontstaan die mogelijk schadelijk zijn voor respondenten. Ook is rekening gehouden met dit principe door veel verschillende basisteams te benaderen. Hierdoor hing het aantal respondenten niet af van een enkel team. Dit is belangrijk, omdat het niet wenselijk is dat agenten overbelast worden. Tot slot is aan respondenten de mogelijkheid geboden om in een open vraag aanvullende opmerkingen achter te laten. Een enkele respondent uitte hierin bedenkelijkheden over de nuttigheid van het experiment. De enige manier van schade die aangericht kan zijn aan respondenten zou dus mogelijk zijn dat zij het gevoel hebben dat hun tijd werd verspild.

Tot slot het principe van ‘eerlijkheid, gelijkheid en rechtvaardigheid’ (Rosenberg, 2015, p. 265). In dit experiment is niet verteld dat onderzoek werd gedaan naar het effect van transparantie op ervaren morele verantwoordelijkheid, omdat dit mogelijk kon zorgen voor sociaal gewenste antwoorden. Wel is aan de respondenten vermeld dat dit onderzoek bijdraagt aan het verder ontwikkelen van de bevragingsassistent. Verder zijn de basisteams uitgekozen op basis van beschikbare contacten. In die zin had niet elk basisteam een gelijke kans om in de steekproef te komen. Wel heeft elke agent van de benaderde basisteams een gelijke kans gehad om deel te nemen aan het experiment, omdat zij zijn benaderd via de mail.

Aanvullend voor de principes van Rosenberg (2015) geeft Bryman (2016, p. 131-133) aan dat de privacy van deelnemers ook niet in het gevaar mag worden gebracht. Om deze reden zijn alle gegevens van deelnemers op geen enkele manier terug te herleiden naar een individu. Dit is gedaan door alleen de nodige demografische gegevens te verzamelen en bij sommige vragen naar een categorie te vragen in plaats van een exact getal. Het programma *Qualtrics*, waarin de dataverzameling plaatsvond, wordt daarnaast als een veilig programma geacht.

Hoofdstuk 4: Resultaten

In dit resultatenhoofdstuk wordt het effect van algoritmische transparantie op ervaren morele verantwoordelijkheid getoetst. De twee experimenten, het experiment van scenario 1 en van scenario 2, worden afzonderlijk van elkaar geanalyseerd. Allereerst wordt de slider vraag geanalyseerd. Dan wordt doormiddel van een Chi kwadraat getoetst of er significante verschillen zijn tussen de groep met een lage mate van transparantie en de groep met een hoge mate van transparantie op de vragen die bepaalden of zij zouden uitkomen bij de stellingen die gingen over het 'juist' of 'onjuist' inschatten van de situatie. Tot slot worden de antwoorden die de respondenten gaven op de vragen over het toeschrijven van een 'juist' of 'onjuist' antwoord aan zichzelf of aan de bevragingsassistent geanalyseerd doormiddel van een independent t-toets en worden de volgende hypotheses besproken:

H1: Een hoge mate van transparantie zorgt ervoor dat succes of falen meer toegekend wordt aan de agent zelf dan aan de bevragingsassistent dan bij een lage mate van transparantie

H2: Een lage mate van transparantie zorgt ervoor dat succes of falen meer toegekend wordt aan de bevragingsassistent dan bij een hoge mate van transparantie

4.1 De wisselwerking tussen de eigen inschatting en de informatie van de bevragingsassistent

Allereerst wordt gekeken naar de slider-vraag. Deze bestond uit de vraag: *'In hoeverre is uw antwoord gebaseerd op uw eigen inschatting en in hoeverre op de informatie van de bevragings-assitent?'* (waarbij 1=eigen inschatting en 100= bevragingsassistent).

Scenario 1

De groep met een lage mate van transparantie had een gemiddelde van (*M*=78.05 ; *SD* =26.45 en de groep met een hoge mate van transparantie had een gemiddelde van (*M*=72.48; *SD* =27.44). Het verschil tussen de groepen is niet significant (*t*(157)= 1.299 ; p = .196). Beide groepen gaven dus in gelijke mate aan dat ze hun antwoord voornamelijk baseerden op informatie weergegeven door de bevragingsassistent.

Scenario 2

In het tweede scenario scoorden respondenten met een lage mate van transparantie een gemiddelde van (*M*=62.95 ; *SD* =29.78) en respondenten met een hoge mate van transparantie een gemiddelde van (*M*=72.86 ; *SD* =25.33). Dit verschil is significant (*t*(145)= -2.232 ; p =.027). Beide groepen gaven dus aan hun antwoord vooral te baseren op de informatie van de bevragingsassistent, maar bij de groep met een lage mate van transparantie was dit in mindere mate het geval. In post experimentele gesprekken gaven deelnemers van het experiment bij het scenario met een lage mate van transparantie dan ook aan dat zij de slider meer naar 'eigen inzicht' schuiven, omdat de bevragingsassistent maar heel weinig informatie geeft. De respondenten die een scenario met een hoge mate van transparantie kregen hadden meer informatie en baseerden hun antwoord dan ook in meerdere mate op de bevragingsassistent. Dit is interessant, omdat je met deze

redenering ook een verschil zou verwachten in scenario 1. Een mogelijke verklaring hiervoor is het verschil in de inhoud van het scenario.

4.2 Het eerste experiment: scenario 1

In het eerste scenario (*N* = 159) hadden 138 respondenten verwacht dat Robin Jansen vluchtgevaarlijk zou zijn en 21 respondenten hadden verwacht van niet. De groep met 138 respondenten kwam daardoor uit op stellingen die gingen over het 'juist' inschatten van de situatie en 21 respondenten kwamen uit op stellingen die gingen over het 'onjuist' inschatten van de situatie. De antwoorden op de vraag of Robin Jansen vluchtgevaarlijk zou zijn is schematisch weergegeven in tabel 4.1.

Tabel 4.1

Op basis van de informatie van de bevragingsassistent, verwacht je dat Robin Jansen vluchtgevaarlijk is?	**Groep lage mate van transparantie**	**Groep hoge mate van transparantie**
Ja, ik verwacht dat Robin Jansen vluchtgevaarlijk is (juist)	89,2%	84,7%
Nee, ik verwacht dat Robin Jansen niet vluchtgevaarlijk is (onjuist)	10,8%	15,3%

Van de groep die een lage mate van transparantie te zien kregen kwam dus uiteindelijk 89,2% uit op de stellingen die gaan over het juist inschatten van de situatie en 10,8% kwam uit op stellingen die gingen over het onjuist inschatten van de situatie. De groep met de lage mate van transparantie en de groep met de hoge mate van transparantie verschillen hierin niet significant (χ 2 = .694, p= .405). Bij het scenario waarin de bevragingsassistent geen fout maakt, komen dus beide groepen voornamelijk uit op het juiste antwoord. Dan worden nu de stellingen geanalyseerd.

Allereerst wordt gekeken naar de stelling: *'Het (on)juist inschatten van de situatie is voornamelijk te danken aan mijn eigen interpretatie van de informatie van de bevragingsassistent.'* Hierbij scoorde de groep met een lage mate van transparantie een gemiddelde van (*M*=4.58 ; *SD* =1.77) en de groep met een hoge mate van transparantie (*M*=4.99 ; *SD* =1.48). Dit verschil is niet significant (*t*(132.986)= -1.526 ; p = .065).

Dan de stelling: *'Een compliment/klacht over het (on)juist inschatten van de situatie zou moeten worden toegeschreven aan mijzelf'.* Ook bij deze stelling liggen de gemiddeldes niet ver uit elkaar. Hierbij scoorde de groep met een lage mate van transparantie een gemiddelde van (*M*=4.09 ; *SD* =1.80) en de hoge mate van transparantie (*M*=4.35 ; *SD* =1.40). Dit verschil was ook niet significant (*t*(126.59)= - .972 ; p = .166).

Bij de stelling *'Het (on)juist inschatten van de situatie komt voornamelijk door de informatie die de bevragingsassistent heeft weergegeven'* is ook geen significant verschil gevonden (*t*(151)= - .368 ; p = .357). Hierbij scoorde de groep met een lage mate van transparantie een gemiddelde van (*M*=5.49; *SD* =1.40) en de hoge mate van transparantie (*M*=5.57; *SD* =1.24).

Tot slot de stelling *'Een compliment over het (on) juist inschatten van de situatie zou moeten worden toegeschreven aan de bevragingsassistent'.* De groep met een lage mate van transparantie scoorde een gemiddelde van (*M*=4.87 ; *SD* =1.75) en de groep met een hoge mate van transparantie scoorde een gemiddelde van (*M*=5.08; *SD* =1.28). Dit verschil was wederom niet significant (*t*(121.86)= - 846; p = .200). In tabel 4.2 staat een overzicht van de getoetste stellingen.

Tabel 4.2

Stellingen	Gemiddelde lage mate van transparantie	Gemiddelde hoge mate van transparantie	Significantie
'Het (on)juist inschatten van de situatie is voornamelijk te danken aan mijn eigen interpretatie van de informatie van de bevragingsassistent.'	4.58	4.99	p = .065
'Een compliment/klacht over het (on)juist inschatten van de situatie zou moeten worden toegeschreven aan mijzelf'.	4.09	4.35	p = .166
'Het (on)juist inschatten van de situatie komt voornamelijk door de informatie die de bevragingsassistent heeft weergegeven'	5.49	5.57	p = .357
'Een compliment over het (on) juist inschatten van de situatie zou moeten worden toegeschreven aan de bevragingsassistent'	4.87	5.08	p = .200

Hieruit kan geconcludeerd worden dat er in het eerste scenario tussen de twee groepen geen significante verschillen zijn gevonden in de stellingen die gaan over de ervaren morele verantwoordelijkheid. De gemiddeldes van alle vier de stellingen in het eerste scenario lagen tussen de 4 en 6. Dat betekent dat respondenten het gemiddeld 'neutraal' tot 'eens' waren met de stellingen. Het juist of onjuist inschatten van de situatie werd dus ongeveer in een gelijke mate toegeschreven aan zelf en aan de bevragingsassistent.

4.3 Het tweede experiment: scenario 2

Terwijl bij scenario 1 de meeste mensen hadden gekozen voor het 'juiste' antwoord, was dit in scenario 2 verschillend. Het 'juist' inschatten van de situatie duidt erop dat de agent dacht dat Renee Visser *niet* wapengevaarlijk zou zijn en dat dit inderdaad ook *niet* het geval bleek te zijn. In tabel 4.3 zijn de gegeven antwoorden schematisch weergegeven.

Tabel 4.3

Op basis van de informatie van de bevragingsassistent, verwacht je dat Robin Jansen vluchtgevaarlijk is?	Groep lage mate van transparantie	Groep hoge mate van transparantie
Ja, ik verwacht dat Renee Visser wapengevaarlijk is (onjuist)	65,3%	9,9%
Nee, ik verwacht niet dat Renee Visser wapengevaarlijk is (juist)	34,7%	90,1%

Zoals te zien is in de tabel dacht 65,3% van de groep met een lage mate van transparantie dat Renee Visser wapengevaarlijk zou zijn, terwijl dit in de groep met een hoge mate van transparantie maar 9,9% is. Het verschil tussen deze twee groepen is significant (χ 2 =51.65, p= <.001). Dit is naar verwachting, omdat de groep met een hoge mate van transparantie meer informatie had dan de groep met een lage mate van transparantie. Daarnaast betekent deze uitkomst dat agenten de informatie die onder de gevarenclassificatie staat ook echt goed tot zich nemen.

Allereerst wordt gekeken naar de stelling: *'het (on)juist inschatten van de situatie is voornamelijk te danken aan mijn eigen interpretatie van de informatie van de bevragingsassistent.'* Hierbij scoorde de groep met een lage mate van transparantie een gemiddelde van (*M*=5.10 ; *SD* =1.56) en de groep met een hoge mate van transparantie (*M*=5.13 ; *SD* =1.41). Dit verschil is niet significant (*t*(151)= -.121 ; p = .452).Een score tussen de 5 en 6 betekent dat respondenten het 'een beetje eens' tot 'eens' zijn met de stelling.

Dan de stelling: *'Een compliment/klacht over het (on)juist inschatten van de situatie zou moeten worden toegeschreven aan mijzelf'.* Hierbij scoorde de groep met een lage mate van transparantie een gemiddelde van (*M*=4.30; *SD* =1.72) en de hoge mate van transparantie (*M*=4.55 ; *SD* =1.51). Dit verschil is niet significant (*t*(151)= - .952; p = .171). Beide groepen zijn het dus in gelijke mate neutraal tot een beetje eens met de stelling.

Bij de stelling *'Het (on)juist inschatten van de situatie komt voornamelijk door de informatie die de bevragingsassistent heeft weergegeven'* scoorde de groep met een lage mate van transparantie een gemiddelde van (*M*=5.00; *SD* =1.70) en de groep met een hoge mate van transparantie een gemiddelde van (*M*=5.35; *SD* =1.26. Dit verschil is niet significant (*t*(132.29)= -1.43 ; p = .077). De groep met een hoge mate van transparantie scoort dus hoger op deze vraag maar dit verschil is niet significant.

Tot slot de stelling *'Een compliment/klacht over het (on)juist inschatten van de situatie zou moeten worden toegeschreven aan de bevragingsassistent'.* De groep met een lage mate van transparantie scoorde een gemiddelde van (*M*=4.33 ; *SD* =1.68) en de groep met een hoge mate van transparantie scoorde een gemiddelde van (*M*=4.51; *SD* =1.52). Dit verschil was niet significant (*t*(151)= -.712 ; p = .239). De bevindingen van de toetsen in scenario 2 zijn weergegeven in tabel 4.4.

Tabel 4.4

Stellingen	Gemiddelde lage mate van transparantie	Gemiddelde hoge mate van transparantie	Significantie
'Het (on)juist inschatten van de situatie is voornamelijk te danken aan mijn eigen interpretatie van de informatie van de bevragingsassistent.'	5.10	5.13	p = .452
'Een compliment/klacht over het (on)juist inschatten van de situatie zou moeten worden toegeschreven aan mijzelf'.	4.30	4.55	p = .171
'Het (on)juist inschatten van de situatie komt voornamelijk door de informatie die de bevragingsassistent heeft weergegeven'	5.00	5.35	p = .077
'Een compliment over het (on) juist inschatten van de situatie zou moeten worden toegeschreven aan de bevragingsassistent'	4.33	4.51	p = .239

Hieruit kan geconcludeerd worden dat in scenario 2 ook geen significante verschillen zijn gevonden bij de stellingen die gaan over de ervaren morele verantwoordelijkheid.

4.3 Bevindingen in context: opmerkingen van respondenten

Om nader te kunnen verklaren waarom er geen effect is gevonden van algoritmische transparantie op ervaren morele verantwoordelijkheid is gekeken naar aanvullende opmerkingen van respondenten. De laatste vraag van het experiment was een open vraag. Bij deze vraag konden respondenten opmerkingen over het onderzoek achterlaten. Daarnaast is de onderzoeker fysiek aanwezig geweest bij twee basisteams. Tijdens deze bezoeken zijn van 33 scenario's en 17 respondenten aanvullende opmerkingen meegeschreven die respondenten hadden bij het maken van het experiment.

Uit het lezen van deze opmerkingen bleek een belangrijke nuance. De bevragingsassistent is slechts één van de informatiebronnen die agenten gebruiken om tot oordeelsvorming te komen. In een niet-experimentele setting zouden zij meerdere afwegingen maken op basis van verschillende informatiebronnen. Zo geeft een respondent bij de open vraag over aanvullende opmerkingen aan: *"Vraagstelling deed mij vermoeden dat het foutief inschatten van de situatie direct leidt tot verkeerd handelen/een klacht. Mijn ervaring leert dat de verstrekte informatie vooral de 'mindset' in positieve*

*zin beïnvloed doordat het je alerter maakt. Uiteindelijk wordt je handelen door meerdere factoren beïnvloed en draagt een adequate informatievoorziening daar aan bij."*Een andere respondent ondersteunt deze visie: *"Gevarenclassificatie is voor mij enkel een indicatie. Hoe ik de ander benader, kan een groot effect hebben op het resultaat. Geen idee of dat aspect terugkomt."* Op basis van deze redeneringen geven respondenten aan dat ze in plaats van "Ja ik verwacht" of "Nee, ik verwacht niet" de volgende woorden zouden gebruiken: "Ja ik hou er rekening mee" of "Nee, ik hou er geen rekening mee". Agenten zouden hun vervolgactie bepalen op basis van de informatie die de bevragingsassistent geeft in combinatie met informatie die zij hebben vanuit andere informatiebronnen, zoals eigen eerdere ervaringen, ervaringen van collega's, de meldkamer of de confrontatie op het moment zelf.

Een andere kwalitatieve bevinding is dat het systeem met een hoge mate van transparantie als prettiger lijkt te worden ervaren. Zo benoemt een respondent die het scenario met de hoge mate van transparantie gezien heeft in de aanvullende opmerkingen: *"In de tweede casus was het een stuk makkelijker de informatie te duiden, doordat er een korte samenvatting stond met betrekking tot de gevarenclassificatie. Dat je dat in 1 oogopslag kunt zien vind ik de grootste meerwaarde".* Het toevoegen van een dergelijke korte samenvatting aan de gevarenclassificaties, inclusief dikgedrukte woorden, wordt door sommige respondenten als een meerwaarde ervaren. In de gesprekken geven agenten ook aan dat een filtersysteem handig zou zijn, waarbij het mogelijk is om te filteren op categorie en rol. Deze wens kan echter ook gerealiseerd worden binnen het huidige BVI-IB systeem en hoeft niet per se betrekking te hebben op de bevragingsassistent. Bij respondenten die alleen de scenario's kregen te zien met een lage mate van transparantie was frustratie voelbaar. De informatie was te summier. Vooral het weglaten van de datum werd als een grote beperking ervaren. Dit is ook terug te lezen in de opmerkingen, zoals: *"Om een beter beeld te geven van de gevaarzetting zou ik ook willen weten hoe lang geleden deze gevarenclassificatie meldingen zijn opgemaakt".* Daarnaast wordt van het systeem een bepaalde mate van accuraatheid verwacht en struikelen veel respondenten over het woord 'mogelijk'.

Tot slot, de kwantitatieve bevindingen wijzen erop dat respondenten morele verantwoordelijkheid toeschrijven aan de bevragingsassistent, doordat ze het met de stellingen die hierover gaan neutraal tot eens zijn. Uit kwalitatieve bevindingen blijkt hier echter wel nuance in te zitten. Zo wijzen sommige respondenten erop dat het systeem gevoed wordt door collega's. Ook spreken zij uit het vreemd te vinden om te spreken over een compliment aan het systeem.

Samengevat, uit de kwalitatieve bevindingen komen drie deelconclusies naar voren: 1) de bevragingsassistent zou slechts één van de vele informatiebronnen zijn waarmee agenten moeten werken, 2) gebruikers werken het liefst met een transparant systeem en 3) het is discutabel of agenten daadwerkelijk morele verantwoordelijkheid toeschrijven aan systemen

5. Conclusie en discussie

In het resultatenhoofdstuk zijn de empirische bevindingen uiteen gezet. In dit hoofdstuk wordt de hoofdvraag beantwoord en worden beperkingen van het onderzoek besproken. De volgende vraag staat in dit onderzoek centraal: *Wat is het effect van algoritmische transparantie op de ervaren morele verantwoordelijkheid van gebiedsgebonden politieagenten?* Om deze vraag te beantwoorden zijn de volgende twee hypotheses getoetst:

H1: Een hoge mate van algoritmische transparantie zorgt ervoor dat succes of falen meer toegekend wordt aan de agent zelf dan aan de bevragingsassistent

H2: Een lage mate van algoritmische transparantie zorgt ervoor dat succes of falen meer toegekend wordt aan de bevragingsassistent dan aan de agent zelf

Beide hypotheses konden niet bevestigd worden. Op basis van de bevindingen in dit onderzoek is het antwoord op de hoofdvraag drieledig, deze punten worden hieronder verder uitgewerkt:

1. Algoritmische transparantie heeft geen effect op de ervaren morele verantwoordelijkheid
2. Algoritmische transparantie zorgt wel voor meer begrijpbaarheid van het systeem en het zorgt ervoor dat fouten in een AI systeem niet onopgemerkt blijven
3. Agenten baseren hun oordeelsvorming op basis van verschillende informatiebronnen. De oordeelsvorming van agenten en de ervaren morele verantwoordelijkheid die daaruit volgt moet daarom in een informatie-ecologie gezien worden.

Allereerst, bleek de groep met een lage mate van transparantie en de groep met een hoge mate van transparantie niet significant te verschillen met betrekking tot de stellingen die gingen over de ervaren morele verantwoordelijkheid. Respondenten schreven bij alle stellingen in dezelfde mate evenveel morele verantwoordelijkheid toe aan de bevragingsassistent als aan henzelf. Wel blijkt uit dit onderzoek dat agenten het met alle stellingen neutraal tot eens zijn en dus zowel aan zichzelf als aan de bevragingsassistent morele verantwoordelijkheid toekennen. Met inachtneming van dit gegeven is het dus niet zorgelijk dat transparantie niet heeft geleid tot meer ervaren morele verantwoordelijkheid. Het zou pas zorgelijk zijn als respondenten het oneens zouden zijn met de stellingen, vooral als deze gaan over henzelf. De conclusie uit dit onderzoek is dus dat het gevoel van ervaren morele verantwoordelijkheid niet vergroot kan worden doormiddel van algoritmische transparantie. Mocht de Politie het gevoel van ervaren morele verantwoordelijkheid onder haar agenten willen vergroten, dan zal er dus naar factoren buiten transparantie om gekeken moeten worden om dit te bewerkstelligen. Dit punt komt terug in de theoretische reflectie en in de wetenschappelijke aanbevelingen.

Ten tweede, algoritmische transparantie zorgt daarentegen wel voor meer begrijpbaarheid van het systeem. Ook zorgt dit middel van algoritmische transparantie ervoor dat *als* het systeem een fout maakt, het merendeel van de agenten deze fout zullen corrigeren. Dit helpt agenten om een betere inschatting van de situatie te kunnen maken. Algoritmische transparantie lijkt dus een effectieve functie te hebben voor het voorkomen van risicovolle situaties, ook op het moment dat het systeem een fout maakt. In de casus van de bevragingsassistent lijkt algoritmische transparantie zelfs een

noodzakelijke functie te hebben, omdat het systeem met een lage mate van transparantie te summier is om een oordeelsvorming op te baseren.

Tot slot, agenten lijken hun oordeelsvorming in meerdere mate te baseren op de bevragingsassistent, als dit hun enige beschikbare informatiebron is. Dat er geen effect is gevonden op ervaren morele verantwoordelijkheid kan dan ook mogelijk verklaard worden doordat de oordeelsvorming en de vervolgactie van de agent, hetgeen waarover de agent morele verantwoordelijkheid zou moeten voelen, van veel verschillende factoren afhankelijk is. Zo is het afhankelijk van: de informatie uit het systeem, maar ook van eerdere ervaringen, ervaringen van collega's, informatie uit de meldkamer en de confrontatie tussen de agent en de betreffende persoon op dat moment. Agenten combineren de informatie die de bevragingsassistent geeft vaak met informatie uit andere informatiebronnen. In een niet-experimentele setting, dus in de context van de 'echte wereld', moeten agenten dan ook meerdere afwegingen maken. Naast de keuzes en afwegingen die agenten moeten maken, heeft ook het gedrag van agenten effect op het uiteindelijke resultaat. De houding en 'stijl' in het benaderen van een verdachte verschilt per agent, maar vaak zorgt een gevarenclassificatie voornamelijk voor meer alertheid en leidt dit niet direct tot verkeerd handelen of een klacht. De oordeelsvorming van de agent moet daarom in een 'informatie-ecologie' worden gezien: een systeem van mensen, praktijken, waarden en technologieën in een specifieke lokale omgeving (Thaens, 2006, p. 35). In een informatie-ecologie staat niet de technologie, maar de menselijke activiteiten die gediend worden door de technologie centraal (Nardi & O'Day, 1999, p. 50). Hier wordt dieper op ingegaan in de theoretische en praktische reflectie.

5.1 Theoretische reflectie

De hypotheses in dit onderzoek zijn gebaseerd op verschillende theorieën die besproken zijn in de wetenschappelijke relevantie en het theoretisch kader. In de volgende alinea's wordt gereflecteerd op deze wetenschappelijke literatuur. De bevindingen uit dit onderzoek krijgen daarmee betekenis voor het bredere wetenschappelijke debat over het gebruik van AI in een context van street-level-bureaucrats.

Een dominante focus op het onderzoeken van nieuwe middelen van algoritmische transparantie

In het theoretisch kader kwam aan bod dat Kim & Hinds (2006) beredeneren dat het effect van algoritmische transparantie in hoge mate afhangt van het gebruikte middel om algoritmische transparantie te realiseren. Hoewel algoritmische transparantie daadwerkelijk een functie heeft en daarom bewust ingezet moet worden bij de Politie, zorgt het, zelfs als het middel bijdraagt aan meer begrijpbaarheid zoals in dit onderzoek, niet voor een effect op ervaren morele verantwoordelijkheid. Met die kennis kan de stelling van Kim & Hinds (2006) dan ook in twijfel worden gebracht. Het risico dat schuilt voor toekomstig onderzoek is dat onderzoekers in een 'tunnelvisie' alsmaar op zoek gaan naar een geschikt middel om algoritmische transparantie te realiseren, om ervaren morele verantwoordelijkheid te bewerkstelligen, terwijl het effect ervan er niet of nauwelijks lijkt te zijn. Het is daarom van meerwaarde om te onderzoeken welke andere factoren de ervaren morele verantwoordelijkheid beïnvloeden. Toekomstig onderzoek naar het effect van algoritmische transparantie op andere vraagstukken, zoals begrijpbaarheid, uitlegbaarheid, vertrouwen of het al dan niet opvolgen van aanbevelingen van het systeem, is daarom wel interessant. Vooral de rol van visualisatie is hierbij nog onderbelicht , terwijl er ook in dit onderzoek (evenals in onderzoek van

Wortham et al. (2017) bevindingen zijn gedaan die wijzen op een positief verband van visualisaties op de begrijpbaarheid van geautomatiseerde systemen.

De toekenning van morele verantwoordelijkheid aan geautomatiseerde systemen

In het theoretisch kader is aangehaald dat er discussie bestaat over de stelling of geautomatiseerde systemen een morele verantwoordelijkheid kunnen hebben (Friedman, 1990; Johnson 2006, Hong et al., 2020). Deze discussie raakt aan dit onderzoek, omdat gekeken is naar de morele verantwoordelijkheid die mensen toeschrijven aan de bevragingsassistent. Hoewel sommige onderzoekers beredeneren dat een systeem geen morele verantwoordelijkheid kan bekleden (Friedman, 1990; Johnson 2006), wijst onderzoek van Friedman et al. (1995) uit dat mensen wel daadwerkelijk verantwoordelijkheid *toeschrijven* aan systemen. Zo beoordeelden respondenten dat systemen besluitvormingscapaciteiten hebben en zelfs intenties (Friedman et al., 1995). In dit onderzoek is de rol van het algoritme niet groot. Het algoritme in de bevragingsassistent doet geen concrete aanbevelingen, maar maakt gebruik van clusteralgoritmes. Toch lijkt dit onderzoek het onderzoek van Friedman et al. (1995) te bevestigen, omdat mensen het neutraal tot eens waren met de stellingen die gingen over het toeschrijven van succes of falen aan de bevragingsassistent. Wel hadden sommige respondenten bedenkingen bij het toeschrijven van verantwoordelijkheid aan het systeem: collega's hebben immers het systeem gevoed. Experimenteel onderzoek van Hong et al. (2020), waaruit blijkt dat mensen geneigd zijn om meer schuld toe te kennen aan systemen dan aan andere mensen, lijkt in dit onderzoek dan weer niet te worden bevestigd. Mensen kennen in dezelfde mate verantwoordelijkheid toe aan de bevragings-asisstent als aan zichzelf. Wat betekenen deze bevindingen nu voor het wetenschappelijk debat?

Naar mijn inziens is een milieu waarin mensen te vaak verantwoordelijkheid toeschrijven aan een systeem niet gewenst. Als dit vaak gebeurd impliceert dit dat het systeem teveel fouten maakt om mee te werken of het impliceert dat gebruikers verantwoordelijkheid van zich afschuiven. Bij de Politie lijkt dit niet het geval. Agenten zijn geneigd om verantwoordelijkheid toe te schrijven aan zichzelf, zelfs als het systeem een fout maakt. Toch denk ik dat het zowel vanuit filosofisch als bestuurskundig perspectief interessant is als toekomstig onderzoek zich focust op de vraag waar mensen precies naar verwijzen als het gaat om het verantwoordelijk houden van een systeem. Verwijzen mensen naar de specifieke functionaliteit van het systeem of verwijzen zij voornamelijk naar de mensen die het systeem hebben gevoed? Als dan toch blijkt dat verwezen wordt naar de specifieke functionaliteit van het systeem moet naar mijn inziens opnieuw overwogen en onderzocht worden of algoritmische transparantie dan niet toch een effect heeft op ervaren morele verantwoordelijkheid.

5.2 Beperkingen van het onderzoek

Het onderzoek kent ook een aantal methodologische beperkingen die in deze paragraaf worden besproken. De wetenschappelijke aanbevelingen bouwen onder andere voort op deze beperkingen.

Ten eerste, het middel dat is ingezet om transparantie te bewerkstelligen bestond zowel uit een korte uitleg als uit een tekstuele visualisatie. Aangezien het gaat om een combinatie van beiden is daarom ook niet gekozen voor het begrip 'uitlegbaarheid' in de onderzoeksvraag maar voor het begrip 'algoritmische transparantie'. De manipulatie in dit experiment varieerde dan ook op

meerdere vlakken. Hierdoor kan alleen geconcludeerd worden dat het middel in zijn geheel gezorgd heeft voor meer begrijpbaarheid van het systeem, maar niet welk specifiek onderdeel ervan. In hoeverre de visualisatie en in hoeverre de uitleg heeft bijgedragen aan de begrijpbaarheid van het systeem is met dit design niet vast te stellen. Hoewel de tekstuele visualisatie summier is, zijn er in de kwalitatieve bevindingen wel aanwijzingen dat de visualisatie een rol heeft gespeeld. De visualisatie in combinatie met gevarenclassificatie werd door sommige respondenten als een meerwaarde ervaren. Nader onderzoek naar het gebruik van visualisatie als vorm van transparantie is daarom nodig.

Een andere beperking binnen dit experiment is dat de stellingen die gingen over het toeschrijven van een juist/onjuist antwoord aan de bevragingsassistent wel redelijk betrouwbaar leken (α = .622, α = .626), maar de stellingen die gingen over het toeschrijven van een juist/onjuist antwoord aan zelf waren twijfelachtig (α = .495, α = .567). Hierdoor is het lastig om te stellen of de vragen die gingen over 'zelf' wel echt hetzelfde concept hebben gemeten. Dit tast de betrouwbaarheid van het onderzoek aan. In paragraaf 6.1 wordt op basis van deze bevinding een aanbeveling gedaan voor vervolgonderzoek.

Tot slot heeft de onderzoeker pas later in het onderzoek een dienst meegelopen met agenten. Hoewel alvorens het onderzoek wel met agenten is gesproken wordt het toch als een beperking ervaren dat pas later in het onderzoek een completere beeldvorming kon worden geschetst van de belevingswereld van agenten. Door het meelopen met een dienst werd de rol van bevragings-systemen in de dagelijkse realiteit van de Politie een stuk duidelijker. Mocht het onderzoek opnieuw worden uitgevoerd, dan zou de onderzoeker er voor kiezen om eerst diensten mee te lopen met agenten. Dit zou namelijk kunnen bijdragen aan de ecologische validiteit van het onderzoek en aan het ontwikkelen van realistische(re) vignetten.

5.3 Aanbevelingen voor vervolgonderzoek

Aanbeveling 1: Nader onderzoek naar visualisatie als vorm van algoritmische transparantie, maar niet met betrekking tot ervaren morele verantwoordelijkheid

De eerste aanbeveling heeft betrekking op het middel om algoritmische transparantie te realiseren. Alhoewel in de theoretische reflectie besproken is dat algoritmische transparantie geen effect heeft op de ervaren morele verantwoordelijkheid, is het wel interessant om te kijken naar het effect van het middel van algoritmische transparantie op andere vraagstukken. Dit kunnen vraagstukken zijn rondom begrijpbaarheid, uitlegbaarheid, vertrouwen of het al dan niet opvolgen van aanbevelingen van een systeem. De vorm van visualisaties als middel is daarbij vaak nog onderbelicht, terwijl er net zoals in het onderzoek van Wortham et al. (2017) ook in dit onderzoek (kwalitatieve) bevindingen zijn gevonden die wijzen op het gegeven dat visualisaties bijdragen aan de begrijpbaarheid van het systeem. Kwantitatief moet deze bevinding nader onderzocht worden, omdat dit onderzoeksdesign alleen iets kan zeggen over de combinatie van een tekstuele visualisatie met een korte uitleg.

Aanbeveling 2: Onderzoek contextfactoren die ervaren morele verantwoordelijkheid kunnen vergroten

De tweede aanbeveling gaat over de contextfactoren die meespelen bij het realiseren van ervaren morele verantwoordelijkheid. Een van de conclusies die uit dit onderzoek naar voren is gekomen is dat ervaren morele verantwoordelijkheid van agenten niet vergroot kan worden doormiddel van algoritmische transparantie. Mocht de Politie het gevoel van ervaren morele verantwoordelijkheid onder haar agenten willen vergroten, dan zal er dus naar factoren buiten transparantie om gekeken moeten worden om dit te bewerkstelligen. Vanuit wetenschappelijk oogpunt is het interessant om nader te onderzoeken welke andere factoren kunnen bijdragen aan het vergroten van ervaren morele verantwoordelijkheid.

Aanbeveling 3: Onderzoek naar de verantwoordelijkheid van systemen

Een ander inzicht dat is opgedaan tijdens dit onderzoek is dat agenten verantwoordelijkheid lijken toe te schrijven aan systemen. Naar mijn inziens is het van belang om te onderzoeken waar agenten precies naar verwijzen als ze het hebben over ervaren morele verantwoordelijkheid van zichzelf of een systeem. Zoals eerder is benoemd verwijzen agenten bij zichzelf vooral naar keuzes, gedrag of houdingen als factoren die hun oordeelsvorming beïnvloeden, maar waar verwijzen agenten precies naar als ze het hebben over de oordeelsvorming van een systeem? Het is dan ook aan te bevelen om nader kwalitatief onderzoek te doen naar de betekenisgeving van agenten aan het begrip *ervaren morele verantwoordelijkheid.*

Aanbeveling 4: Ontwikkel een betrouwbare schaal voor het meten van ervaren morele verantwoordelijkheid

Tijdens het ontwikkelen van dit experiment is de schaal om ervaren morele verantwoordelijkheid te meten meerdere malen aangepast. Zo werd als eerst de schaal van Hong et al. (2020) getest. Echter, de stellingen van deze schaal werden door de agenten als onduidelijk en vaag ervaren. Hierdoor was de schaal gebaseerd op de schaal van Hong et al. (2020) niet bruikbaar. De uiteindelijke schaal is gebaseerd op de schaal van Kim & Hinds (2006). De stellingen die in deze schaal stonden vonden agenten wel begrijpbaar en de Cronbach's Alpha in het onderzoek van Kim & Hinds (2006) was bijna bij alle concepten hoger dan 0,8. Deze schaal is op verschillende manieren vertaald in het Nederlands en uiteindelijk is gekozen voor de meest geschikte optie. In dit onderzoek bleek echter uit de Cronbach's Alpha dat de schaal om ervaren morele verantwoordelijkheid te meten twijfelachtig was. Dit was vooral zo bij de stellingen die gingen over het toeschrijven van succes of falen aan zelf. Op basis van deze bevinding is het relevant dat er meer onderzoek gedaan wordt naar een betrouwbare schaal voor het meten van ervaren morele verantwoordelijkheid. Het is hierbij interessant om schalen die gebruikt worden in verschillende organisatiecontexten met elkaar te vergelijken. Daarnaast kan gekeken worden naar het toevoegen van vragen aan de schaal van Kim & Hinds (2006), om zo de betrouwbaarheid te waarborgen. Daarnaast kun je door het toevoegen van meer vragen een bredere blik werpen op het vraagstuk, wat kan resulteren in een genuanceerder beeld.

5.4 Maatschappelijke reflectie

Een dominante focus op de inzet van nieuwe technologische middelen

Op basis van de bevindingen in dit onderzoek lijkt er zowel in wetenschappelijk onderzoek als in de praktijk een dominante focus te liggen op het onderzoeken en inzetten van nieuwe technologische middelen. Hieronder wordt uiteengezet hoe deze focus tot uiting komt in de praktijk bij de Politie.

Bij de Politie lijkt een sterke focus te liggen op het inzetten van nieuwe technologische middelen zoals de kennisassistent of de bevragingsassistent. Dit is opvallend, omdat huidig bestaande systemen zoals BVI-IB nauwelijks tot niet aan innovatie toekomen. De focus lijkt te liggen op de inzet van nieuwe middelen, terwijl de bestaande technologische middelen nog niet geoptimaliseerd zijn. Een ander inzicht dat is opgedaan tijdens dit onderzoek, is dat agenten, vaak in een korte tijd, afwegingen moeten maken tussen veel verschillende informatiebronnen. Dit doet de vraag rijzen of het toevoegen van een nieuwe informatiebron de agenten daadwerkelijk ondersteund.

Mocht de bevragingsassistent gerealiseerd worden, dan is het belangrijk om er rekening mee te houden dat de bevragingsassitent slechts één van de vele informatiebronnen is waar een agent mee te maken heeft. De Politie moet goed nagaan in welke specifieke casussen de bevragingsassistent, zoals die vormgegeven is in dit onderzoek, een meerwaarde vormt voor de agent of in hoeverre het misschien juist zorgt voor meer onduidelijkheid of een slechtere besluitvorming. Het toevoegen van een informatiebron kan onbedoeld meer bureaucratie tot gevolg hebben, omdat ook het gebruik van de bevragingsassistent administratief moet worden vastgelegd en verantwoord. Een bureaucratische omgeving zorgt paradoxaal genoeg juist voor een slecht milieu om innovaties van de grond te krijgen (Thaens, 2006, p. 33). Hoewel het inzetten van nieuwe technologische middelen de werksituatie van agenten op korte termijn lijkt te verbeteren, kan het dus op lange termijn ervoor zorgen dat oude systemen nog moeilijker geoptimaliseerd kunnen worden.

Als men toekomstig wil gaan werken met de bevragingsassistent, dan moet goed onderzocht worden in welke specifieke casuïstiek het middel daadwerkelijk de werksituatie van de agent beter maakt. De Politie moet naar mijn inziens een heroverweging maken of zij de focus willen houden op het ontwikkelen van een nieuwe technologische middelen zoals de bevragingsassistent, of dat zij de focus willen verleggen naar het onderzoeken van achterliggende oorzaken die de innovatie van huidige systemen (zoals BVI-IB) belemmeren.

Samenvattend, de dominante focus op technologie in zowel de wetenschap als in de praktijk moet naar mijn inziens meer ruimte maken voor een holistische blik, waarbij de mens die door de technologie gediend wordt centraal staat. Dit wordt ook wel een ‘informatie-ecologie’ genoemd (Nardi & O’Day, 1999, p. 50).

5.5 Praktische aanbevelingen

Aanbeveling 1: Wat draagt daadwerkelijk bij aan innovatie?

Zoals besproken is in de praktische reflectie moet de Politie naar mijn inziens een heroverweging maken of zij de focus willen houden op het ontwikkelen van nieuwe technologische middelen zoals de bevragingsassistent, of dat zij de focus willen verleggen naar het onderzoeken van achterliggende

oorzaken die de innovatie van huidige systemen (zoals BVI-IB) belemmeren. Mogelijke oorzaken voor de belemmering van innovatie kunnen worden gevonden in bureaucratisering, maar ook zaken als de organisatiestrategie, de cultuur en de structuur van de organisatie en onderlinge machtsverhoudingen kunnen een rol spelen (Thaens, 2006, p. 26- 36). Het is hierbij interessant om te kijken naar de achterliggende oorzaak achter de wens om nieuwe technologische middelen in te zetten. Welke betekenis geeft het Projectteam AI eigenlijk aan het begrip *innovatie*? Draagt het ontwikkelen van de bevragingsassistent daadwerkelijk bij aan innovatie binnen de Politie of is het onbedoeld een voedingsbodem voor meer bureaucratie, wat resulteert in een slecht milieu om innovatie van de grond te krijgen? Het is mijn aanbeveling vanuit bestuurskundig perspectief om deze vragen met elkaar te bediscussiëren en zaken te heroverwegen waar nodig.

Aanbeveling 2: Zet de bevragingsassistent bewust in de informatie-ecologie van de agent

Mijn tweede aanbeveling is dat niet het inzetten van nieuwe technologie centraal moet staan, maar het optimaliseren van de werkomgeving van de agent. Op dit moment lijkt de dominante focus op technologie de sociale context waarin de technologie gebruikt wordt te overschaduwen. De bevragingsassistent is slechts één onderdeel van een soort web met allemaal verschillende informatiebronnen die samen in relatie staan met de agent. Er moet goed gekeken worden naar specifieke casuïstiek waarbij de bevragingsassistent daadwerkelijk een meerwaarde oplevert. Hoe ziet de dynamiek tussen de verschillende informatiebronnen er precies uit? Welke afwegingen maken agententen in welke situatie en waarom? Is er een wens om te werken met een nieuwe bevragingsassistent of ligt daaronder een wens om te werken met geoptimaliseerde al bestaande systemen? Onder agenten heerst een behoefte voor een soort filtersysteem waarin je kunt filteren op registratie en rol, maar moet dit bij een nieuw systeem of kan dit ook toegepast worden op het huidige systeem? Op basis van deze gedachtegangen is mijn specifieke aanbeveling voor de Politie om meer aandacht te schenken aan het gebruikersperspectief en de informatie-ecologie van de agent waarin de bevragingsassistent een rol zou moeten spelen. De agent moet zich niet aanpassen aan het systeem, maar het systeem moet zich aanpassen aan de agent.

Aanbeveling 3: Algoritmische transparantie van de bevragingsassistent heeft een noodzakelijke functie

Een derde en daarmee laatste praktische aanbeveling is specifiek voor het design van de bevragingsassistent. Mocht de Politie voornemens zijn te werken met een dergelijk AI-systeem, dan moet het een slim ontwerp zijn, waarin: 1) de visuele vormgeving bijdraagt aan de begrijpbaarheid van het systeem, en 2) er geen ruimte mag zijn voor semantische discussies (de agenten moeten allemaal dezelfde mate van begrip hebben over de gebruikte definities). Algoritmische transparantie heeft niet alleen een functie, maar is noodzakelijk voor het inschatten van risicovolle situaties. Daarnaast wordt van de bevragingsassistent een bepaalde mate van accuraatheid verwacht en kunnen agenten met interpreteerbare woorden zoals ‘mogelijk’ niet werken. Het valt dan ook aan te raden om voor de visuele vormgeving een grafisch ontwerper in te schakelen, die met zijn expertise kan nadenken over hoe een boodschap het beste overgebracht kan worden op de gebruiker. Daarnaast is het belangrijk dat de Politie, alvorens het systeem geïmplementeerd wordt, trainingen gaat geven aan haar agenten over de gebruikte definitiestellingen en hoe een en ander geïnterpreteerd dient te worden.

6. Literatuur

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. Journal of Operations Management, 53(1), 63-70.

Amnesty International. (2021, 6 december). *Toeslagenschandaal is mensenrechtenschending, zegt*. Geraadpleegd op 17 mei 2022, van https://www.amnesty.nl/actueel/toeslagenaffaire-is mensenrechtenschending-zegt-amnesty-international

Aristoteles (1985), The Nicomachean Ethics, trans. by Terence Irwin. Hackett Publishing Co, 1985.

Blom-Hansen, J., Morton, R., & Serritzlew, S. (2015). Experiments in public management research. International Public Management Journal, 18(2), 151-170.

Boer, A. (2009). Legal theory, sources of law and the semantic web. IOS Press. https://doi.org/10.3233/978-1-60750-003-2-i

Bryman, A. (2016). Social Research Methods (5e ed.). Oxford: Oxford University Press

Burrell, J. (2016). How the machine 'thinks:' Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 205395171562251.

Coleman, J. S. 1990. Foundations of Social Theory. Cambridge, Mass: Harvard University Press

Cucciniello, M., Porumbescu, G. A., & Grimmelikhuijsen, S. (2016). 25 Years of Transparency Research: Evidence and Future Directions. *Public Administration Review*, *77*(1), 32–44. https://doi.org/10.1111/puar.12685

Curtin, D., & Meijer, A. J. (2006). Does transparency strengthen legitimacy? *Information Polity*, *11*(2), 109–122. https://doi.org/10.3233/ip-2006-0091

De Fine Licht, K., & De Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision making. *AI & SOCIETY*, *35*(4), 917–926. https://doi.org/10.1007/s00146-020-00960-w

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3064761

Eshleman, A. (2009), Moral Responsibility, The Stanford Encyclopedia of Philosophy (Winter 2009 Edition), Zalta, E. N. (editor), http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/.

Europese Commissie. (2019, 8 april). Ethische Richtsnoeren voor Betrouwbare KI. https://www.betabit.nl/media/4614/ethicsguidelinesfortrustworthyai-nl.pdf

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, *7*(1). https://doi.org/10.1057/s41599-020-0494-4

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, *17*(4–5), 663–671. https://doi.org/10.1080/09614520701469955

Friedman, B., Moral Responsibility and Computer Technology, Erin Document Reproduction Services, April 1990.

Friedman, B and Millett, L, "It's the computer's fault: reasoning about computers as moral agents", Proceedings of the CHI 1995, Conference on Human Factors on Computer Systems, ACM, New York, 1995.

Grimmelikhuijsen, S. (2022). Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making. *Public Administration Review*. https://doi.org/10.1111/puar.13483

Grimmelikhuijsen, S. G., 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., Jilke, S., Olsen, A. L., & Tummers, L. (2017). Behavioral public administration: Combining insights from public administration and psychology. *Public Administration Review*, *77*(1), 45-56.

Hansen, H. K., & Flyverbom, M. (2015). The politics of transparency and the calibration of knowledge in the digital age. Organization, 22, 872–889.

Haugeland, J. (1989). Artificial Intelligence: The Very Idea. Bradford, Engeland: Bradford: A Bradford Book.

Heckman,C., and Wobbrock, J (1998) "Liability for autonomous agent design", Proceedings of the second international conference on Autonomous agents, Minneapolis, Minnesota, United States, pp.392-399.

Heimstädt, M. (2017). Openwashing: A decoupling perspective on organizational transparency. Technological Forecasting and Social Change, 125, 77–86

Hellström, T. (2012). On the moral responsibility of military robots. *Ethics and Information Technology*, *15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2

Hengst, M. D., & Ter Mors, J. (2017). *Informatiegestuurd politiewerk in de praktijk* (1ste editie). Vakmedianet.

High-Level Expert Group on AI. (2018, juni). *Ethische richtsnoeren voor betrouwbare KI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hochwarter, W. A., G. R. Ferris, M. B. Gavin, P. L. Perrewé, A. T. Hall, and D. D. Frink. 2007. "Political Skill as Neutralizer of Felt Accountability—job Tension Effects on Job Performance Ratings: A Longitudinal Investigation." Organizational Behavior and Human Decision Processes 102 (2): 226–239. doi:10.1016/j.obhdp.2006.09.003.

Holmquist, L. E. (2017). Intelligence on tap. *Interactions*, *24*(4), 28–33. https://doi.org/10.1145/3085571

Hong, J. W., Wang, Y., & Lanz, P. (2020). Why Is Artificial Intelligence Blamed More? Analysis of Faulting Artificial Intelligence for Self-Driving Car Accidents in Experimental Settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693

Hood C (2006) Transparency in historical perspective. In: Hood C, Heald D (eds) Transparency: the key to better governance?. Oxford University Press, Oxford, pp 3–23

Kenniscentrum Data & Maatschappij. (2020). *Ethische principes en (niet-)bestaande juridische regels voor AI*. Geraadpleegd op 12 juni 2022, van https://data-en-maatschappij.ai/publicaties/ethische principes-en-niet-bestaande-juridische-regels-voor-ai

Kim, T., & Hinds, P. (2006, 1 september). *Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction*. IEEE Conference Publication | IEEE Xplore. Geraadpleegd op 11 april 2022, van https://ieeexplore.ieee.org/abstract/document/4107789/

Kim, T. W., & Routledge, B. R. (2018). Informational Privacy, A Right to Explanation, and Interpretable AI. In 2018 IEEE Symposium on Privacy-Aware Computing (PAC), 64-74.

Koene, A., European Parliament. Directorate-General for Parliamentary Research Services, Clifton, C. W., Hatada, Y., Webb, H., Patel, M., Machado, C., LaViolette, J., Richardson, R., & Reisman, D. (2019). *A Governance Framework for Algorithmic Accountability and Transparency*. UTB.

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2017) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 2017:1–17

Meijer, A., & Grimmelikhuijsen, S. (2020). Responsible and accountable algorithmization. *The Algorithmic Society*, 53–66. https://doi.org/10.4324/9780429261404-5

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. https://doi.org/10.1037/h0076486

Ministerie van Algemene Zaken. (2022, 17 mei). *Jaarverantwoording 2021 Politie Nederland*. Jaarverslag | Rijksoverheid.nl. Geraadpleegd op 17 juni 2022, van

https://www.rijksoverheid.nl/documenten/jaarverslagen/2022/05/18/nationale-politie-2021

Ministerie van Justitie en Veiligheid. (2019). *Evaluatierapport over de aanslag in Utrecht op 18 maart 2019* (deel B). Rijksoverheid.nl.https://www.rijksoverheid.nl/documenten/rapporten/2021/05/26/tk bijlage-6-rapport-aanslagutrecht-deel-b

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M. A. (2021). Traceability for Trustworthy AI: A Review of Models and Tools. *Big Data and Cognitive Computing*, *5*(2), 20. https://doi.org/10.3390/bdcc5020020

Nardi, B.A. & V.L. O'Day (1999). Information Ecologies. Using Technology with Heart, Cambridge, Massachusetts: The MIT Press.

Nationale ombudsman. (2010, 11 november). *Ombudsman maakt zich zorgen over onjuiste gegevens in politiesysteem*. Geraadpleegd op 15 juni 2022, van https://www.nationaleombudsman.nl/nieuws/2010/ombudsman-maakt-zich-zorgen-over-onjuiste gegevens-in-politiesysteem

Oshana, M. (2004). Moral Accountability. *Philosophical Topics*, *32*(1), 255–274. https://doi.org/10.5840/philtopics2004321/22

Overman, S., Schillemans, T., & Grimmelikhuijsen, S. (2020). A validated measurement for felt relational accountability in the public sector: gauging the account holder's legitimacy and expertise. *Public Management Review*, *23*(12), 1748–1767. https://doi.org/10.1080/14719037.2020.1751254

Park, S., & Gil-Garcia, J. R. (2022). Open data innovation: Visualizations and process redesign as a way to bridge the transparency-accountability gap. *Government Information Quarterly*, *39*(1), 101456. https://doi.org/10.1016/j.giq.2020.101456

Peters, B. Guy. 2014. Accountability in Public Administration. In The Oxford Handbook of Public Accountability, 1st ed., edited by Mark Bovens, Thomas Schillemans, and Robert E. Goodin, 211–25. Oxford: Oxford Univ. Press

Porumbescu, 2015 . Linking Transparency to Trust in Government and Voice. American Review of Public Administration. Published electronically on October 5. doi:10.1177/0275074015607301

Ployhart, R. E., & Ryan, A. M. (1997). Toward an Explanation of Applicant Reactions: An Examination of Organizational Justice and Attribution Frameworks. *Organizational Behavior and Human Decision Processes*, *72*(3), 308–335. https://doi.org/10.1006/obhd.1997.2742

Roberts, N. C., & Wargo, L. (1994). The Dilemma of Planning in Large-Scale Public Organizations: The Case of the United States Navy. *Journal of Public Administration Research and Theory*, 469–491. https://doi.org/10.1093/oxfordjournals.jpart.a037227

Romzek, B. S., K. LeRoux, and J. M. Blackmar. 2012. "A Preliminary Theory of Informal Accountability

among Network Organizational Actors." Public Administration Review 72 (3): 442–453. doi:10.1111/j.1540-6210.2011.02547.x.

Rosenberg, A. (2015). Philosophy of Social Science. Amsterdam: Athenaeum Uitgeverij.

Russell, S, J. & Norvig, P. (2013). Artificial intelligence: a modern approach. London: Pearson Education Limited.

Santoni De Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Schiff, D. S., Schiff, K. J., & Pierson, P. (2021). Assessing public value failure in government adoption of artificial intelligence. *Public Administration*, 1–21. https://doi.org/10.1111/padm.12742

Schlenker, B. R., & Weigold, M. F. (1989). Self-identiWcation and accountability. In P. Rosenfeld & R.
Slootweg, P. (2016). *De implementatie van Hohfeldian legal concepts, ambiguïteit en traceerbaarheid met Semantic Webtechnologieën*. Open Universiteit. https://core.ac.uk/download/pdf/55539462.pdf

A. Giacalone (Eds.), Impression management in the organization (pp. 21–43). Hillsdale, NJ: Lawrence Erlbaum.

Sinha, R., & Swearingen, K. (2002). The Role of Transparency in Recommender Systems. *School of Information Management & Systems*, 830–831.
https://dl.acm.org/doi/abs/10.1145/506443.506619?casa_token=PjbvwuerM4gAAAAA:0vPFRt5U4Y
Yn7mVf6iP-TmAe0qKCOb767x93rod5mouFwm2Kc7LiZK1NBcXmr5BrPo9m7mStJtK

Shrum, K., Gordon, L., Regan, P., Maschino, K., Shark, A.R. & Shropshire, A. (2019) AI and its impact on public administration. Washington, D.C.: National Academy of Public Administration. Available at: https://www.napawash.org/uploads/ Academy_Studies/9781733887106.pdf.

Stahl, B. C. (2004). Responsible management of information systems. Hershey: Idea-Group Publishing

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems* (pp. 1–9). Glasgow, Scotland, UK.

Thaens, M. (2006). *Verbroken verbindingen hersteld?* LEMMA.
https://repub.eur.nl/pub/8213/ThaensOratie%20eindversie%20gedrukt.pdf

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, *29*(3), 230–241. https://doi.org/10.1080/09540091.2017.1310182

Thompson, D.F. 1980. "Moral responsibility and public officials: The problem of many hands." American Political Science Review 74(4): 905–916.

Van Eck, M. (2018). Geautomatiseerde ketenbesluiten & rechtsbescherming: Een onderzoek naar de praktijk van geautomatiseerde ketenbesluiten over een financieel belang in relatie tot rechtsbescherming. PhD Dissertation, Tilburg University. https://www.academia.edu/35955556/Geautomatiseerde_ketenbesluiten_and_rechtsbescherming_Automated_Administrative_Chain_Decisions_and_Legal_Protection

Van Eck, M., Bovens, M., & Zouridis, S. (2018). Algoritmische rechtstoepassing in de democratische rechtsstaat. *NEDERLANDS JURISTENBLAD*, *40*, 3008–3017. https://scholarlypublications.universiteitleiden.nl/access/item%3A2978354/view

Weiner, I.B. (1986). Conceptual and Empirical-Perspectives on the Rorschach Assessment of Psychopathology. Journal of Personality Assessment, 50 (3), 472-479.

Wong, W., and E. Welch. 2004. "Does E-Government Promote Accountability? A Comparative Analysis of Website Openness and Government Accountability." Governance 17 (2): 275–297. doi:10.1111/j.1468-0491.2004.00246.x.

Whittington, R., & Yakis-Douglas, B. (2020). The grand challenge of corporate control: Opening strategy to the normative pressures of networked professionals. Organization Theory, 1. https:// doi.org/10.1177/2631787720969697

Wortham, R. H., Theodorou, A., & Bryson, J. J. (2017). Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 1–9. https://doi.org/10.1109/roman.2017.8172491

Young, M., Katell, M., & Krafft, P. M. (2019). Municipal surveillance regulation and algorithmic accountability. *Big Data & Society*, *6*(2), 205395171986849. https://doi.org/10.1177/2053951719868492

Zarsky T (2016) The trouble with algorithmic decisions: an analytic road map to examine efciency and fairness in automated and opaque decision making. Sci Technol Human Values 41(1):118–132

Zerilli J, Knott A, Maclaurin J, Gavaghan C (2018) Transparency in algorithmic and human decision making: is there a double standard? Philos Technol 32(4):661–683

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. Minds and Machines, 29(4), 555-578.

7. Bijlagen

Bijlage 7.1 Vignetten

Vignette van scenario 1- Lage mate van transparantie

MOGELIJKE GEVAARS CATEGORIEËN

- 2 REGISTRATIES MOGELIJK GERELATEERD AAN 'VERZET'
- 3 REGISTRATIES MOGELIJK GERELATEERD AAN 'VLUCHTGEVAARLIJK'
- 5 REGISTRATIES MOGELIJK GERELATEERD AAN 'DRUGSGEBRUIK'

Vignette van scenario 1- Hoge mate van transparantie

MOGELIJKE GEVAARS CATEGORIEËN

2 REGISTRATIES MOGELIJK GERELATEERD AAN 'VERZET'

*20-05-2019: '**WERKT NIET MEE** MET OMDOEN VAN HANDBOEIEN…'*
EN NOG 1 MELDING…

3 REGISTRATIES MOGELIJK GERELATEERD AAN 'VLUCHTGEVAARLIJK'

*05-12-2001 'PERSOON **NEGEERT STOPTEKEN** VAN DE POLITIE IN RODE...'*
EN NOG 2 MELDINGEN…

5 REGISTRATIES MOGELIJK GERELATEERD AAN 'DRUGSGEBRUIK'

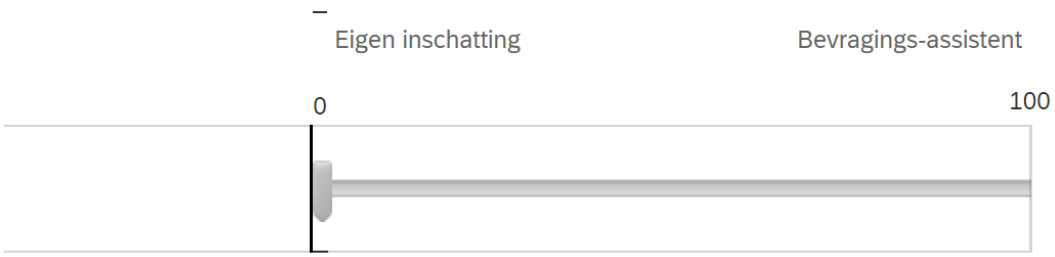
*16-03-2000 'MELDING MOGELIJK **DRUGS** VERKOOP DICHTBIJ DE SUPERMARKT'*
EN NOG 4 MELDINGEN…

Na de vignetten volgden de volgende twee vragen:

Op basis van de informatie van de bevragings-assistent, verwacht u dat Robin Jansen vluchtgevaarlijk is?

- ○ Ja, ik verwacht dat Robin Jansen vluchtgevaarlijk is
- ○ Nee, ik verwacht niet dat Robin Jansen vluchtgevaarlijk is

In hoeverre is uw antwoord gebaseerd op uw eigen inschatting en in hoeverre op de informatie van de bevragings-assistent?



Gebaseerd op het antwoord op de eerste vraag kreeg de respondent het onderstaande scherm te zien:

Bij scenario 1 met een lage mate van transparantie:

Uw inschatting blijkt juist te zijn. U dacht dat Robin vluchtgevaarlijk zou zijn en dat is ook zo. Vlak voor u arriveerde op de locatie van het incident blijkt Robin Jansen te zijn gevlucht naar een onbekende locatie.

Of

Uw inschatting blijkt onjuist te zijn. U dacht dat Robin niet vluchtgevaarlijk zou zijn maar het bleek van wel. Vlak voor u arriveerde op de locatie van het incident blijkt Robin Jansen te zijn gevlucht naar een onbekende locatie.

Of bij scenario 1 met een hoge mate van transparantie:

Uw inschatting blijkt juist te zijn. U dacht dat Robin vluchtgevaarlijk zou zijn en dat is ook zo. De bevragingsassistent had de woorden ‘negeert stopteken’ gevonden in een melding en heeft daarom de categorie ‘vluchtgevaarlijk’ laten zien. Vlak voor u arriveerde op de locatie van het incident blijkt Robin Jansen te zijn gevlucht naar een onbekende locatie.

Of

Uw inschatting blijkt onjuist te zijn. U dacht dat Robin niet vluchtgevaarlijk zou zijn maar het bleek van wel. De bevragingsassistent had de woorden ‘negeert stopteken’ gevonden in een melding en heeft daarom de categorie ‘vluchtgevaarlijk’ laten zien. Vlak voor u arriveerde op de locatie van het incident blijkt Robin Jansen te zijn gevlucht naar een onbekende locatie.

Vignette van scenario 2- Lage mate van transparantie

Vignette van scenario 2- Hoge mate van transparantie

MOGELIJKE GEVAARS CATEGORIEËN

1 REGISTRATIE MOGELIJK GERELATEERD AAN 'BEDREIGING MET WAPEN'

*03-11-2020 'VISSER WORDT BEDREIGD DOOR AGRESSIEVE VROUW MET **SCHROEVENDRAAIER**'*

8 REGISTRATIES MOGELIJK GERELATEERD AAN 'ALCOHOLGEBRUIK'

*05-12-2000 'PERSOON AANGEHOUDEN WEGENS **ALCOHOLGEBRUIK** IN PARK WAAR VERBOD GELDT'*
EN NOG 14 MELDINGEN…

10 REGISTRATIES MOGELIJK GERELATEERD AAN 'BEKEURINGEN'

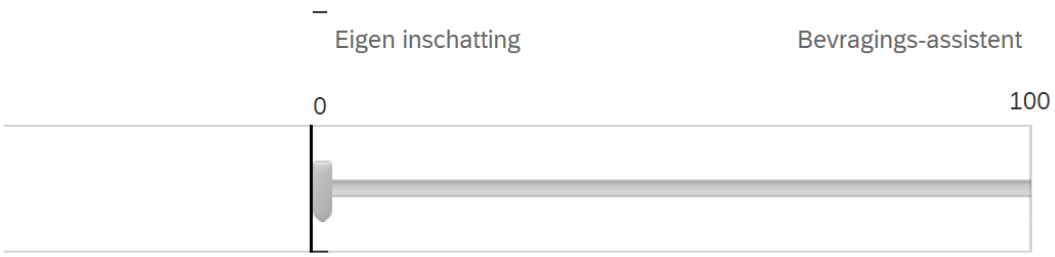
*16-03-2022 'PERSOON KRIJGT BEKEURING VOOR **DOOR ROOD LOPEN** BIJ VERKEERSLICHT…'.*
EN NOG 9 MELDINGEN'…

Na de vignetten volgden de volgende twee vragen:

Op basis van de informatie van de bevragings-assistent, verwacht u dat Renee Visser wapengevaarlijk is?

- ○ Ja, ik verwacht dat Renee Visser wapengevaarlijk is
- ○ Nee, ik verwacht niet dat Renee Visser wapengevaarlijk is

In hoeverre is uw antwoord gebaseerd op uw eigen inschatting en in hoeverre op de informatie van de bevragings-assistent?



Gebaseerd op het antwoord op de eerste vraag kreeg de respondent het onderstaande scherm te zien:

Bij scenario 2 met een lage mate van transparantie:

Uw inschatting blijkt onjuist. Je verwachtte dat Renee Visser wapengevaarlijk is, maar eenmaal op locatie aangekomen bleek die dat niet te zijn.

Of

Je inschatting blijkt juist te zijn. Renee Visser was inderdaad niet wapengevaarlijk.

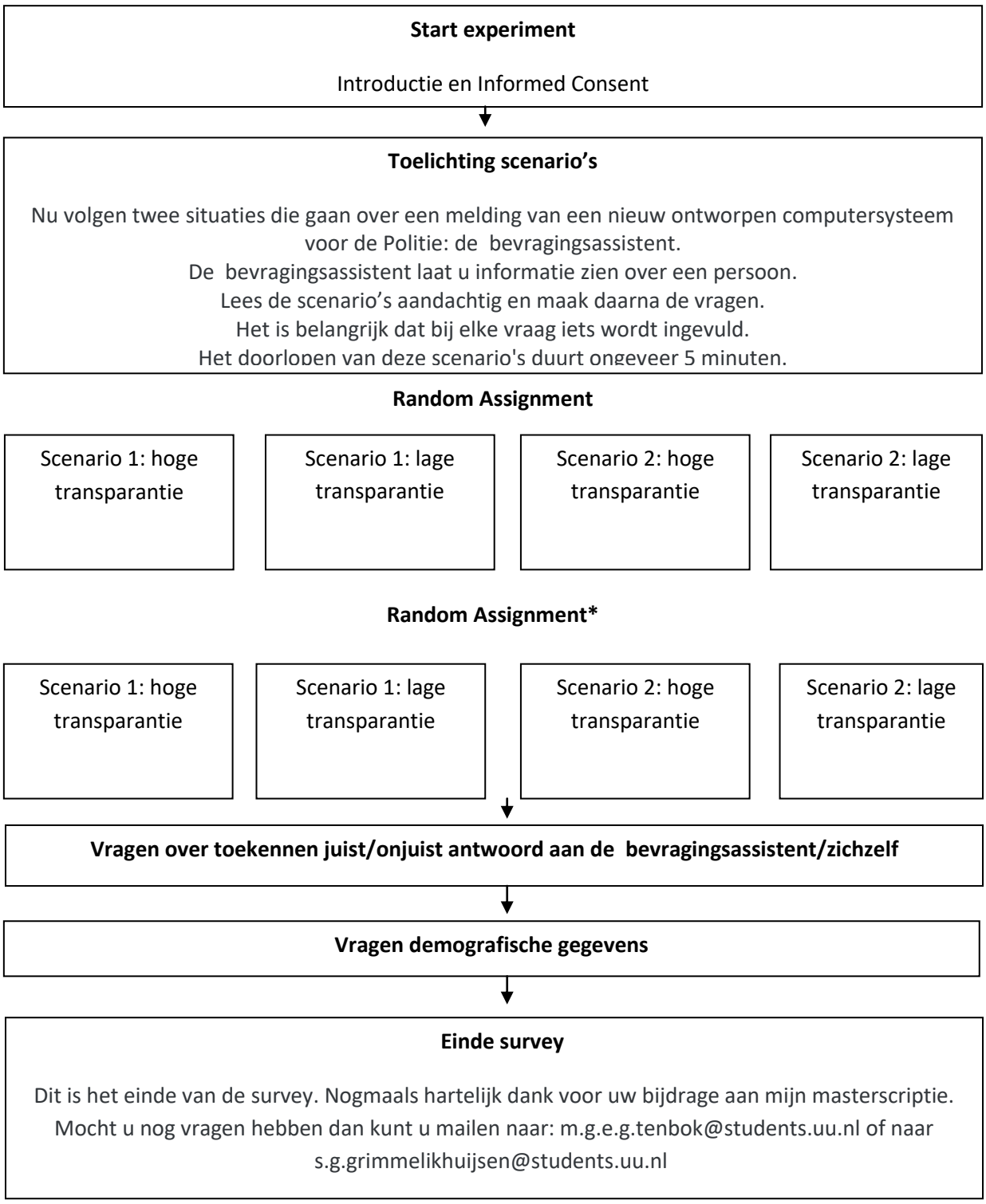
Bij scenario 2 met een hoge mate van transparantie:

Je inschatting blijkt onjuist. Je verwachtte dat Renee Visser wapengevaarlijk is, maar eenmaal op locatie aangekomen bleek die dat niet te zijn. De bevragingsassistent had de categorie ‘wapengevaarlijk’ laten zien doordat het systeem het woord ‘schroevendraaier’ in een melding had gevonden, maar Renee Visser bleek toen zelf de melder te zijn.

Of

Je inschatting blijkt juist te zijn. Renee Visser was inderdaad niet wapengevaarlijk. De bevragingsassistent had de categorie ‘wapengevaarlijk’ laten zien doordat het systeem het woord ‘schroevendraaier’ in een melding had gevonden, maar Renee Visser bleek toen zelf de melder te zijn.

Bijlage 7.2 Survey flow



*Iedere respondent krijgt scenario 1 en scenario 2 te zien. Je kunt niet twee keer achter elkaar scenario 1 of scenario 2 te zien krijgen. Je kunt bijvoorbeeld wel twee keer achter elkaar een scenario te zien krijgen met een hoge mate van transparantie.

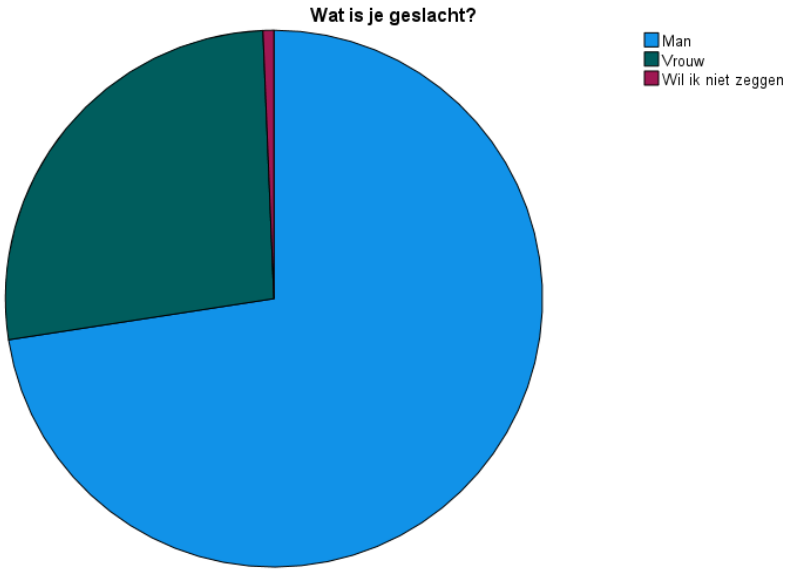
Bijlage 7.3 Demografische gegevens

Bijlage beschrijvende statistiek respondenten

Hieronder volgt een overzicht van de beschrijvende statistiek (*N* =152)

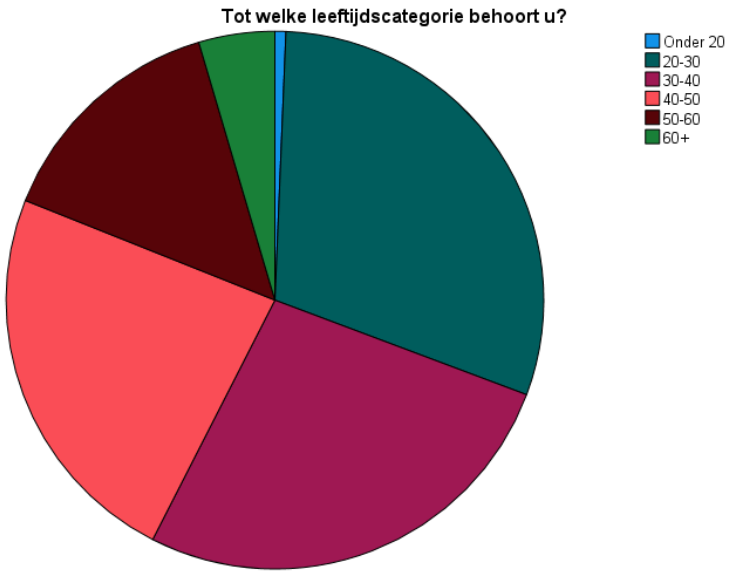
Tabel 7.1: Geslacht respondenten

Geslacht	Aantal	Percentage (%)
Man	111	72,5
Vrouw	41	26,8
Wil ik niet zeggen	1	0,7
Totaal	153	100



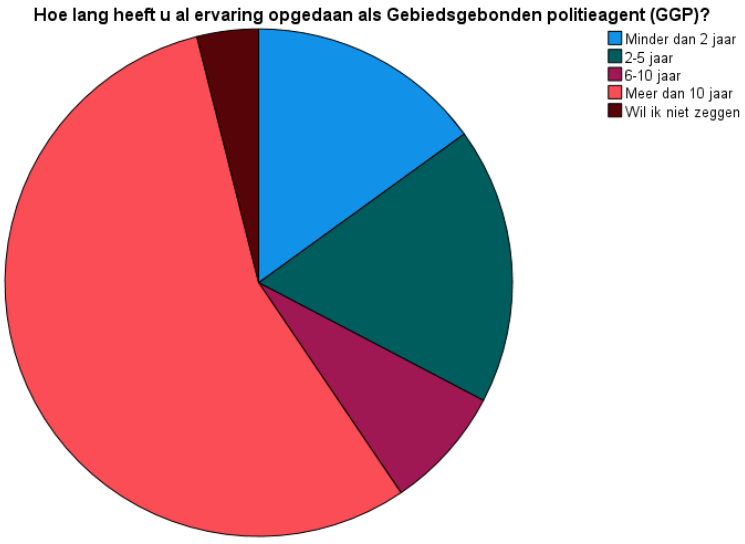
Tabel 7.2: Leeftijdscategorie respondenten

Leeftijdscategorie	Aantal	Percentage (%)
Onder 20	1	0,7
20-30	46	30,1
30-40	41	26,8
40-50	36	23,5
50-60	22	14,4
60+	7	4,6
Totaal	153	100



Tabel 7.3: Aantal jaar ervaring respondenten

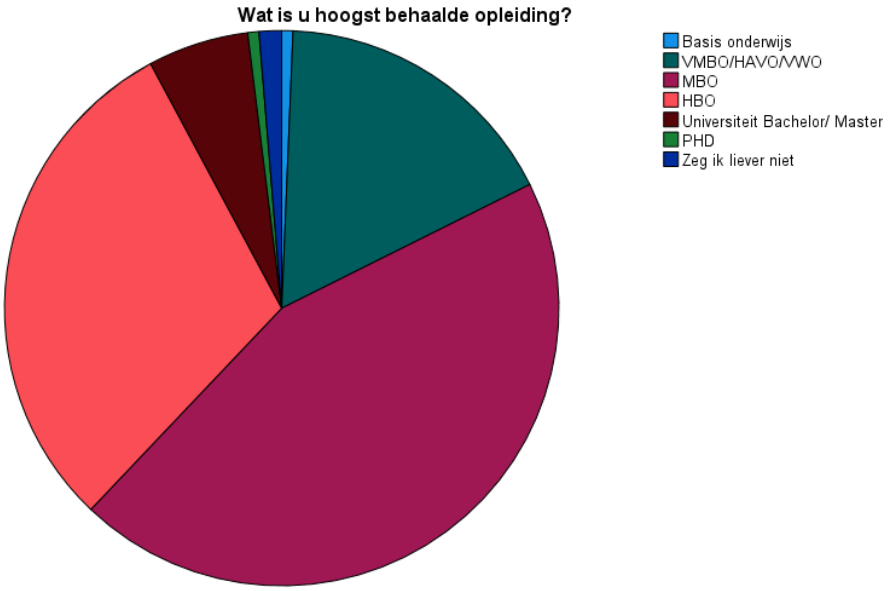
Jaren ervaring	Aantal	Percentage (%)
Minder dan 2 jaar	23	15
2-5 jaar	27	17,6
6-10 jaar	12	7,8
Meer dan 10 jaar	85	55,6
Wil ik niet zeggen	6	3,9
Totaal	153	100



Tabel 7.4: Opleidingsniveau respondenten

Opleidingsniveau	Aantal	Percentage (%)
Basis onderwijs	1	0,7

VMBO/HAVO/VWO	26	17
MBO	68	44,4
HBO	46	30,1
Universiteit Bechelor/Master	9	5,9
PHD	1	0,7
Zeg ik liever niet	2	1,3
Totaal	153	100



Bijlage 7.4 Randomisatiecheck

Om te controleren of de gevonden verschillen daadwerkelijk zijn toe te schrijven aan de behandeling uit het experiment, wordt gebruik gemaakt van een randomisatiecheck. Met de randomisatiecheck wordt gekeken of er significante verschillen zijn in demografische gegevens tussen de twee groepen.

Scenario 1

Variabele	Statistiek	Significante (*p*)
Geslacht	χ 2 = 1,087	0,58
Leeftijd	t= -0,527	0,30
Aantal jaar ervaring	t= -1,047	0,15
Opleidingsniveau	t= -0,229	0,41

Scenario 2

Variabele	Statistiek	Significante (*p*)
Geslacht	χ 2 = 1,127	0,57
Leeftijd	t= -0,173	0,43
Aantal jaar ervaring	t= -0,459	0,32
Opleidingsniveau	t= 0,021	0,49

In de tabellen van beide scenario's is te zien dat de controlegroep en de experimentele groep niet significant van elkaar verschillen. De randomisatiecheck in dit experiment is dus geslaagd.