Verantwoord beslissen

De rol van verantwoording bij gepast vertrouwen in AI-adviezen binnen de politie

Mitch Eddy Opperhuizen

6481302

Eerste begeleider	**Tweede begeleider**	**Praktische begeleider**
dr. Stephan Grimmelikhuijsen	dr. Merijn Bruijnes	dr. Bas Testerink
s.g.grimmelikhuijsen@uu.nl	m.bruijnes@uu.nl	b.j.g.testerink@uu.nl

Ingediend ter gedeeltelijk vervulling van de vereisten van een
Master of Science in Publiek Management aan de
Faculteit Rechten, Economie, Bestuurs- en organisatiewetenschap

Universiteit Utrecht
23 augustus 2024

Abstract

AI-systemen worden steeds vaker ingezet om contactambtenaren te ondersteunen bij complexe besluitvormingsprocessen. Het gebruik van AI-systemen in besluitvormingsprocessen biedt risico's, waardoor het nog steeds gebruikelijk is dat contactambtenaren het beslissingsproces overzien en het uiteindelijke besluit nemen. Echter, net als het systeem is de mens niet feilloos. Mensen tonen zowel te veel vertrouwen in AI-systemen als te weinig vertrouwen. Het is van belang dat contactambtenaren streven naar een ‘gepaste’ hoeveelheid vertrouwen in AI-systemen. In de context van politiewerk toetst dit onderzoek het effect van verantwoording afleggen op het bereiken van een gepast vertrouwen in AI-adviezen. Hiervoor is een surveyexperiment met een fictief AI-systeem uitgevoerd bij een steekproef van politieagenten (N=131). De resultaten van dit experiment laten zien dat: (1) gepast vertrouwen op AI-adviezen moeilijk te bereiken is, (2) verantwoording afleggen geen rol speelt in het voorkomen van te weinig vertrouwen in AI-adviezen, en (3) ervaring geen modererende factor is die de relatie tussen verantwoording afleggen en gepast vertrouwen versterkt. Deze resultaten hebben belangrijke implicaties voor het werk van contactambtenaren.

Inhoud

Abstract ..... **iii**

1. Inleiding **1**

1.1 Aanleiding 1

1.2 Probleemstelling 3

1.3 Deelvragen 6

1.4 Maatschappelijke relevantie 6

1.5 Wetenschappelijke relevantie 7

2. Theoretisch kader **8**

2.1 AI en discretie 8

2.2 Appropriate reliance 11

2.3 Verantwoording 15

2.3.1 Underreliance en de politie 15

2.3.2 Het afleggen van verantwoording 17

2.3.3 Ervaring en moderatie 20

3. Methode en metingen **22**

3.1 Experimenteel design 22

3.2 Materialen 23

3.2.1 Algemene opzet 24

3.2.2 Verantwoording: Controlegroepen vs. verantwoordingsgroepen 26

3.2.3 Appropriate reliance 27

3.2.4 Samengestelde schalen 30

3.3 Dataverzameling 31

3.4 Steekproef samenstelling 33

3.4.1 Representativiteit ten opzichte van de populatie 33

3.4.2 Randomisatie 34

3.5 Manipulatiecheck 35

3.6 Onderzoeksethiek 37

3.7 Kwaliteitscriteria: Validiteit en betrouwbaarheid 38

3.7.1 Validiteit 38

3.7.2 Betrouwbaarheid 40

4. Resultaten **41**

4.1 Beschrijvende statistiek 41

4.2 Effect van verantwoording op underreliance 43

4.3 Modererend effect van ervaring op de relatie 44
4.4 Exploratieve analyses 45
4.4.1 Tijd 45
4.4.2 Vertrouwen in technologie 46
4.4.3 De gegeven verantwoording 47
4.5 Samenvattende conclusie resultaten 50
5. Discussie 52
5.1 Beperkingen 54
5.2 Literatuur implicaties 57
5.2.1 AI-besluitvorming 57
5.2.2 Appropriate reliance en verantwoording 59
5.3 Implicaties Werkveld 61
5.4 Afsluiting 63
Bronnen 64
Bijlage 71
Bijlage A. Factor analyses 71
Bijlage B. Representativiteit toets 72
Bijlage C. Crosstabs hoofdvraag 74
Bijlage D. Crosstabs moderator 75
Bijlage E. Grafieken codering 76
Bijlage F. Enquête 78

1. Inleiding

1.1 Aanleiding

Stel je het volgende scenario voor: Een politieagent surveilleert tijdens een tot dan toe stille nacht binnen hun regio. Plotseling komt er een melding binnen op hun telefoon. Deze melding is gegenereerd door een geavanceerd AI-systeem dat vluchtroutes van misdadigers berekent. *"Inbraak op heterdaad betrapt in de winkelstraat. Eenheden worden opgeroepen om wegen af te zetten. Uw eenheid wordt verzocht om de ingang van de snelweg te blokkeren bij Locatie A."* De navigatie van de politieauto springt automatisch aan en lijkt de agent direct te leiden naar de locatie die door het AI-systeem wordt geadviseerd. Het systeem biedt zelfs inzicht in de routes die andere eenheden zullen nemen en welke wegen zij zullen afzetten. De agent beseft dat het AI-systeem een logische keuze heeft gemaakt die is gebaseerd op een grote hoeveelheid data over vluchtroutes. Desondanks blijft er een knagend onderbuikgevoel hangen. De agent kent de regio goed en weet dat een bepaalde binnendoorweg niet wordt afgezet met deze formatie, een route die de misdadigers een gemakkelijke ontsnapping zou kunnen bieden. Eveneens weet de agent dat hij of zij later verantwoording zal moeten afleggen aan hun leidinggevende over de beslissing die deze nacht wordt genomen. Wat is een gepaste hoeveelheid vertrouwen in AI-systemen en waarop vertrouwt de agent? Op de jaren aan verzamelde politiedata die door een AI-systeem is geanalyseerd, of op 15 jaar ervaring binnen de regio en persoonlijke intuïtie?

Het gebruik van algoritmes, waaronder Artificiële Intelligentie (AI), verandert en automatiseert het werk van ambtenaren. Een belangrijke taak van ambtenaren waar AI invloed op uitoefent is het nemen van besluiten en het proces van totstandkoming van die besluiten (Alon-Barkat & Busuioc, 2023; Binns, 2022; Bovens & Zouridis, 2002; Buffat, 2015; J. B. Bullock, 2019; de Boer & Raaphorst, 2023; Peeters, 2020). Een domein dat voorop loopt in de ontwikkeling van AI-systemen die ondersteunen bij besluitvormingsprocessen is de politie.

AI ontwikkelt snel, zo zien de Rijksoverheid en de wetenschap in die ontwikkeling kansen, maar ook risico's (Grimmelikhuijsen & Meijer, 2022;

Rijksoverheid, 2024; Tsamados et al., 2022). Allereerst bieden AI-systemen kansen binnen beslissingsprocessen. AI-systemen zijn goed en snel in het herkennen van patronen. Een voorspelling van een AI-systeem is niet gebaseerd op een enkel oordeel of casus, maar berekent in hoeverre een casus kenmerken deelt met een bredere groep (Hannah-Moffat, 2013; Van Eijk, 2020). Dit zorgt ervoor dat AI-systemen de mogelijkheid bieden om het besluitvormingsproces te verbeteren rondom efficiëntie, effectiviteit, objectiviteit, consistentie en nauwkeurigheid (Alon-Barkat & Busuioc, 2023; Binns, 2022; Bullock, 2019; Peeters, 2020).

Echter, aan de toepassing van AI-systemen zit ook een keerzijde aan. Het gebruik van AI kent een scala aan risico's. In de volgende alinea's worden voorbeelden van risico's benoemd.

Ten eerste, AI-systemen bieden weinig transparantie in de beslissingen die zij maken. De interne databases die de systemen gebruiken om te komen tot beslissingen zijn complex, lastig te interpreteren en kunnen zelfs veranderen (Burrell, 2016; Giest & Grimmelikhuijsen, 2020; Grimmelikhuijsen & Meijer, 2022; Tsamados et al., 2022).

Ten tweede, AI biedt weinig ruimte voor maatwerk voor individuele cases. Voorbeelden laten zien dat patroonherkenning kan leiden tot situaties met ernstige gevolgen voor groepen burgers, waaronder discriminatie, zoals gebeurde bij de toeslagenaffaire van de Belastingdienst (Binns, 2022; Peeters, 2020; Peeters & Widlak, 2023).

Ten derde, AI-systemen bestaan uit kansberekeningen en in kansberekeningen kunnen fouten voorkomen. Daarom is het in veel gevallen wenselijk dat AI-systemen gebruik maken van menselijk overzicht en inzicht die deze fouten kunnen verbeteren (Burrell, 2016; Peeters, 2020).

Als laatste, bij toepassing van AI-systemen geldt het ontstaan van verschillende vormen van bias bij algoritmische besluitvorming. Hieronder vallen *automation bias* en *selective adherence* (Alon-Barkat & Busuioc, 2023; Meijer et al., 2021; Selten et al., 2023). *Automation bias* leidt tot het altijd opvolgen van AI-

systeem adviezen en *selective adherence* is het alleen opvolgen van adviezen wanneer die het eigen inzicht bevestigen.

Al met al biedt AI binnen beslissingsprocessen kansen, maar ook uitdagingen. Zo is het van belang dat ambtenaren die besluiten maken op basis van AI-adviezen niet blindelings vertrouwen hebben in de systemen, gezien de uitdagingen. Eveneens is het van belang dat ze niet nooit op AI-systemen vertrouwen, gezien de kansen. Daarom richt dit onderzoek zich op het concept *appropriate reliance* (Benda et al., 2022; Lee & See, 2004; Schemmer et al., 2022, 2023).

1.2 Probleemstelling

Appropriate reliance draait om het hebben van de 'gepaste' hoeveelheid vertrouwen in en afhankelijkheid van AI-systemen en is daarmee een alternatief voor gewoonweg vertrouwen in het gebruik van AI (Benda et al., 2022). Er is veel onderzoek gedaan naar vertrouwen in data gedreven systemen, waaronder AI, maar nog weinig onderzoek naar *appropriate reliance* (Benda et al., 2022; Dzindolet et al., 2003; Schemmer et al., 2022; Selten et al., 2023). Het doel bij *appropriate reliance* is niet het behalen van zoveel mogelijk vertrouwen, maar een 'gepaste' hoeveelheid vertrouwen. Onderzoek dat gedaan wordt naar *appropriate reliance* is nog generiek, met steekproeven die bestaan uit studenten of niet praktijkrelevante casussen. Daarmee is het onderzoeken van *appropriate reliance* binnen een realistische en belangrijke maatschappelijke context zoals het politiedomein erg interessant.

Appropriate reliance is belangrijk, want de huidige wetenschap gaat in basis ervan uit dat zowel ambtenaren als AI-systemen fouten kunnen maken in hun beslisvermogen (Schemmer et al., 2022). Vertrouwen in AI is belangrijk en wenselijk wanneer een systeem een, naar menselijke en door een organisatie vastgestelde maatstaven, goed advies geeft. Tegelijkertijd is het niet wenselijk dat een AI-advies wordt opgevolgd wanneer het een onjuist advies geeft (Schemmer et al., 2023). *Appropriate reliance* kan gedefinieerd worden als het tonen van vertrouwen in AI wanneer het een goed advies geeft en het afwijken van AI wanneer het een fout advies geeft. De ambtenaar toont een gepaste hoeveelheid vertrouwen in het

systeem, zodat ze elkaar aan kunnen vullen om te komen tot de best mogelijke beslissingen.

Het creëren van *appropriate reliance* is niet eenvoudig, want het is sterk contextafhankelijk. In huidig onderzoek wordt getoetst voor *appropriate reliance*, maar worden contextfactoren nog genegeerd (Benda et al., 2022; Lee & See, 2004; Schemmer et al., 2023). Het vertrouwen op AI-advies verschilt bijvoorbeeld sterk door de organisatiecontext. Duitse politieagenten zijn sneller geneigd deze adviezen op te volgen en zien het als onderdeel van de bureaucratie en hiërarchie. Terwijl Nederlandse agenten de AI-systemen zien als een collega die een suggestie geeft (Meijer et al., 2021). Net als de relatie tussen de beslisruimte van contactambtenaren en AI-adviezen wordt *appropriate reliance* beïnvloed door diverse contexten, waaronder de aard van de technologie (Peeters, 2020), de context van de organisatie (J. Bullock et al., 2020; Giest & Grimmelikhuijsen, 2020; Grimmelikhuijsen & Meijer, 2022; Meijer et al., 2021; Peeters & Widlak, 2023), de aard van de taak (J. Bullock et al., 2020; J. B. Bullock, 2019; de Boer & Raaphorst, 2023; Young et al., 2019) en de aard van de individuele ambtenaar (Alon-Barkat & Busuioc, 2023; Binns, 2022; Selten et al., 2023). Aangezien *appropriate reliance* nog niet diepgaand is onderzocht en toegepast, is niet bekend welke van deze contextfactoren bijdragen aan het creëren van de gepaste hoeveelheid vertrouwen.

De contextfactor die centraal staat in dit onderzoek is de mens en hun verwachtingen rondom het besluitvormingsproces. Een contextfactor die voorheen licht is onderzocht is *explainable AI* ofwel XAI. De verwachting is dat een beter begrijpelijk systeem kan helpen om de juiste hoeveelheid vertrouwen te creëren (Schemmer et al., 2023). XAI is een verklarend systeem dat onderliggende factoren voor de besluitvorming onthult en daarover uitleg geeft aan een gebruiker van dat systeem (Miller, 2019b). XAI helpt in beperkte mate het tegengaan van te weinig vertrouwen, maar niet tegen het vertonen van te veel vertrouwen (Schemmer et al., 2023). Bovendien volgen ambtenaren vooral hun eigen instinct volgen en vertrouwen AI voornamelijk als het de eigen bias bevestigt (Selten et al., 2023).

Een bewezen factor die helpt tegen het vertonen van bias in besluitvormingsprocessen uit de psychologie literatuur, is het afleggen van

verantwoording rondom de keuze van de ambtenaar (Aleksovska et al., 2019; Tetlock, 1983). Ambtenaren denken langer na over hun keuzes en vertonen ook minder gedragingen zoals *overconfidence*. In tegenstelling tot XAI, waarin systemen hun besluitvormingsprocessen moeten onthullen, wordt in dit onderzoek het effect van verantwoording afleggen op *appropriate reliance* onderzocht. In plaats van naar het systeem te kijken, moet in dit onderzoek de mens zijn besluitvormingsproces verantwoorden.

Het domein waarin dit onderzoek plaatsvindt is dat van de politie. De politie heeft sterk te maken met de kansen als de risico's bij het gebruik van AI-systemen (Digitale Overheid, 2023b). Ethische thema's zijn belangrijk binnen deze context en besluiten kunnen zwaarwegende consequenties hebben. Dat maakt het onderzoeken van verantwoording en *appropriate reliance* rondom AI-systemen interessant. Bovendien is het afleggen van verantwoording in de context van AI-keuzeprocessen nog weinig onderzocht. Tegelijkertijd is het afleggen van verantwoording een belangrijk thema binnen de AI-literatuur, door de ambiguïteit van verantwoordelijkheid bij computersystemen. Het is veelal onduidelijk bij wie de verantwoordelijkheid komt te liggen zodra fouten gemaakt worden in het handelen met algoritmes, waaronder AI-systemen, ook bij de politie (Wieringa, 2020). Volgens Wieringa (2020) bestaat algoritmische verantwoordelijkheid uit een netwerk van meerdere actoren die verantwoording moeten afleggen, waaronder de besluitnemers zelf. Dit onderzoek kijkt naar het effect van de verantwoordelijkheid op *appropriate reliance* bij AI-besluitvormingsprocessen. Daarin staat de volgende vraag centraal:

Wat is het effect van het afleggen van verantwoording door politieagenten op gepast vertrouwen in AI-systemen bij besluitvormingsprocessen met AI-adviezen?

Het doel is daarmee om bij te dragen aan betere besluitvorming met AI-systemen binnen het politiewerk. Dit door het creëren van randvoorwaarden (zoals verantwoording afleggen) waardoor ambtenaren de juiste hoeveelheid vertrouwen in AI aanbevelingen hebben.

1.3 Deelvragen

Dit onderzoek focust op het effect van verantwoording bij besluitvorming met AI-systemen op *appropriate reliance* in de context van de politie. Hierdoor draagt het onderzoek bij aan het ontdekken van factoren die *appropriate reliance* beïnvloeden bij contactambtenaren.

Het beantwoorden van deze vraag en het op stellen van hypotheses vergt de beantwoording van een aantal deelvragen. De eerste deelvraag is: *wat is AI en hoe wordt dit toegepast binnen publieke organisaties?* Hierbij wordt gekeken naar de verschillende kenmerken van AI en hoe artificiële discretie vorm heeft gekregen. Dit is mede van belang om het begrip *'appropriate reliance'* te begrijpen. De tweede deelvraag is vervolgens *wat is appropriate reliance en hoe kan dit geoperationaliseerd worden in het politiedomein? Appropriate reliance* is een gelaagd concept dat ontstaan is vanuit onderzoek naar vertrouwen in AI in combinatie met de beslisruimte van ambtenaren. Daarom is het van belang om een goede definitie vast te leggen in de context van het experiment. De derde deelvraag luidt, *wat wordt verstaan onder het afleggen van verantwoording in de context van AI aanbevelingen?* Om het effect van verantwoording goed te toetsen is een definitie nodig, passend in deze context.

1.4 Maatschappelijke relevantie

De digitaliseringsstroom binnen de overheid biedt kansen en risico's. Dit onderzoek draagt bij aan het identificeren van risico's en het omzetten in kansen door inzicht te bieden in hoe (politie)ambtenaren beter kunnen samenwerken met AI applicaties. Specifiek door te kijken naar wat de 'gepaste hoeveelheid vertrouwen is in AI-adviezen en wat bijdraagt aan creëren van deze 'gepaste hoeveelheid vertrouwen (Benda et al., 2022; Lee & See, 2004; Schemmer et al., 2022). Dit helpt het voorkomen van blindelings vertrouwen in AI-adviezen en overtollige aversie tegen AI-adviezen. Blindelings vertrouwen kan leiden tot gevaarlijke situaties, zoals de toeslagenaffaire, waarin algoritmes een rol speelde in het verkeerd 'labelen' van mensen als fraudeurs. Hierdoor leek het alsof de menselijke maat verdwenen was (Chavannes, 2020; Peeters & Widlak, 2023). Tegelijkertijd is het belangrijk om de

kansen en mogelijkheden van AI-systemen te blijven benutten, zoals de potentiële effectiviteit, efficiëntie en betere besluiten die het belooft (Peeters, 2020; Young et al., 2019). Door gebruik te maken van de 'juiste' hoeveelheid vertrouwen hoopt dit onderzoek bij te dragen aan het voorkomen van onnodige aversie tegen AI. Zo wordt een beter AI-besluitvormingsproces bij de politie gecreëerd.

Specifiek brengt dit onderzoek in kaart of de implementatie van verantwoordingsstructuren bijdragen aan het creëren van een gepaste hoeveelheid vertrouwen in AI-adviezen. Bovendien biedt dit onderzoek inzicht in de huidige staat van *appropriate reliance* bij de politie.

1.5 Wetenschappelijke relevantie

Dit onderzoek levert een bijdrage aan de publiek management literatuur over digitalisering en besluitvorming met AI-systemen (Binns, 2022; J. Bullock et al., 2020; Eck et al., 2018; Meijer et al., 2021; Peeters, 2020; Young et al., 2019). Specifiek wordt een bijdrage geleverd rondom besluitvorming van contactambtenaren (nog specifieker de politie) en eventuele bias die zij laten zien (de Boer & Raaphorst, 2023; Grimmelikhuijsen & Meijer, 2022; Meijer et al., 2021; Selten et al., 2023). Daarbij levert dit onderzoek een bijdrage aan de summiere *appropriate reliance* literatuur en onderzoekt mogelijke contexten hoe *appropriate reliance* bereikt kan worden (Benda et al., 2022; Lee & See, 2004; Schemmer et al., 2022). De context die onderzocht wordt is het geven van verantwoording bij besluiten met AI-systemen in relatie tot *felt accountability*, ofwel gevoelde verantwoording door de besluitenmakers (Aleksovska et al., 2019; Hochwarter et al., 2007; Overman et al., 2021; Overman & Schillemans, 2022; Schillemans et al., 2021; Wieringa, 2020). Het onderzoek probeert hiermee een bijdrage te leveren aan de verantwoordingsliteratuur door in kaart te brengen of contactambtenaren 'gepaster' vertrouwen tonen in AI-systemen door verantwoording af te leggen.

2. Theoretisch kader

In het onderstaande theoretische kader worden de theoretische deelvragen beantwoord. Ten eerste wordt ingegaan op AI en de toepassing daarvan binnen publieke organisaties, waaronder de manier waarop AI en de beslisruimte van een contactambtenaar met elkaar in verband staan. Ten tweede wordt stilgestaan bij het concept *appropriate reliance* en hoe dit geoperationaliseerd kan worden in een politie domein. Als laatste element wordt onderzocht wat verantwoording afleggen inhoudt in de context van besluitvormingsprocessen met AI. Dit wordt gerelateerd aan de mogelijke rol die verantwoording kan spelen om te komen tot *appropriate reliance*.

2.1 AI en discretie

Wat is AI? AI kan worden gedefinieerd als: *"de groepering van technologieën die gegevens, algoritmes en hardware gebruiken om relaties te identificeren, voorspellingen te doen of semi complexe taken uit te voeren. AI bezit of vertoont menselijke intelligentie of is in staat om problemen rationeel op te lossen om hun doelen te bereiken"* (Young et al., 2019). Om beter te begrijpen wat dit inhoudt, moet de term algoritme worden gedefinieerd. Een AI-systeem bestaat namelijk uit vele algoritmes die gezamenlijk zelfs nieuwe algoritmes kunnen creëren. Een algoritme is een set met regels die een probleem oplost of doel bereikt (Grimmelikhuijsen, 2023). Een recept, een routebeschrijving en een wiskundesom zijn allemaal algoritmes. In het kader van dit onderzoek wordt gekeken naar computersystemen bestaande uit algoritmes die gezamenlijk intelligent genoeg zijn om aanbevelingen te geven bij besluitvormingsprocessen (Digitale Overheid, 2023a). Volgens de definitie van Young et al. (2019) wordt dit beschouwd als AI.

Een belangrijke toevoeging op de definitie van Young et al. (2019) is dat de rationaliteit een aanname is. AI wordt verwacht, in tegenstelling tot mensen 'neutraal' te zijn en 'feilloze beslissingen' te kunnen maken door grote hoeveelheden data te analyseren (Peeters, 2020). Echter, AI is niet neutraal en niet feilloos, zelfs al konden organisaties een overeenstemming bereiken wat de term neutraal inhoudt (Bannister & Connolly, 2020; Eck et al., 2018; Peeters & Widlak, 2023). AI-systemen

worden gemaakt door mensen, dat houdt in dat de vertoonde biasen, mogelijke fouten en waardeoordelen, niet meer gemaakt worden bij een keuze van een ambtenaar, maar ingebed zijn in de codering van het systeem (Bannister & Connolly, 2020). Dit verplaatst de mogelijkheid tot fouten maken van een individuele ambtenaar naar systeem design fouten (Eck et al., 2018; Peeters & Widlak, 2023). Dit roept de vraag op waarom worden AI-systemen dan gebruikt als ze niet compleet rationeel zijn? AI biedt namelijk een alternatief voor pure menselijke discretionaire ruimte (Eck et al., 2018; Meijer et al., 2021; Young et al., 2019).

Om het belang hiervan te begrijpen wordt eerst stilgestaan bij wat de discretionaire ruimte van ambtenaren inhoudt. AI automatiseert een deel van deze bureaucratische besluitvormingsprocessen en beïnvloedt daarmee de relatie van de ambtenaar en zijn discretionaire ruimte (Young et al., 2019). Specifieker gezegd: politieagenten functioneren als contactambtenaren (Noordegraaf, 2015). De kerntaak van contactambtenaren is het maken van beslissingen over andere mensen in bureaucratische besluitvormingsprocessen. Zij zijn de uitvoerders van wet- en regelgeving (Lipsky, 2010). Geschreven regels zijn niet altijd perfect afgestemd voor elke situatie en daarom maken contactambtenaren voortdurend gebruik van hun discretionaire ruimte (Lipsky, 2010). Discretionaire ruimte is de vrijheid, met gegeven beperkingen, die contactambtenaren hebben om te bepalen hoe zij wetgeving implementeren in hun besluitproces (Bovens & Zouridis, 2002; de Boer & Raaphorst, 2023; Evans, 2011; Thomann et al., 2018; Tummers & Bekkers, 2014).

Het gebruik van discretionaire ruimte kan zowel positieve als negatieve gevolgen hebben. Veelal gebruiken contactambtenaren discretionaire ruimte voor de goede zaak. Contactambtenaren kunnen met hun discretionaire ruimte effectief gebruik maken van overheidsmiddelen, terwijl zij publieke waarde proberen te creëren en compassie kunnen tonen (J. B. Bullock, 2019; Evans & Hupe, 2019; Maynard-Moody & Musheno, 2000; Moore, 2013). Tegelijkertijd zijn mensen imperfecte wezens. De menselijke besluitmakingsprocessen zitten vol met bias die leidt tot oneerlijke of zelfs racistische uitkomsten, corrupte gedragingen of cognitieve beperkingen (J. B. Bullock, 2019; Evans & Hupe, 2019; Maynard-Moody & Musheno, 2000; Pierson et al., 2020; Selten et al., 2023). Regels kunnen anders

geïnterpreteerd worden door de contactambtenaren. Dit leidt tot ruis tussen de beleidsvormers en de uitvoerders van het beleid (Maynard-Moody & Musheno, 2000).

AI biedt een alternatief voor pure menselijke discretionaire ruimte (Eck et al., 2018; Meijer et al., 2021; Young et al., 2019). De discretionaire ruimte van mensen kan worden overgenomen of aangetast door besluitvormingsprocessen waarin AI een adviserende rol speelt. Young et al. (2019) omschrijven dit als artificiële discretie. Artificiële discretie hoort de negatieve gevolgen van menselijke discretie te beperken, door consistente en op data gebaseerde beslissingen te maken. Tegelijkertijd, zoals eerder is geconstateerd, is AI-besluitvorming eveneens niet feilloos. Bovendien neemt AI menselijke discretie niet volledig over zoals voorheen werd gedacht (Bovens & Zouridis, 2002; Buffat, 2015).

Recente bevindingen laten zien dat AI een tool kan zijn om de discretionaire ruimte van contactambtenaren op een positieve manier te vergroten: dit noemt men de *enablement thesis* (Buffat, 2015; de Boer & Raaphorst, 2023; Meijer et al., 2021; Peeters, 2020; Selten et al., 2023). De huidige literatuur en realiteit laten zien dat contactambtenaren veelal '*in the loop*' zitten bij AI-besluitvorming (Meijer et al., 2021; Peeters, 2020; Selten et al., 2023). *In the loop*, of op de hoogte van, houdt in dat ambtenaren de besluitvorming overzien en de gemaakte keuzes van AI-systemen kunnen tegenhouden (Peeters, 2020). Binnen het Nederlandse politiedomein heeft AI een adviserende rol, de keuzebevoegdheid ligt bij de agenten zelf (Meijer et al., 2021). Dit is in gesprekken bevestigd door agenten en ontwikkelaars bij het Nationaal Politielab AI. In dit onderzoek is het daarom relevant dat het besluitvormingsproces gepresenteerd wordt als samenspel tussen de contactambtenaar en het AI-systeem.

Samenvattend wordt de eerste deelvraag als volgt beantwoord en de volgende definitie van AI gehanteerd: De groepering van technologieën die gegevens, algoritmes en hardware gebruikt om relaties te identificeren, voorspellingen te doen of semi complexe taken uit te voeren. AI bezit of vertoont menselijke intelligentie en is in staat om problemen tot zover hun systeem dat toe laat rationeel op te lossen om hun doelen te bereiken. (Bannister & Connolly, 2020; Young et al., 2019). De

belangrijkste toevoeging is dat AI en mensen niet feilloos zijn in hun besluitvorming. Daarom vullen AI-systemen aan op de discretionaire ruimte van de ambtenaar in een proces dat artificiële discretie wordt genoemd. (Buffat, 2015; Peeters, 2020; Young et al., 2019). Daarbij is die mogelijkheid tot fouten de reden waarom contactambtenaren niet blind moeten vertrouwen op AI, maar waarom ontworpen moet worden voor *appropriate reliance* (Benda et al., 2022; Lee & See, 2004).

2.2 Appropriate reliance

Onderstaand wordt stilgestaan bij het concept *appropriate reliance* en hoe dit geoperationaliseerd kan worden binnen het politie domein. Daarbij wordt voortgebouwd op de conclusies rondom artificiële discretie.

Het samenspel tussen de contactambtenaar en AI-systemen bij besluitvorming is niet goed uitvoerbaar als ambtenaren geen vertrouwen hebben in de adviezen van AI (Dzindolet et al., 2003; Schemmer et al., 2022; Selten et al., 2023). Er zijn verschillende redenen waarom contactambtenaren in bepaalde gevallen weinig vertrouwen vertonen in de adviezen van AI. In de eerste plaats, kan dat aan de aard en opzet van het systeem liggen. AI-systemen zijn over het algemeen een ‘black box’ die weinig transparantie bieden aan de gebruiker (Burrell, 2016; Giest & Grimmelikhuijsen, 2020; Miller, 2019a). Ten tweede kan het liggen aan sociologische factoren zoals de aard van de persoon of de organisatorische context. Denk hierbij aan de inherente bias die contactambtenaren hebben bij het gebruik van AI (Meijer et al., 2021; Selten et al., 2023). Sommige ambtenaren zullen positievere of negatievere ervaringen hebben met AI. Als laatste kunnen contactambtenaren sceptisch zijn om AI-systemen te vertrouwen uit ethische overwegingen. Nieuwsberichtgeving rondom ingebouwde bias of discriminerende AI-systemen zijn redenen om AI te wantrouwen (Peeters & Widlak, 2023; Tsamados et al., 2022). Veel onderzoek focust daarom op het vergroten van vertrouwen in AI. Hiervoor worden concepten zoals XAI onderzocht die beloven transparantere en beter interpreteerbare systemen te kunnen bouwen, waardoor vertrouwen toeneemt (Miller, 2019b). Dit onderzoek stelt echter dat ondanks dat vertrouwen een

belangrijke rol speelt, het vertrouwen niet het einddoel hoort te zijn in besluitvorming met AI-systemen.

Het einddoel is dus niet het verkrijgen van zoveel mogelijk vertrouwen in AI. Het einddoel is het behalen van een 'juiste' hoeveelheid vertrouwen, ofwel *appropriate reliance* (Benda et al., 2022; Bruijnes et al., 2024; Schemmer et al., 2023). Dit moet leiden tot evenwichtige en betere besluitvorming. Zoals eerder is geconcludeerd, AI-systemen zijn niet feilloos en maken fouten, maar kunnen eveneens helpen om tot efficiëntere en/of betere beslissingen te komen (Bannister & Connolly, 2020; Burrell, 2016; Tsamados et al., 2022). Wanneer een AI-systeem een fout maakt, is het wenselijk dat de contactambtenaar een andere keuze maakt dan het systeem. Het is tegelijkertijd wenselijk dat een contactambtenaar een goed advies opvolgt (Schemmer et al., 2022).

Ter illustratie, achteraf gezien was het onwenselijk dat ambtenaren van de belastingdienst veel van de door algoritmes gelabelde fraudeurs, daadwerkelijk labelden als fraudeurs. Tijdens de toeslagenaffaire kwam aan het licht dat de algoritmes het vaak fout hadden en ook geen rekening konden houden met persoonlijke omstandigheden van de zogenaamde fraudeurs (Peeters & Widlak, 2023). Tegelijkertijd is het wenselijk om fraudeurs te kunnen herkennen. Dit bleek ook uit de politieke keuzes die volgden na de Bulgaren fraude die leidden tot strengere controles binnen het toeslagenstelsel (Peeters & Widlak, 2023). Wanneer een algoritme terechte redenen heeft om iemand als fraudeur te bestempelen is het wenselijk om hier onderzoek naar te doen. Een contactambtenaar moet genoeg vertrouwen hebben om het AI-advies goed te kunnen interpreteren. Vanuit die positie dient de contactambtenaar het advies weloverwogen over te nemen, ten delen over te nemen of weloverwogen ter discussie te stellen en eventueel besluiten om niet over te nemen (Schemmer et al., 2022).

Om *appropriate reliance* te begrijpen is het van belang om beide concepten, *appropriate* en *reliance* apart van elkaar te begrijpen. Daarbij is het van belang om te benoemen dat *appropriate reliance* een nieuw concept is binnen de publiek management literatuur waar weinig over geschreven is (Bruijnes et al., 2024). Daarvoor bij de conceptdefiniëring ook uit oudere literatuur geput. Dit omdat de

literatuur rondom het concept schaars is. Ten eerste: wat is *reliance*? *Reliance* verschilt van vertrouwen. Vertrouwen beïnvloedt gedrag, maar leidt niet in alle gevallen tot gedrag. Wellicht vertrouw je een vriend, maar volg je zijn advies niet altijd op. In dat geval leidt vertrouwen niet tot gedragsaanpassing. Reliance daarentegen is gedrag en kan dus geobserveerd worden (Dzindolet et al., 2003; Lee & See, 2004). Scharowski et al. (2022) definiëren *reliance* daarmee als "*gedrag dat volgt na advies vanuit het systeem*".

Ten tweede wat houdt *appropriate*, of juist in? *Appropriate* is de mate waarin iets gepast of juist is. In de context van beslissingen en de juiste hoeveelheid vertrouwen hebben in adviezen van AI-systemen, zet de 'gepaste' hoeveelheid vertrouwen zich af tegen twee andere vormen. Daarmee bestaan de drie vormen: *Underreliance*, *overreliance* en *appropriate reliance* (Schemmer et al., 2022).

Underreliance houdt in dat een contactambtenaar nagenoeg geen rekening houdt met het resultaat van het AI-systeem. De contactambtenaar is sceptisch ten aanzien van AI-advies en volgt alleen AI-advies op als het overeenkomt met de eigen bias.

Overreliance is het altijd volgen van AI-adviezen. Een voorbeeld hiervan is te vinden in Meijer et al. (2021) waarin de Duitse politieagenten algoritmische adviezen zien als onderdeel van het bureaucratische systeem en de besluiten van die algoritmes produceren bijna altijd opvolgen, zelfs bij incorrecte besluiten.

Appropriate reliance is daarmee het volgen van AI-systemen wanneer die het correcte advies weergeven, maar de contactambtenaar blijft voldoende sceptisch om incorrecte AI-adviezen niet op te volgen (Schemmer et al., 2022). Het draait daarmee om het gepast omgaan met AI-adviezen. Hieronder is een schematische weergave gegeven van *appropriate reliance*.

	Keuze besluitnemer In overeenstemming met het AI-systeem	**Keuze besluitnemer** Niet overeenstemming met het AI-systeem
Advies AI: Het advies klopt niet	**Overreliant**	**Appropriate reliant**
Advies AI: Het advies klopt	**Appropriate reliant**	**Underreliant**

Figuur 1. De binaire typering van appropriate reliance

Dit onderzoek kiest expliciet voor de binaire definitie van *appropriate reliance* van Schemmer et al. (2022), waarin een beslissing juist of niet juist is, om de volgende redenen: Deze operationalisering past bij de context van dit onderzoek en de methodiek die gebruikt wordt om *appropriate reliance* te toetsen. In veel gevallen is het vertonen van een juiste hoeveelheid vertrouwen minder binair, omdat er niet één 'juiste' keuze bestaat. In dat geval is *appropriate reliance* een spectrum, in plaats van een matrix tabel zoals weergegeven hierboven (Bruijnes et al., 2024). Ook in het politiedomein is vooraf lastig te zien welke keuze de juiste is. In sommige gevallen blijkt dit pas achteraf. Daarom wordt in dit onderzoek gesproken van een waarschijnlijk juist en waarschijnlijk onjuist antwoord. Dit behoudt de binaire definitie die past bij de onderzoeksmethodiek, maar houdt ruimte voor de onzekerheid die het politiedomein omvat.

Ter illustratie, wanneer een AI-systeem 95% zeker weet dat het antwoord A is, dan is dat het waarschijnlijk juiste antwoord en dus *appropriate reliant*. Tegelijkertijd wanneer een politie agent zeker weet dat dit het moment is wanneer de 5% kans opduikt voor onwaarschijnlijkheid door een expliciet detail dat het AI-systeem over het hoofd heeft gezien, dan is de andere optie het waarschijnlijk juiste antwoord.

Dat in acht nemende is de definitie van appropriate reliance en de beantwoording van de tweede deelvraag als volgt: *"De menselijke vaardigheid om te*

differentiëren tussen een waarschijnlijk correct en waarschijnlijk incorrect AI-advies en te kunnen handelen op dat verschil" (Schemmer et al., 2022).

2.3 Verantwoording

2.3.1 Underreliance en de politie

In het onderstaande hoofdstuk wordt de deelvraag beantwoord wat verantwoording afleggen inhoudt in de context van AI-besluitprocessen.

Het ideaalbeeld is dat contactambtenaren altijd *appropriate reliant* op AI-systemen zijn, maar dit blijkt uitdagend. Eerdere onderzoeken bij de politie laten zien dat zowel *over-* als *underreliance* voorkomt. In het onderzoek van Meijer et al. (2021) vertonen de Berlijnse politieagenten consistent vormen van *overreliance*, waardoor zij niet meer sceptisch tegenover de adviezen van algoritmes staan. Terwijl het onderzoek van Selten et al. (2023) het tegenovergestelde aantoont bij Nederlandse politieagenten. De Nederlandse politieagenten lijken consequent hun eigen ingevingen op te volgen in plaats van de algoritmische aanbevelingen.

Het meest wenselijke scenario is wanneer AI en mens samenwerken om tot de best mogelijke besluiten te komen. Bij de discretionaire ruimte van contactambtenaren bestaat namelijk ruimte voor empathie en flexibiliteit per situatie (J. B. Bullock, 2019; Evans & Hupe, 2019; Maynard-Moody & Musheno, 2000; Moore, 2013). Daarbij kunnen AI-systemen de mens aanvullen met consistentie, efficiëntie en effectiviteit (Binns, 2022; J. Bullock et al., 2020; Young et al., 2019). Daarom wordt de samenwerking tussen mens en AI als wenselijk gezien door de mogelijk aanvullende rol. Zo heeft de contactambtenaar de rol van een *human in the loop* ofwel dat ambtenaren de besluitvorming overzien en de gemaakte keuzes van AI-systemen kunnen tegenhouden (Peeters, 2020). Tegelijkertijd heeft deze rol beperkingen. De uitdagingen bij de besluitnemer stipt Peeters (2020) aan als de menselijke controle problemen.

"Menselijke controle problemen verwijzen naar de neiging van mensen om bij mens-machine interacties zelfingenomen te zijn of door te veel vertrouwen te tonen bij het omgaan met een betrouwbaar autonoom systeem" (Peeters, 2020). Specifiek leidt dit tot overreliance op AI-systemen. Het eerste omschreven probleem is

begrensde rationaliteit, de lastig te begrijpen systemen leiden tot misinterpretaties of een te korte aandacht inspanning. Hierdoor zijn mensen minder in staat om fouten te detecteren (Peeters, 2020). Het tweede menselijke controle probleem is bevredigend gedrag vertonen. Dit houdt in dat ambtenaren de AI-besluiten niet onderzoeken door druk vanuit de organisatie. Ter illustratie: het management kan beweren dat AI-systemen efficiënter en/of effectiever zijn. Een goed voorbeeld dat deze druk illustreert is de toeslagenaffaire bij de belastingdienst, waar het de norm was om de algoritmische besluiten niet in twijfel te trekken (Peeters, 2020; Peeters & Widlak, 2023). Dit leidde ertoe dat vele burgers onterecht als fraudeur werden bestempeld. Automatisering bias is het derde controle probleem en betekent het geloof in de wetenschappelijke neutraliteit en rationaliteit van algoritmische procedures (Alon-Barkat & Busuioc, 2023; Peeters, 2020). Dit leidt eveneens tot overreliance, want zoals eerder is geconcludeerd: algoritmes en dus AI-systemen zijn niet neutraal en niet volledig rationeel.

Echter, de uitdaging bij contactambtenaren is niet *overreliance* in AI-advies, maar het is juist *underreliance* dat herhaaldelijk voorkomt. Peeters (2020) stipt aan dat contactambtenaren het tegenbewijs leveren van de vorige drie menselijke controle problemen, waarbij het probleem *overreliance* het meest voorkomend is. Een belangrijke reden hiervoor is dat contactambtenaren weinig inzicht hebben in het besluitvormingsproces van een AI-systeem, maar tegelijkertijd worden zij wel verantwoordelijk gehouden voor de uiteindelijke beslissingen (Binns, 2022; Elish, 2019; Keddell, 2019; Peeters, 2020). Een voorbeeld hiervan is het onderzoek van Keddell (2019) waarin contactambtenaren van de Nieuw-Zeelandse kinderbescherming niet kunnen uitleggen aan ouders waarom een AI-systeem tot een bepaalde beslissing komt. Dit soort gevoelige onderwerpen en thema's, hebben een menselijke hand nodig die compassie kan tonen indien nodig. Bovendien zijn contactambtenaren behoedzaam om verstrikt te raken in wat Elish (2019) 'morele kreukelzones' noemt. *"Een morele kreukelzone is de manier waarop een verantwoordelijkheid van een actie verkeerd kan worden toegeschreven aan een mens die beperkte controle heeft over het gedrag van een autonoom systeem"* (Elish, 2019). Denk hierbij aan een zelfrijdende auto die crasht en de chauffeur krijgt de

boete. Uit angst om in de morele kreukelzone te belanden, zetten contactambtenaren zich af tegen AI-besluitvorming, dus ontstaat *underreliance*.

Een belangrijke nuance is, dat de aanwezigheid van *underreliance* sterk contextafhankelijk is, doordat elk beroep van contactambtenaar een unieke taak, organisatie setting, institutionele omgeving en management context biedt (Noordegraaf, 2015; Peeters, 2020). Eerder onderzoek heeft al aangetoond dat Nederlandse politieagenten eerder vertrouwen op hun eigen professionele kennis dan AI-adviezen (Meijer et al., 2021; Selten et al., 2023). De agenten zien de AI-systemen meer als een collega dan iets dat ze 'moeten' volgen en wanneer die collega afwijkt van hun eigen inzichten, kiezen ze ervoor om niet naar het AI-systeem te luisteren. (Meijer et al., 2021). Schemmer et al. (2023) hebben aangetoond dat *Explainable AI*, ofwel XAI bij kan dragen aan het voorkomen van *underreliance*. XAI kan de uitkomst van een AI-advies uitleggen inclusief de daarbij horende redenering en rechtvaardiging (Miller, 2019a). Echter, datzelfde onderzoek binnen de politie van Selten et al. 2023) zag dat uitlegbare AI-systemen niet bijdragen aan een toename in vertrouwen van AI-systemen. Recent onderzoek toont aan dat Nederlandse politieagenten *underreliant* zijn op AI-adviezen door hun professionele bias en het zien van AI-systemen als een collega (Meijer et al., 2021).

2.3.2 Het afleggen van verantwoording

In de relatie tussen AI en mens, legt XAI de bal bij het systeem om de aanbevelingen te rechtvaardigen tegenover een besluit nemende contactambtenaar. Maar wat als de contactambtenaar degene is die verantwoording moet afleggen? Volgens de morele kreukelzonetheorie zal een contactambtenaar zich afzetten tegen AI-systemen, omdat zij weinig controle hebben over de keuzes van het systeem (Elish, 2019). Echter, dit onderzoek kijkt expliciet naar AI-adviezen, ofwel de ambtenaar heeft controle of zij het advies opvolgen of niet. Wat gebeurt er dan wanneer contactambtenaren van te voren weten dat zij verantwoording moeten afleggen over hun keuze? Zal de contactambtenaar zich blijven afzetten tegen het advies van een AI? Of zal hij *appropriate reliant* proberen te handelen om te komen tot de best mogelijke beslissingen, omdat hij verantwoordelijk wordt gehouden?

Uit de psychologische literatuur blijkt dat het afleggen van verantwoording gedrag en keuzeprocessen beïnvloeden. Tetlock (1983) als grondlegger van de psychologische literatuur rondom verantwoording, ontdekte in de jaren 80' al dat verantwoording afleggen een positief effect had op het verminderen van eerste indruk bias. Een van de reacties die Tetlock vond genaamd preventieve zelfkritiek, zorgt voor beter overwogen keuzes en kritisch zijn op eigen gedrag (Aleksovska et al., 2019; Schillemans et al., 2021; Tetlock, 1983). Doordat proefpersonen van te voren wisten dat zij verantwoording moesten afleggen bleven zij langer stilstaan bij de eigen overwegingen en potentiële bias.

Sindsdien is de literatuur rondom het onderwerp verantwoording afleggen, zich binnen de psychologie en publiek management verder gaan ontwikkelen. Onder andere is meermaals bewezen dat verantwoording afleggen een positief effect heeft op het maken van meer doordachte keuzes en het voorkomen van bias (Aleksovska et al., 2019). Uit een onderzoek van Ličen & Slapničar (2022) blijkt het dat verantwoording afleggen managers helpt om hun bias te overkomen voor korte termijn profijt dat ten kosten gaat van de lange termijn. Managers die wisten dat ze verantwoording af moesten leggen spendeerde meer tijd om hun keuzes inzichtelijker te maken. Uit de verworven nieuwe inzichten die ze creëerden bleek dat de lange termijn kosten groter waren dan dat ze origineel dachten (Ličen & Slapničar, 2022). Eveneens helpt verantwoording afleggen automatisering bias te voorkomen. Ondanks dat de proefpersonen een monotone geautomatiseerde taak moesten overzien bleek uit een studie van Skitka et al. (2000) dat de proefpersonen die verantwoording aflegden, oplettender en accurater waren in hun beslissingen. Mede omdat deze proefpersonen grondiger informatie doorzochten en het belangrijker vonden om accuraat te reageren. Kortom als verantwoording afleggen bijdraagt aan het maken van meer doordachte keuzes en het voorkomen van bias, dan is het wellicht datgene dat nodig is om contactambtenaren te sturen richting *appropriate reliance*.

Om te toetsen of verantwoording afleggen effect heeft op *appropriate reliance*, moet het soort verantwoording dat wordt afgelegd passen binnen het politiedomein. De politieagenten hebben mandaat om bepaalde besluiten te nemen,

maar bevinden zich tegelijkertijd in een hiërarchische organisatie met strikte kaders vanuit wet- en regelgeving (Nederlandse Politie Bond, 2023; Politie, n.d.; Van Koetsveld et al., 2016). Dit zijn twee kanten die kunnen conflicteren. Het mandaat geeft de politie unieke en bijzondere rechten en een bepaalde macht. Tegelijkertijd is dit mandaat sterk ingekaderd binnen de hiërarchische organisatie. Daarom is het van belang dat de agent van tevoren aanvoelt dat hij verantwoordelijk is voor besluiten gemaakt op basis van AI-adviezen. Het type verantwoording dat hierbij past is *felt accountability*, ofwel gevoelde verantwoordelijkheid (Hochwarter et al., 2007; Overman & Schillemans, 2022).

Gevoelde verantwoordelijkheid is: (1) de impliciete of expliciete verwachting dat de beslissingen of acties van iemand worden onderworpen aan evaluatie door een prominent publiek, (2) met de overtuiging dat de mogelijkheid bestaat dat diegene beloningen of sancties ontvangt op basis van deze verwachte evaluatie (Hochwarter et al., 2007; Overman & Schillemans, 2022; Schillemans et al., 2021).

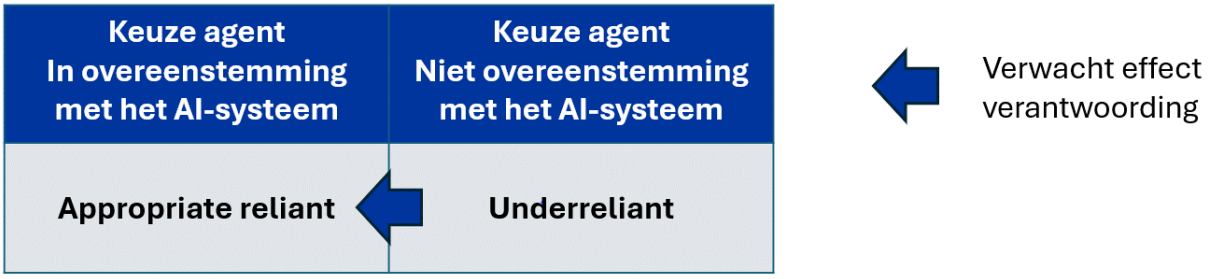
De redenen waarom gevoelde verantwoordelijkheid specifiek is gekozen zijn als volgt: Ten eerste bouwt het verder voort op publiek management literatuur over verantwoordelijkheid. Ten tweede wordt een hoge graad van gevoelde verantwoordelijkheid gevoeld bij contactambtenaren waaronder de politie (Overman & Schillemans, 2022). Ten derde focust gevoelde verantwoordelijkheid op factoren binnen de organisatie, de verandering die AI teweeg brengt geldt ook voornamelijk binnen de organisatie. Andere vormen van verantwoording focussen vaak ook op externe factoren. Dit maakt dat gevoelde verantwoordelijkheid past bij het politie domein en dit onderzoek (Deng & Sun, 2024).

Daarmee is het antwoord op de deelvraag: *“wat wordt verstaan onder het afleggen van verantwoording in de context van AI aanbevelingen?”*, dat politieagenten voelen alsof zij verantwoording moeten afleggen over het AI-besluit. Om hieraan te voldoen moet van te voren (1) de impliciete of expliciete verwachting gewekt worden dat de beslissingen of acties van iemand worden onderworpen aan evaluatie door een prominent publiek (zoals een teamleider of inspectie), (2) met de overtuiging dat er de mogelijkheid bestaat dat diegene beloningen of sancties

ontvangt op basis van deze verwachte evaluatie (Hochwarter et al., 2007; Overman & Schillemans, 2022; Tetlock, 1983).

In overeenstemming met eerder bevonden literatuur over de het verantwoordingsmechanisme wordt verwacht dat het effect van verantwoording afleggen professionele bias tegengaat en dat de politie doordachte keuzes maakt (Aleksovska et al., 2019). In relatie tot *appropriate reliance*, wordt de eerste hypothese opgesteld:

H1. Het afleggen van verantwoording door politieagenten aan een prominent publiek (zoals een teamleider) draagt bij aan het voorkomen van underreliance op AI-advies



Figuur 2. Verwacht effect H1. Appropriate reliance wanneer het AI-systeem een waarschijnlijk juiste keuze maakt

2.3.3 Ervaring en moderatie

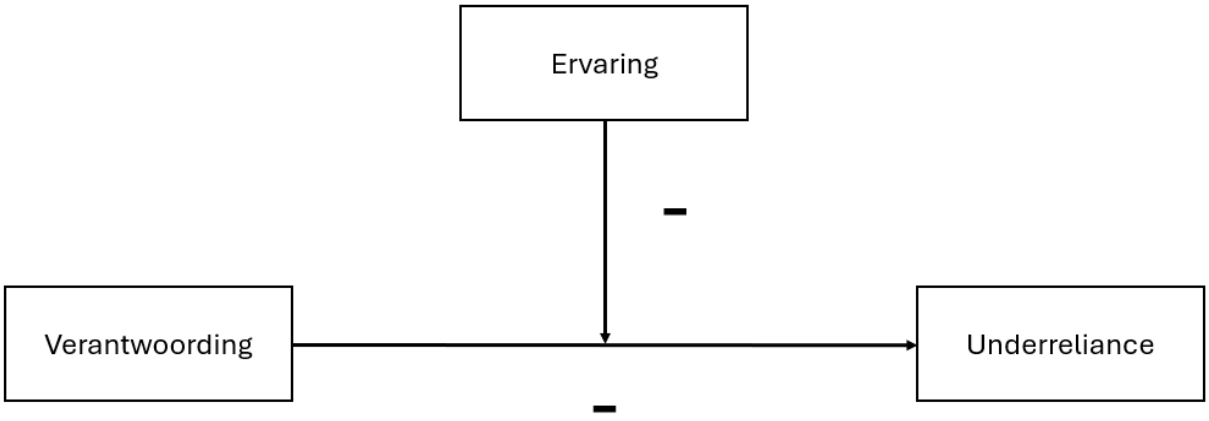
Naast het hoofdeffect wordt een modererend effect van ervaring verwacht. Dit verwachte effect is gebaseerd op de rol die bias speelt in de beïnvloeding van *appropriate reliance*. H1 is gebaseerd op het feit dat verantwoording afleggen ertoe leidt dat contactambtenaren beter doordachte keuzes maken (Aleksovska et al., 2019; Skitka et al., 2000; Tetlock, 1983). Een van de mechanismen hierachter is dat de inherente bias wordt overwonnen door het maken van beter doordachte keuzes (Aleksovska et al., 2019). Echter, het kan voorkomen dat de interne bias van een contactambtenaar sterker is dan dit effect.

Contactambtenaren die in het veld werken, hebben de specifieke ervaringen, trainingen en intuïtie die bijdragen aan hun bestaande bias (Maynard-Moody & Musheno, 2000; Noordegraaf, 2015; Thomann et al., 2018; Tummers & Bekkers, 2014). Ervaring draagt bij aan het versterken van deze bias. Een agent die de politieacademie heeft gevolgd, gewend is geraakt aan de cultuur binnen de politie en

jaren van veranderingen binnen de politie heeft meegemaakt, heeft bij voorhand een mening over de implementatie van digitalisering. Deze agent heeft soortgelijke veranderingen al eerder in hun carrière ervaren. Bovendien heeft die agent normen en waarden opgebouwd die hun mening zullen vormen met betrekking tot het krijgen van advies van een AI-systeem (Noordegraaf, 2015; Selten et al., 2023).

De rol die bias speelt in het wel of niet opvolgen van keuzes van AI is eerder bewezen in het onderzoek van Selten et al. (2023). Hierin is aangetoond dat Nederlandse politieagenten herhaaldelijk hun persoonlijke bias volgen in plaats van het advies van een algoritme. Daarom wordt in dit onderzoek rekening gehouden met ervaring, aangezien deze de persoonlijke bias kan versterken(Noordegraaf, 2015). Hiermee wordt een negatieve modererende invloed van ervaring verwacht op de relatie van verantwoording en *appropriate reliance*. Dit leidt tot de tweede hypothese:

H2. Ervaring binnen de politie heeft een negatief effect op de relatie tussen verantwoording en underreliance



Figuur 3. Het negatieve (causale) effect tussen verantwoording en underreliance, met ervaring als negatieve (correlerende) moderator.

3. Methode en metingen

In het methode hoofdstuk wordt inzicht geboden in de manier waarop data verzameld en geanalyseerd is. Eerst wordt ingegaan op het experimentele ontwerp, de gebruikte materialen en de operationalisering van de variabelen. Vervolgens wordt aandacht besteed aan de methode van dataverzamelingsmethode en de steekproef. Tot slot komen de manipulatiecheck, de onderzoeksethiek, en de validiteit en betrouwbaarheid van het onderzoek aan bod.

3.1 Experimenteel design

De centrale vraag van dit onderzoek is: *Wat is het effect van het afleggen van verantwoording door contactambtenaren over AI-keuzeprocessen op gepast vertrouwen (appropriate reliance) in AI-systemen?* Om deze vraag te beantwoorden is in samenwerking met het Nationaal Politielab AI een survey experiment opgezet en gehouden binnen een populatie van politieagenten.

Het specifieke soort survey experiment dat is uitgevoerd, is een vignettenexperiment. Een vignet biedt een korte omschrijving van een persoon of situatie met een verwijzing naar de belangrijkste factoren in het besluitvormingsproces van de respondent (Peters & Guedes-Neto, 2020). In dit experiment worden politieagenten gepresenteerd met twee verschillende vignetten, ofwel scenario's. Beide scenario's toetsen dezelfde variabelen, maar verschillen van casus om de ecologische validiteit te vergroten (Anderson-Cook, 2005). In elk scenario wordt een misdaad gepleegd. De agent krijgt een melding van de misdaad en wordt gevraagd te helpen bij het afzetten van wegen met blokkades om de daders tegen te houden. De agenten worden ondersteund door het advies van een AI-systeem dat mogelijke vluchtroutes in kaart brengt. Het AI-systeem geeft twee opties voor het plaatsen van de blokkades en biedt een advies, inclusief een uitleg waarom een bepaalde keuze als beste wordt beschouwd. De agent moet vervolgens besluiten tussen één van de twee locaties om de blokkade te plaatsen.

Net als bij een klassiek experiment is een onderscheid gemaakt tussen de twee groepen: De verantwoordinggroep die wordt gevraagd om verantwoording af te leggen over hun keuze, en de controlegroep die geen verantwoording hoeft af te

leggen. Door enkel de verantwoordingsvariabele te variëren tussen de groepen, kan het causale effect van het afleggen van verantwoording als onafhankelijke variabele op het voorkomen van *underreliance* als afhankelijke variabele worden getoetst (Baekgaard et al., 2015). Om uit te sluiten dat een gevonden causaal effect tussen de variabelen te wijten is aan andere derde variabelen, is het van belang dat de respondenten willekeurig zijn verdeeld tussen de verschillende groepen (Druckman et al., 2011). Hierdoor ontstaat een willekeurige verdeling van achtergrondvariabelen zoals technologische kennis, leeftijd, enzovoort. Hierdoor wordt het risico van *non-spuriousness* uitgesloten, waarin de gevonden causale relatie wordt veroorzaakt door een derde, niet gemeten variabele (McDermott, 2002).

Dit experiment is gebaseerd op een experiment van Selten et al. (2023). Het zogenaamde AI-systeem is gebaseerd op een AI ontwikkeld door het Nationale Politielab AI, bedoeld om agenten en de meldkamer te ondersteunen. Dit systeem is de Intelligente Geografische Meldkamer Assistent (IGMA). Bijna alle politieagenten krijgen te maken met een vorm van het gepresenteerde scenario in hun training of in hun politiewerk. Hierdoor is het mogelijk om de vragenlijst af te nemen bij veel verschillende agenten. Dit draagt bij aan de ecologische validiteit van het experiment (Anderson-Cook, 2005). De ecologische validiteit wordt verder versterkt door de gemaakte aannames tijdens het ontwerpproces te toetsen bij politieagenten. Daarnaast is het experiment geëvalueerd en van feedback voorzien door datawetenschappers van het Nationaal Politielab AI, studenten bestuurs- en organisatiewetenschap en wetenschappelijke experts.

3.2 Materialen

In het onderstaande stuk wordt stilgestaan bij de invulling van het survey experiment. Onder andere worden de verschillen tussen de controlegroep en de verschillende behandelingsgroepen uitgelegd. Eerder is genoemd dat dit experiment is gebaseerd op een experiment van Selten et al. (2023). Eveneens zijn de materialen door dit onderzoek geïnspireerd. Tegelijkertijd zitten er twee belangrijke verschillen tussen de onderzoeken. Ten eerste was het doel van Selten et al. (2023) om het effect

van *XAI* te toetsen op vertrouwen in AI-adviezen, terwijl dit experiment het effect van verantwoording afleggen op het voorkomen van *underreliance* toetst. Ten tweede gebruikt dit onderzoek andere scenario's dan degene die voorgelegd waren in het onderzoek van Selten et al. (2023). Deze scenario's zijn beter afgestemd op de hoofdvraag van dit onderzoek en in samenwerking ontwikkelt met het Nationaal Politielab AI.

3.2.1 Algemene opzet

De algemene opzet van het experiment is als volgt. De respondent begint bij een korte introductie van het onderzoek, inclusief de geïnformeerde toestemming die afgegeven moet worden om mee te doen. Vervolgens wordt een uitleg gegeven over wat de respondent precies zal doen in de vragenlijst (zie tekstvak 1). Na de uitleg volgt een aantal demografische vragen. Als laatste zal de respondent aan de slag gaan met de opgestelde scenario's; hier vindt ook de volledige randomisatie

Wat ga je doen

De eerste 10 minuten na de melding van een strafbaar feit zijn van cruciaal belang voor het aanhouden van een verdachte. Een centralist bij de meldkamer kan in deze eerste tien minuten niet alle eenheden op een ideale locatie positioneren. **Daarom moet jij als agent eerst een eigen inschatting maken op welke plek je het best positie kunt innemen.**

Om je te ondersteunen bij het maken van deze inschatting heeft de politie een computersysteem ontwikkeld: **de Hulp Onderschepping Verdachten (HOV)**. De HOV maakt een **inschatting van de vluchtroutes** die verdachten kunnen gebruiken en adviseert waar jij het beste positie kan innemen
om hen te onderscheppen.

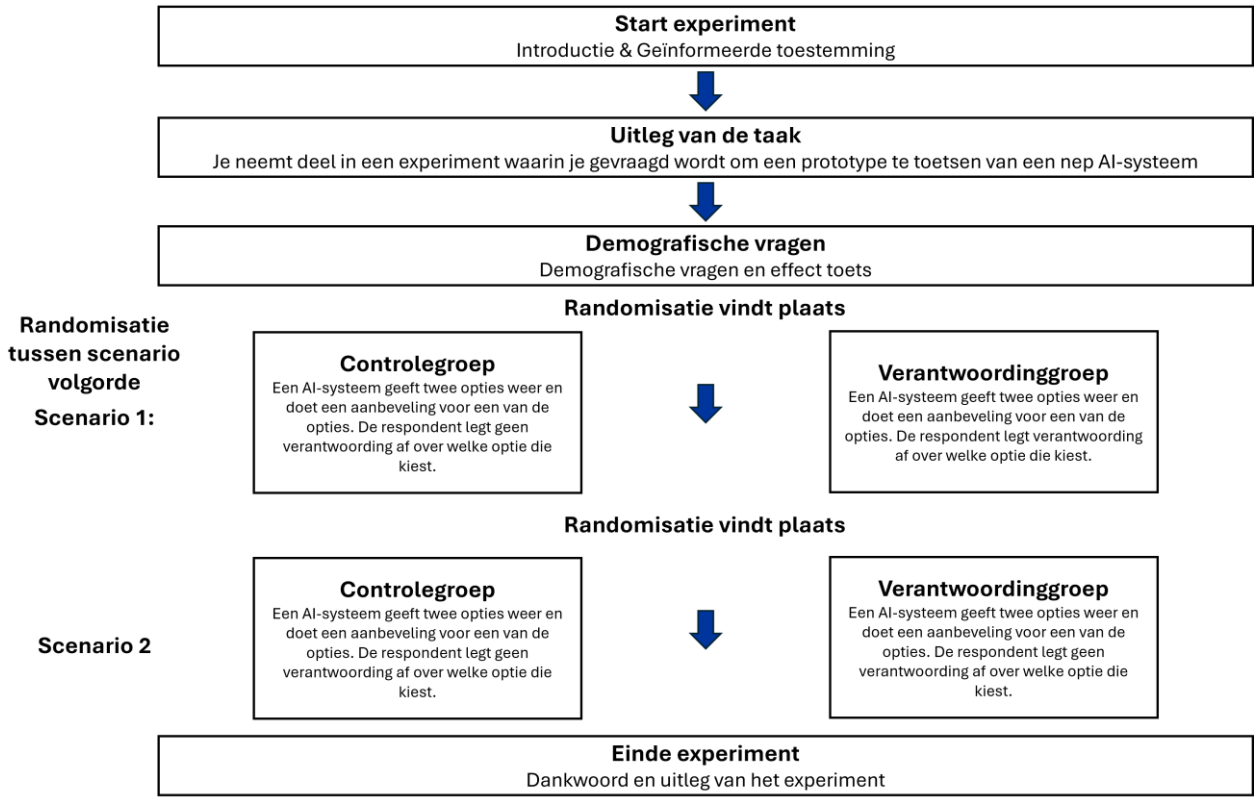
Je gaat zo meedoen aan een experiment waarin je wordt gevraagd om te werken met de HOV. Over de werking van de HOV moet je weten dat:

- Tests hebben uitgewezen dat in de meeste gevallen de HOV een betere inschatting maakt over de vluchtroute van verdachten dan een mens. We vragen je daarom het advies van de HOV serieus te nemen.
- Jouw eigen kennis, ervaring en intuïtie blijven echter belangrijk. De HOV geeft je een gefundeerd advies, maar je kunt ervoor kiezen dit advies niet te volgen.

Let op: De HOV bestaat niet echt, maar is wel gebaseerd op een systeem dat wordt ontwikkeld door de politie.

Tekstvak 1. Uitleg die respondent ontvangt voor het experiment

plaats. Zowel de groepsindeling als de volgorde van scenario's in random. Voor een volledige flowchart van het experiment zie figuur 4.



Figuur 4. Flowchart experiment

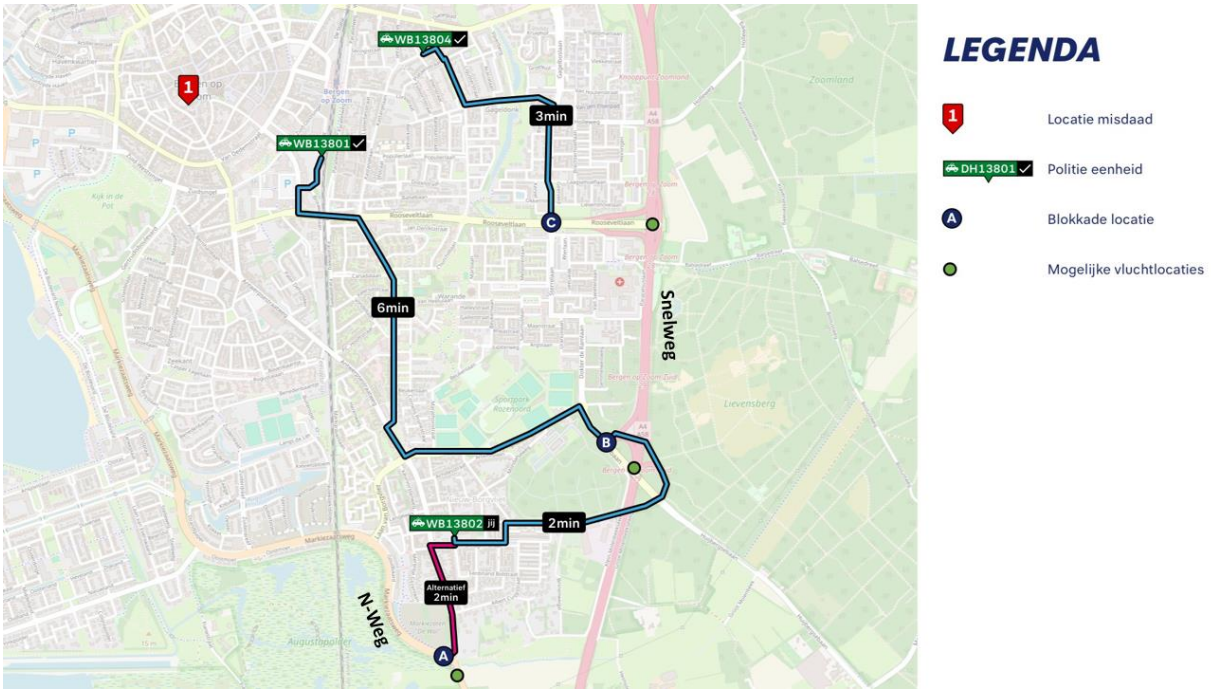
In het survey experiment worden participanten twee verschillende scenario's voorgelegd (Bergen op Zoom en Sassenheim, de volledige scenario's zijn te vinden in de enquête in Bijlage F). De scenario's worden uitgelegd in stukken tekst en zijn vervolgens afgebeeld op een kaart (zie figuur 5). In elk scenario vindt een misdaad plaats. De criminelen vluchten weg van de scene delict in een aangegeven richting en het is de taak van de participant om een politieauto blokkade neer te zetten op de juiste plek. Ter ondersteuning suggereert de Hulp Onderschepping Verdachten (HOV), een verzonnen AI-systeem, twee plekken om de politieblokkade neer te zetten. Beide locaties dat het de HOV aangeeft krijgen een korte uitleg waarom deze locatie gekozen zijn. Daarbij doet de HOV een aanbeveling op welke van de twee plekken de participant de blokkade het beste kan neerzetten. Na het lezen van de aanbeveling die het AI-systeem doet krijgt de participant de optie om te kiezen voor locatie A of B.

3.2.2 Verantwoording: Controlegroepen vs. verantwoordingsgroepen

De controlegroepen toetsen het gedrag van de politieagenten wanneer zij geen verantwoording afleggen. De controlegroep werkt net als de verantwoordingsgroep met het AI-systeem dat een aanbeveling geeft en een uitleg bij deze aanbeveling. In de controlegroep wordt expliciet weergegeven dat een teamleider geen verantwoording eist: *"Om de kans te vergroten dat de verdachten worden aangehouden wordt de Hulp Onderschepping Verdachten (HOV) ingeschakeld.* ***Je teamleider vraagt in dergelijke situaties nooit om je keuzes achteraf uit te leggen."***

De verantwoordingsgroepen krijgen dezelfde scenario's voorgelegd als de controlegroep met een belangrijk verschil. Participanten in de verantwoordingsgroep moeten verantwoording afleggen over hun keuze. Dit is als volgt weergegeven: *" Om de kans te vergroten dat de verdachten worden aangehouden wordt de Hulp Onderschepping Verdachten (HOV) ingeschakeld.* ***Bij gebruik van de HOV vindt jouw teamleider het belangrijk dat jij je keuze achteraf verantwoord door uit te leggen waarom jij wel of niet hebt geluisterd naar het algoritme."*** Vervolgens worden de participanten ook om uitleg gevraagd over hun locatie keuze. Het geven van uitleg over die keuze wordt niet gevraagd bij de controlegroep. Deze manier van verantwoording afleggen past bij de definitie uit het theoretisch kader. Bij die definitie was het belangrijk dat van te voren (1) de impliciete of expliciete verwachting gewekt wordt dat de beslissingen of acties van iemand worden onderworpen aan evaluatie door een prominent publiek (zoals een teamleider of inspectie), (2) met de overtuiging dat de mogelijkheid bestaat dat diegene beloningen of sancties ontvangt op basis van deze verwachte evaluatie (Hochwarter et al., 2007; Overman & Schillemans, 2022; Tetlock, 1983). In de verantwoordingsgroep wordt expliciet van te voren de verwachting gecreëerd dat verantwoording afgelegd moet worden. Dit wordt gedaan door een prominent publiek (de teamleider) die de acties aan een evaluatie onderwerpt, met mogelijkheid tot beloning of sancties. Teamleiders maken namelijk evaluaties over het handelen van hun team. Het idee van deze evaluaties sluit ook aan bij de criteria omtrent de verwachte beloning of sancties.

Bij het eerste scenario (welke dat is, wordt bepaald door de randomisatie) wordt gevraagd naar in welke mate de respondent zich verantwoordelijk voelt voor keuzes die hij neemt in verband met AI-adviezen. Dit is de manipulatiecheck naar verantwoording. Deze vraag zal dus zowel bij de controlegroep als de verantwoordingsgroep verschijnen. De keuze is expliciet gemaakt om de manipulatiecheck alleen na het eerste scenario uit te voeren, want het zal vreemd overkomen als diezelfde vragen twee keer gesteld worden. Verdere informatie en de resultaten van de manipulatiecheck wordt weergegeven in 3.5.



Figuur 5. Scenario Bergen op Zoom, liquidatie in het theater

3.2.3 Appropriate reliance

De manier waarop dit onderzoek *appropriate reliance* en *underreliance* operationaliseert is door de manier waarop de scenario's zijn opgesteld. *Appropriate reliance* is eerder gedefinieerd in het theoretisch kader als "*De menselijke vaardigheid om te differentiëren tussen een waarschijnlijk correct en waarschijnlijk incorrect AI-advies en te kunnen handelen op dat verschil*" (Schemmer et al., 2022). *Underreliance* is daarmee gedefinieerd als het niet volgen van een waarschijnlijk correct AI-advies. Dit onderzoek toetst specifiek *underreliance*. Daarom geeft de HOV in beide scenario's een waarschijnlijk correct

advies. Specifiek geeft de HOV een correct advies dat op eerste blik onwaarschijnlijk is voor iemand met professionele kennis, maar waarschijnlijk wordt gemaakt door de uitleg van het AI-systeem.

Ter illustratie: in het Bergen op Zoom scenario worden aan de participant twee opties gepresenteerd voor de blokkadelocaties die de vluchtroutes afzetten. De snelweg en een landweg. De snelweg zal ook geblokkeerd worden door een collega, jij kan de landweg afsluiten die de laatste uitgang uit het dorp blokkeert of bij de snelweg gaan staan samen met de collega. Normaliter is de landweg de waarschijnlijke blokkade locatie. Door de landweg te kiezen zal de respondent het dorp volledig af kunnen zetten. Dit is daarom volgens 'professionele kennis' de waarschijnlijke locatie. Het advies van de HOV is, in tegenstelling tot de professionele kennis, om de snelweg af te zetten. In de uitleg geeft de HOV weer dat jij namelijk twee keer zo snel bij de snelweg kan zijn als jouw collega. Snelwegen zijn over het algemeen belangrijkere vluchtroutes om af te zetten en bij het afzetten van wegen telt elke seconde. Dus de echte 'waarschijnlijk' juiste keuze is om bij de snelweg te gaan staan. Hier komt de respondent alleen achter als hij de casus beter bekijkt en stilstaat bij het advies van de HOV.

Het effect van verantwoording op *underreliance* wordt daarmee getoetst door te kijken of de controlegroep in hogere mate kiest voor de landweg, ofwel meer *underreliant* is dan de verantwoordingsgroep. De verwachting is dat de verantwoordingsgroep meer tijd besteedt om na te denken over het meest waarschijnlijke antwoord. Hierdoor nemen zij de tijd om het antwoord dat waarschijnlijk is door hun professionele kennis (de landweg) te heroverwegen en het detail van de collega die er langer over doet, te heroverwegen (Aleksovska et al., 2019; Tetlock, 1983). Figuur 6 geeft de operationalisatie van *appropriate reliance* binnen dit experiment weer.

Keuze agent In overeenstemming met het AI-systeem	Keuze agent Niet overeenstemming met het AI-systeem
Appropriate reliant Agent heeft voor het waarschijnlijke blokkade punt gekozen in overeenstemming met het AI-systeem	**Underreliant** Agent heeft voor het onwaarschijnlijke blokkade punt gekozen ondanks de aanbeveling van het AI-systeem

Figuur 6. Appropriate reliance operationalisatie

De twee scenario's toetsen in essentie hetzelfde en de manipulatie tussen de controle en verantwoordingsgroepen blijft eveneens hetzelfde. Het verschil tussen de scenario's is de casus. In het Bergen op Zoom scenario wordt een liquidatie gepleegd bij het theater, waarbij de agent moet kiezen tussen de landroute en de snelweg. In het Sassenheim scenario wordt een inbraak gepleegd en moet de agent kiezen tussen het afsluiten van een snelweg waarbij een brug dicht gaat of het afsluiten van een alternatieve landweg die naar het volgende dorp leidt. Door twee scenario's te gebruiken wordt een grotere N behaald en kan hetzelfde getoetst worden in een andere context. Dit vergroot de betrouwbaarheid van het onderzoek (Bryman, 2015).

Er vindt eenmalig een volledige randomisatie plaats. De randomisatie is tussen groepen en tussen de volgorde van scenario's. Dat houdt in dat het volgende kan voorkomen: een respondent die bij het Sassenheim scenario de controlegroep krijgt, kan vervolgens bij het Bergen op Zoom scenario de verantwoordingsgroep krijgen. Dit kan leiden tot aangepast gedrag van respondenten omdat zij zien dat de vragenlijst tussendoor verandert door de wisseling tussen controle- en verantwoordingsgroep. Echter, een onderzoek naar *demand effects* door Mummolo & Peterson (2019) bewijst dat participanten die de hypothese van een experiment weten, alsnog geen gewenste gedragingen vertonen die de hypothese expres zal bewijzen. Het gedrag van de participanten die de hypothese wel wisten verschilde amper van de controlegroep. Dit bewijs en de voordelen van een volledige

randomisatie hebben bijgedragen aan de keuze om het experiment volledig te randomiseren, waaronder tussen groepen.

3.2.4 Samengestelde schalen

De enquête maakt gebruik van twee samengestelde schalen. De schaal van vertrouwen in technologie en de schaal voor verantwoordingsgevoel. Beide schalen bestaan uit drie stellingvragen en maakten gebruik van een zevenpuntschaal (1 = Helemaal mee oneens; 7 = Helemaal mee eens). Voordat deze schaal gebruikt wordt, wordt de validiteit getoetst middels een factoranalyse en Cronbach's Alpha (Bryman, 2015; Field, 2018; Samuels, 2016).

Bij de factoranalyse is de verwachting dat beide schalen op één factor laden. Hieruit kan een samengestelde schaal gevormd worden, aangezien de variantie tussen de vragen genoeg overeen komt (Brandler & Roman, 2020). Wanneer een vraag onvoldoende laadt kan ervoor gekozen worden om deze niet mee te nemen in de samengestelde schaal. De twee voorwaardes voor een goede lading zijn als volgt. Ten eerste wordt een minimale lading van 0,3 verwacht (0,5 is wenselijk). Ten tweede mogen de items niet op een ander concept laden met een kleiner veschil dan 0,2. (Brandler & Roman, 2020; Field, 2018; Samuels, 2016).

De factoranalyse is gedaan middels een maximum likelihood extractie en een direct oblimin rotatie. Deze extractie en rotatie zijn passend, want er wordt een bepaalde hoeveelheid samenhang verwacht tussen de items. Aangezien de vragen gebaseerd zijn op schalen die eerder getoetst zijn, kan deze samenhang ook verondersteld worden. De volledige factoranalyses zijn weergegeven in Bijlage A.

De Kaiser-Meyer-Olkin (KMO) scores tonen aan dat de data geschikt zijn om een factoranalyse op uit te voeren. De factoranalyses scoren hoger dan nodig, namelijk 0,6. Eveneens tonen significante Barlett's toetsen (<0,001) aan dat de variantie tussen groepen gelijk zijn. Daarmee zijn beide factoranalyses geslaagd, want in beide factoranalyses laden de vragen op dezelfde dimensie met waardes hoger dan 0,5 (Samuels, 2016).

Als laatste wordt de interne betrouwbaarheid getoetst middels Cronbach's Alpha. Deze toets kijkt naar hoe consistent de verschillende vragen één concept

meten (Bryman, 2015). De resultaten van deze toets vallen tussen de 0 (inconsistent) en 1 (consistent). De waarde voor technologie vertrouwen is (α = 0,778). Dit is lager dan gewenst. Eveneens is (α = 0,760) ook lager dan gewenst. Vanaf 0,7 wordt gezien als acceptabel, maar hoger dan 0,8 is gewenst. Aangezien beide waarden acceptabel zijn, worden deze schalen gebruikt binnen het onderzoek.

3.3 Dataverzameling

De data voor dit onderzoek is verzameld in samenwerking met Nationaal Politielab AI. Met behulp van een contactpersoon bij het AI lab zijn diverse dataverzameling strategieën gebruikt. De eerste strategie omvat het gebruik van mailingslijsten binnen de politie en het contact via sociale media kanalen. Na de initieel lage respons is van strategie veranderd. De nieuwe strategie was het direct telefonisch benaderen van teamchefs voor verdere verspreiding binnen hun team; dit werkte erg goed. De enquête is verspreid onder straatagenten, centralisten en agenten in opleiding. Deze groepen hebben direct te maken met het omschreven scenario. De eerste enquête was verstuurd op 22 mei en de data verzameling is gestopt op 21 juni. 288 respondenten hebben gereageerd op de enquête, hiervan hebben 142 respondenten de enquête volledig ingevuld. De lage retentie is mede verklaarbaar doordat de enquête software iemand die op de link klikt direct registreert als respondent, ook al zijn nog geen vragen ingevuld.

Tabel 1.
Dataverzameling strategieën

Tijd	Voornaamste strategieën	Respondenten (cumulatief)
22 mei – 4 juni	Enquête verspreiden via mailinglijsten door contactpersoon Nationaal Politielab AI Berichten naar officiële politie sociale media kanalen	11
5 juni – 21 juni	Persoonlijk belcontact teamchefs en mails met de enquête via hen verspreiden Bij toestemming van teamchef langsgaan bij bureau en pitch houden bij de dagstart	142

Onderzoek naar de invloed van AI-adviezen bij contactambtenaren, met een steekproef bestaande uit contactambtenaren, zijn belangrijk, maar lastig uit te voeren. Vele contactambtenaren, waaronder de politie, hebben een hoge werkdruk. Hierdoor is het uitdagend om hen te bereiken met enquêtes. Alsnog is een steekproef die representatief is voor de populatie, binnen dit onderwerp belangrijk. Dit onderzoek gaat ervan uit dat de opgebouwde ervaringen van politieagenten een rol spelen binnen het overwegend *underreliant* gedragen (Alon-Barkat & Busuioc, 2023; Selten et al., 2023). Contactambtenaren die in het veld werken, hebben de specifieke ervaringen, trainingen en intuïtie die bijdragen aan de bestaande bias (Maynard-Moody & Musheno, 2000; Noordegraaf, 2015; Thomann et al., 2018; Tummers & Bekkers, 2014). Aangezien het onderzoek rekent op deze kwaliteiten, is het gebruik van een steekproef die representatief is voor de populatie essentieel voor een goede externe validiteit en generaliseerbaarheid (Anderson-Cook, 2005).

Tegelijkertijd is het moeilijk om te claimen dat de dataverzameling volledig representatief is. Door de gekozen strategieën heeft de enquête geen volledig willekeurige groep mensen bereikt, waardoor sprake is van een *convenience sample* (Bryman, 2015). Daarnaast is het lastig om een responspercentage te berekenen, omdat het onbekend is hoeveel agenten de enquête daadwerkelijk hebben ontvangen van hun teamchef. Volgens Mutz & Pemantle (2015) is het minder relevant om het responspercentage te berekenen bij een steekproef die niet pretendeert volledig willekeurig te zijn.

Volledige representativiteit is echter bijzonder moeilijk te realiseren in een onderzoekssetting zoals de politie. Veel onderzoeken maken gebruik van studenten of online panels om hypotheses te toetsen. De waarde van een omvangrijke, realistische steekproef van beroepsagenten is daarom aanzienlijk. Hoewel de claim op volledige representativiteit moeilijk te onderbouwen is, draagt de realistische aard van de steekproef sterk bij aan de externe validiteit en generaliseerbaarheid van het onderzoek (Anderson-Cook, 2005).

Binnen het experiment zijn aandachtchecks, een timer of een progressiebalk bewust weggelaten. Aandachtschecks kunnen gezien worden als controlerend. Dit zorgt vervolgens voor valse negatieve antwoorden bij de vraag die de aandacht

checkt (Silber et al., 2022). Daarbij kan de check, net als een timer, de variabelen beïnvloeden. De bias van politieagenten die dit onderzoek veronderstelt komt gedeeltelijk voort uit het feit dat politieagenten doeners zijn. In diverse gesprekken met agenten werd dit beaamd. De verantwoordingsvariabele moet hen binnen het experiment hun bias en acties laten overwegen. Echter, aandachtschecks en een timer kunnen soortgelijke effecten hebben. Het promoot namelijk systematisch denken over casussen om te voorkomen dat ze "*in de val gelokt worden*" (Hauser & Schwarz, 2015). Een progressiebalk is weggelaten omdat de enquête software de langer durende vragen verkeerd vertegenwoordigen met een progressiebalk. Het weglaten ervan voorkomt het misleiden van respondenten. Door de bovenstaande redenen zijn de checks, timers en progressiebalken bewust weggelaten.

De steekproef neemt iedereen mee die de volledige enquête heeft ingevuld en voldoet aan de criteria van een executieve status en het invullen van alle vragen. Van de 142 respondenten die de enquête volledig hebben ingevuld, voldoen 11 respondenten niet aan de criteria van een executieve status. Mensen die niet voldoen aan een executieve status hebben geen ervaring als straatagent en worden daarom uitgesloten van de enquête. Hierdoor komt de totale steekproef uit op 131 respondenten (N = 131).

Tabel 2.
Respondenten

Totaal aan enquête begonnen	Totaal enquête afgemaakt	Totaal enquête afgemaakt en executieve status
288	142	131

3.4 Steekproef samenstelling

In het onderstaande hoofdstuk wordt gekeken naar de representativiteit van de steekproef in vergelijking met populatiedata en wordt de randomisatie gerapporteerd.

3.4.1 Representativiteit ten opzichte van de populatie

Bijlage B laat zien of een representatieve steekproef is afgenomen ten opzichte van de populatiedata van de politie. De twee variabelen die gebruikt

worden om te controleren of de steekproef overeenkomt met de populatiedata zijn geslacht en leeftijd. Ervaring is meegenomen in de descriptieve statistiek, maar hiervoor was geen populatiedata beschikbaar. Ook loopt ervaring veelal gelijk aan leeftijd, daarom wordt het niet hebben van de populatiedata niet van uiterst belang geacht. Bovendien is de beschrijvende variabele van executieve status meegenomen in de tabel.

Uit de toets blijkt dat de steekproef qua geslacht niet volledig representatief is voor de populatiedata, maar qua leeftijd wel. Dit heeft invloed op de externe validiteit van het onderzoek (Anderson-Cook, 2005). De volledige uitwerking van de toets is weergegeven in Bijlage B.

3.4.2 Randomisatie

Om een zuiver causaal verband aan te tonen is het belangrijk dat de randomisatie tussen de verschillende groepen succesvol is. Dit betekent dat de gemiddelden en varianties van de achtergrondvariabelen tussen de controlegroep en de verantwoordingsgroep niet sterk van elkaar verschillen (Anderson-Cook, 2005; Baekgaard et al., 2015; Druckman et al., 2011). De variabelen waar de randomisatiechecks op uitgevoerd worden zijn: geslacht, ervaring, leeftijd en de samengestelde schaal van vertrouwen in technologie. Geslacht en ervaring worden getoetst met behulp van de Chi-kwadraat-goodness-of-fit (GOF) toets. De Chi-kwadraat GOF toets kan significante verschillen tussen de verdeling in groepen van nominale of ordinale data aantonen. Leeftijd en vertrouwen in technologie worden getoetst middels t-toetsen (Field, 2018) Daarmee kunnen de gemiddelden en varianties tussen groepen worden vergeleken.

Tabel 3.
Randomisatie toets

	Scenario 1 (Bergen op Zoom)	Scenario 2 (Sassenheim)
% Vrouw	$\chi 2(1, n = 129) = 0{,}138$, $p = 0{,}711$	$\chi 2(1, n = 129) = 0{,}323$, $p = 0{,}570$
% Meer dan 5 jaar ervaring	$\chi 2(1, n = 131) = 1{,}139$, $p = 0{,}268$	$\chi 2(1, n = 131) = 0{,}332$, $p = 0{,}565$
Leeftijd	$t(129, n = 131) = 1{,}081$, $p = 0{,}056$	$t(129, n = 131) = 1{,}436$, **$p = 0{,}017$***

Vertrouwen in technologie	t(129, n = 131) = 5,192, p = 0,323	t(129, n = 131) = 0,395, p = 0,538

Noot 1. Om anonimiteit te behouden is leeftijd en ervaring gemeten in groepen. De bovenstaande berekeningen zijn gebaseerd op groepsgemiddelden. Leeftijd gebruikte een 5 puntenschaal, ervaring is gecodeerd op basis van een dummy variabel: 5 < year experience = 0, > 5 year experience = 1.
Noot 2. *p < 0.025; **p < 0.005; ***p < 0.0005 (tweezijdig)
Noot 3. Bij het scenario 1 vertrouwen in technologie was de Levene's toets significant, daarom zijn de waardes afgelezen bij *equal variances not assumed* in SPSS (Field, 2018).

Uit tabel 3 blijkt dat de randomisatie grotendeels succesvol is voor de variabelen die een potentieel effect hebben op de hoofdvraag. De variabele leeftijd binnen het scenario Sassenheim is echter niet gelijk verdeeld. Dit blijkt ook uit de gemiddelden tussen de controlegroep (39,3 jaar) en de verantwoordingsgroep (44,1 jaar). De t-toets toont aan dat dit verschil significant is. Desondanks is de randomisatie voor alle andere variabelen, waaronder ervaring, succesvol. Daarom wordt verwacht dat dit geen significant effect zal hebben op de uiteindelijke resultaten.

3.5 Manipulatiecheck

De survey heeft na het eerste scenario een manipulatiecheck ingebouwd om te toetsen of het verschil tussen de verantwoordingsgroep en controlegroep effectief was. Als de manipulatie tussen de verschillende groepen niets beïnvloedde dan hebben de verschillende behandeling in groepen weinig uitgemaakt (Baekgaard et al., 2015). Het toevoegen van een manipulatiecheck voorkomt zo een mogelijke type-II fout (Mutz & Pemantle, 2015). De manipulatiecheck toetst in dit experiment of de mensen in de verantwoordingsgroep echt voelen of denken dat zij verantwoording afleggen. De verantwoordingsgroep hoort zich verantwoordelijk gehouden te voelen en de controlegroep significant minder.

De check rondom verantwoording is gebaseerd op de 'felt accountability' schaal van (Hochwarter et al., 2007). De schaal is in het Nederlands uitgewerkt en aangepast op basis van de enquête. De vragen zijn in de context van AI-adviezen gezet. Een vraag zoals *"Top management holds me accountable for all of my decisions"* (Hochwarter et al., 2007) is vertaald in de enquête naar *"Mijn teamleider houdt mij verantwoordelijk voor mijn besluiten die ik neem op basis van een algoritmisch advies"*. De verandering van topmanagement naar teamleider is

gebaseerd op het feit dat de respondent, in de verantwoordingsgroep, verantwoording aflegt aan de teamleider. Daarbij wordt de context van handelingen gebaseerd op AI-advies toegevoegd. Alle vragen zijn gesteld op basis van een zevenpuntschaal (1 = Helemaal oneens; 7 = Helemaal eens). Dit leidt tot de volgende stellingen die de samengestelde schaal vormen voor de 'afleggen van verantwoording' manipulatiecheck:

- Ik word verantwoordelijk gehouden voor besluiten die ik neem op basis van algoritmische adviezen zoals het HOV
- Ik moet vaak uitleggen waarom ik een bepaalde keuze maak op werk na een algoritmisch advies
- Mijn teamleider houdt mij verantwoordelijk voor mijn besluiten die ik neem op basis van algoritmische advies.

Noot 1. De schaal is getoetst op samenhang, zie hiervoor 3.2.4

In tabel 4 is te zien dat de manipulatie tussen de groepen niet significant is. Dit was een eenzijdige toets, met de verwachting dat de groepen significant van elkaar zou verschillen, waarbij de verantwoordingsgroep een hoger gemiddelde zal hebben (meer verantwoordingsgevoel). In de tabel is het tegenovergestelde te zien. Het gemiddelde van de controlegroep ligt hoger dan de verantwoordingsgroep. De groep die bij het eerste scenario het volgende te zien kreeg: ***"Je teamleider vraagt in dergelijke situaties nooit om je keuzes achteraf uit te leggen."*** Voelden zich op de bovenstaande samengestelde schaal gemiddeld meer verantwoordelijk dan de groep die het volgende te zien kreeg: ***"Bij gebruik van de HOV vindt jouw teamleider het belangrijk dat jij je keuze achteraf verantwoord door uit te leggen waarom jij wel of niet hebt geluisterd naar het algoritme."*** Daarbij moest de verantwoordingsgroep ook daadwerkelijk verantwoording afleggen door een uitleg bij hun antwoord te schrijven. Dit zijn opvallende resultaten.

Ter controle is nog gecontroleerd of alleen de vraag: *"Mijn teamleider houdt mij verantwoordelijk voor mijn besluiten die ik neem op basis van algoritmische advies*" andere resultaten leverde. Deze vraag komt namelijk direct overeen met de toegevoegde zin aan de verantwoordingsgroep, waarin een leidinggevende om verantwoording vraagt. Echter, de resultaten blijven onveranderd t(129, n = 131) = -

0,635, p = 0,263. Al met al blijft de conclusie dat de manipulatie niet geslaagd is. In de discussie zullen deze resultaten verder besproken worden.

Tabel 4.
Manipulatiecheck

	N	Gemiddelde	SD	t-toets
Het eerste scenario				t(129, n = 131) = 0,063, p = 0,475
Controlegroep	70	3,895	1,459	
Verantwoordinggroep	61	3,880	1,336	

Noot 1. *p < 0.05; **p < 0.01; ***p < 0.001 (eenzijdig)

3.6 Onderzoeksethiek

In het uitvoeren van het onderzoek zijn de standaard regels gebruikt rondom data- en onderzoekethiek. De data is verzameld via de surveytool van de politie, Limesurvey. De surveytool is niet toegankelijk voor iemand buiten de politieorganisatie. Verder wordt de verzamelde data van de respondenten geanonimiseerd. Om te voldoen aan het recht om vergeten te worden, wordt de data set een jaar na inlevering van dit onderzoek verwijderd. Vóór het invullen van de enquête, is duidelijk bekend gemaakt dat de deelnemers onderdeel zijn van een onderzoek. In datzelfde onderdeel wordt het doel van het onderzoek duidelijk gemaakt, eveneens dat de ingevulde antwoorden gebruikt worden om een masterscriptie te schrijven. Elke respondent is schriftelijk om toestemming gevraagd of de persoon deel wil nemen aan de vragenlijst. Ook is de optie gegeven om vragen te mailen naar de onderzoeker.

De data wordt opgeslagen in een programma genaamd Yoda. Dit is een datamanagement systeem vanuit de Universiteit Utrecht. Yoda verzekert een veilige opslag van data binnen Europese servers, dat in lijn is met Europese en Nederlandse AVG wetgeving.

Verder worden, in lijn met de onderzoeksethiek, de hypotheses van te voren vastgelegd. Dit is gedaan binnen de pre-registratie in Open Research (https://osf.io/wkxyf/). Dit verzekerd dat de hypotheses niet aangepast kunnen worden na het analyseren van de resultaten uit de enquête.

3.7 Kwaliteitscriteria: Validiteit en betrouwbaarheid

Twee belangrijke kwaliteitscriteria voor een goed experiment zijn validiteit en betrouwbaarheid. Onderstaand wordt gereflecteerd hoe dit onderzoek om is gegaan met deze kwaliteitscriteria.

3.7.1 Validiteit

Validiteit houdt het volgende in: Meet het onderzoek datgene dat het wilt meten (Bryman, 2015)? Validiteit bestaat uit vier verschillende soorten statistische, interne, construct en externe validiteit (Anderson-Cook, 2005).

Statistische validiteit houdt in dat de uitkomsten van een onderzoek gebaseerd zijn op een goede en secure data analyse. Hierbij worden de juiste statistische methoden gebruikt (García-Pérez, 2012). Om dit te bewaken zijn de aannames achter de verschillende uitgevoerde toetsen gecontroleerd. Daarbij worden factoranalyses en Chonbach's Alpha gebruikt om de schaalvaliditeit van variabelen te toetsen. Ook worden de statistische bevindingen in de conclusie en discussie onderbouwd met literatuur. Dit borgt de statistische validiteit (García-Pérez, 2012).

De interne validiteit draait om het kunnen vaststellen van een causaal verband (Bryman, 2015). Hiervoor moet een onderzoek aan drie belangrijke voorwaarden voldoen, oorzaak voor gevolg, covariatie en geen alternatieve verklaringen. Een experiment heeft over het algemeen een sterke interne validiteit, door het exclusief veranderen van een enkele variabele tussen groepen en randomisatie (Anderson-Cook, 2005). Oorzaak en gevolg worden bewaakt door de plaatsing van manipulatie van de onafhankelijke variabele vóór de meting van de afhankelijke variabel. Covariatie wordt bewaakt door uitsluiting van alternatieve verklaringen voor de relatie. Daarom wordt in een experiment maar een enkele variabele, namelijk verantwoording aangepast (McDermott, 2002). Voor de mogelijkheid tot geen alternatieve verklaringen, is het van belang dat de random toekenning van groepen geslaagd is. Dit vangt de mogelijkheid op dat externe variabelen de verklaring zijn voor het wel of niet vinden van een verband (McDermott, 2002). Omdat het experiment twee scenario's gebruikt, elk met random

toedeling in groepen bestaat het risico dat men testeffecten laat zien, ofwel ze gedragen zich anders omdat ze in een experiment zitten. De respondent doet dan wat van hen verwacht wordt (Mummolo & Peterson, 2019). In dit geval kunnen de respondenten kennis op doen over de potentiële hypothese door het verschil tussen scenario 1 en 2 te zien als ze ook in andere groepen worden geplaatst. Echter is bewezen dat zelfs met kennis van de hypothese testeffecten amper voorkomen (Mummolo & Peterson, 2019). De volledige randomisatie intact houden werd waardevoller geacht dan het risico van testeffecten.

De construct validiteit houdt in dat het experiment een sterke link heeft met de theoretische concepten (Anderson-Cook, 2005; Bryman, 2015). De onderbouwing hiervoor is terug te vinden in het theoretisch kader en het methode hoofdstuk. Deze hoofdstukken laten uitgebreid zien hoe het experiment onderbouwd wordt door de theoretische achtergrond. Een belangrijke beperking in dit experiment is de versimpeling van *appropriate reliance* tot een binair waarschijnlijk of onwaarschijnlijk antwoord. De theoretische onderbouwing geeft weer dat *appropriate reliance* zich meer voordoet als een uitgebreid spectrum met gradaties van juistheid van de adviezen van AI-systemen en de besluiten daarvan (Benda et al., 2022; Bruijnes et al., 2024; Schemmer et al., 2023). Dit was een bewuste keuze voor de uitvoerbaarheid van het experiment. Bovendien zijn de aannames achter ‘waarschijnlijkheid’ en ‘onwaarschijnlijkheid’ van de antwoorden getoetst bij straatagenten die hebben bevestigd dat het terechte aannames zijn.

De externe validiteit is de generaliseerbaarheid van het experiment naar de onderzoekspopulatie (Anderson-Cook, 2005; Bryman, 2015). Dit bestaat ten eerste uit de ecologische validiteit, namelijk of het experiment gegrond is in de realiteit. Zoals eerder is benoemd zijn diverse maatregelen genomen om de ecologische validiteit te garanderen. De enquête is getoetst op de realisme bij twee straatagenten, vier datawetenschappers van het Nationaal Politielab AI, een hoogleraar en meerdere universiteitsstudenten. Externe validiteit bestaat ten tweede uit generaliseerbaarheid, ofwel houdt de causale relatie stand in andere domeinen. Aangezien de casus zich specifiek binnen een politie domein afspeelt, kunnen hier geen garanties voor gegeven worden. Tegelijkertijd kan dit onderzoek

dienen als springplank om soortgelijke onderzoeken uit te voeren in andere domeinen binnen en buiten de publieke sector. Hier wordt verder op ingegaan in de discussie.

3.7.2 Betrouwbaarheid

Betrouwbaarheid gaat over de consistentie van het onderzoek. Onder betrouwbaarheid vallen de criteria herhaalbaarheid, stabiliteit en interne betrouwbaarheid (Bryman, 2015, p. 157). Herhaalbaarheid kijkt naar de vraag of het onderzoek herhaald kan worden en of dit dezelfde resultaten oplevert bij herhaling. Hier ligt de focus op de uiteindelijke consistentie van de resultaten. De stabiliteit kan bewaakt worden door het onderzoek herhaalbaar te maken, ofwel zijn de achterliggende keuzes en gebruikte materialen goed uitgelegd in het methode hoofdstuk? Daarom is de volledige enquête in Bijlage F toegevoegd en wordt in het methode hoofdstuk uitgelegd hoe het onderzoek is uitgevoerd.

De interne betrouwbaarheid kijkt naar de samenhang van samengestelde schalen, ofwel vragen, die in de enquête gesteld worden. Dit wordt op drie manieren bewaakt. Ten eerste worden aan elke respondent twee scenario's voorgelegd. Zo worden meerdere keren dezelfde variabelen in een ander scenario gemeten. Ten tweede, worden theoretisch onderbouwde schalen voor verantwoording en vertrouwen in technologie gebruikt (Chi et al., 2021; Hochwarter et al., 2007). Als laatste wordt nog een exploratieve factor analyse op deze schalen losgelaten om te zien of alle vragen op dezelfde factor laden.

4. Resultaten

In het onderstaande hoofdstuk worden de toetsen uitgevoerd die antwoord geven op de hoofdvraag. De data die gebruikt wordt in dit hoofdstuk is verzameld middels het eerder benoemde vragenlijstexperiment. De analyses en statistische toetsen zijn uitgevoerd middels het statistiekprogramma SPSS. Eerst wordt stilgestaan bij de beschrijvende statistiek. Daarna wordt gekeken naar het antwoord op de hoofdvraag, ofwel H1. Vervolgens wordt het modererende effect getoetst van H2. Als laatste wordt stilgestaan bij een aantal exploratieve analyses.

4.1 Beschrijvende statistiek

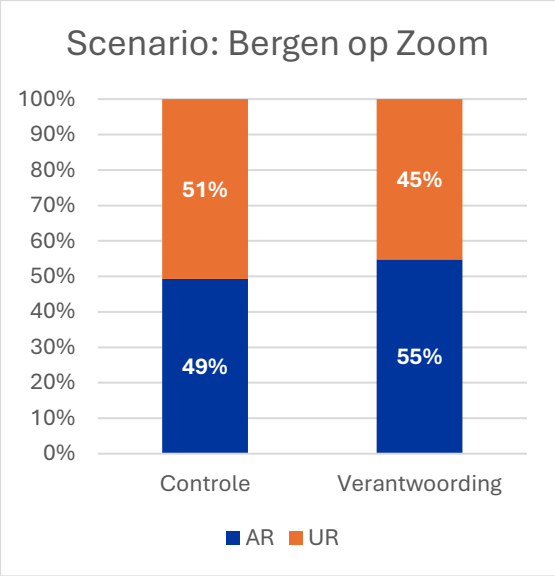
Voordat gericht wordt op de hoofdvraag is het van belang om de beschrijvende statistiek te behandelen. Dit geeft een beginnend beeld van de karakteristieken van de dataset. Specifiek wordt gekeken naar de verdeling van agenten die hebben gekozen om het advies van de HOV op te volgen en de agenten die ervoor gekozen hebben om het advies naast zich neer te leggen. Daarnaast wordt stilgestaan bij de keuzezekerheid tussen de verschillende groepen.

In figuur 7 wordt de verdeling tussen *appropriate reliant* (AR) agenten en *underreliant* (UR) agenten geschetst in het scenario Bergen op Zoom. Hier valt op dat de verdeling tussen AR en UR vrijwel gelijk is tussen beide groepen. Alsnog kantelt de verdeling licht richting a*ppropriate reliant* in de verantwoordingsgroep. Figuur 8 schetst de verdeling tussen *appropriate reliant* (AR) agenten en *underreliant* (UR) agenten in het scenario Sassenheim. De verdeling tussen de groepen gaat hier nog gelijker op. In tegenstelling tot het Bergen op Zoom scenario, is in dit scenario de verantwoordingsgroep voor een groter deel *underreliant*.

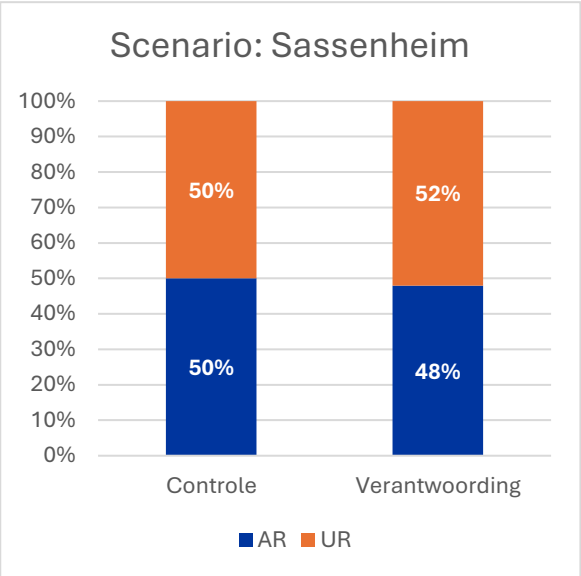
Tabel 5 geeft de gemiddeldes en standaarddeviatie weer van de mate waarin de agent zeker was over hun keuze. Hierin kreeg de agent de mogelijkheid om te kiezen uit een tienpuntschaal (1 = onzeker; 10 = zeker). Het valt op dat de groepen die verantwoording afleggen gemiddeld zekerder zijn van hun keuzes, dan de groepen die geen verantwoording afleggen. Tegelijkertijd zijn geen significante verschillen aanwezig tussen de groepen. Een uitschieter in deze trend is de groep die het AI-advies niet heeft opgevolgd bij het scenario Bergen op Zoom. Bovendien

lijken de respondenten die het AI-advies hebben opgevolgd over het algemeen minder zeker te zijn van hun keuze, behalve in de verantwoordingsgroep bij het scenario Bergen op Zoom.

Deze beschrijvende resultaten geven belangrijke context voor de hoofdvraag toetsen die in het volgende hoofdstuk behandelt worden.



Figuur 7. Keuze agenten scenario BoZ



Figuur 8. Keuze agenten scenario Sas

Tabel 5.
Gemiddelde van zekerheid van keuze (SD tussen haakjes)

	Scenario	
	Scenario 1. (Bergen op Zoom)	Scenario 2. (Sassenheim)
Controlegroep		
Underreliant	5,43 (1,929)	6,20 (1,605)
Appropriate reliant	5,32 (1,870)	5,79 (1,934)
Totaal controlegroep	5,38 (1,887)	6,02 (1,759)
Verantwoordingsgroep		
Underreliant	4,82 (1,806)	6,29 (1,844)
Appropriate reliant	6,15 (1,617)	6,03 (1,899)
Totaal verantwoordinggroep	5,55 (1,817)	6,18 (1,858)
Verschil toetst (t-toets)	t(129, n = 131) = 0,005, p = 0,598	t(129, n = 131) = 0,295, p = 0,606

Noot 1. *p < 0.025; **p < 0.005; ***p < 0.0005 (tweezijdig)

4.2 Effect van verantwoording op underreliance

Om H1 (Het afleggen van verantwoording door politieagenten aan een prominent publiek (zoals een teamleider) draagt bij aan het voorkomen van underreliance op AI-advies) te toetsen zijn drie Chi-kwadraat-goodness-of-fit (GOF) toetsen gebruikt. Voor elk scenario (Bergen op Zoom en Sassenheim) is apart een GOF toets uitgevoerd. Daarbij is er een aparte GOF toets uitgevoerd die beide scenario's samen neemt. De GOF toetsen vergelijken de controlegroep en de verantwoordingsgroep. Specifiek worden de frequenties vergeleken waarin een agent heeft besloten om het algoritmisch advies te volgen (appropriate reliant) of heeft besloten om het algoritmisch advies niet op te volgen (underreliant).

Tabel 6.
Verschil *appropriate reliance* tussen controle- en verantwoordinggroep

Scenario's	Chi-kwadraat toetsen
Bergen op Zoom	$\chi 2(1, n = 131) = 0{,}405, p = 0{,}525$
Sassenheim	$\chi 2(1, n = 131) = 0{,}055, p = 0{,}815$
Beide scenario's	$\chi 2(1, n = 262) = 0{,}057, p = 0{,}812$

Noot 1. *p < 0.05; **p < 0.01; ***p < 0.001 (eenzijdig)

H1 verwachtte dat het afleggen van verantwoording bijdraagt aan het voorkomen van *underreliance*. Echter, de statistische toetsen uit tabel 6 laten zien dat geen enkel van de scenario's een significant verschil aantoont, dit met een power van 0,9. Aangezien geen significant resultaat aan te tonen is bij elk van de scenario's, wordt H1 verworpen.

Uit de crosstabs in bijlage C valt af te lezen dat de verschillen tussen groepen heel klein zijn. Dit houdt in dat de manipulatie tussen groepen, het afleggen van verantwoording, amper effect heeft gehad op het wel of niet opvolgen van het advies van de HOV. Zowel in de controle als de verantwoordingsgroep legt ongeveer de helft van de agenten het AI-advies naast zich neer en is daarmee *underreliant* op het AI-systeem. Dit ondanks, dat het AI-systeem het waarschijnlijk juiste antwoord gaf.

4.3 Modererend effect van ervaring op de relatie

Naast de eerste hypothese was een tweede hypothese opgesteld voor een negatief modererend effect, namelijk H2 (*Ervaring binnen de politie heeft een negatief effect op de relatie tussen verantwoording en underreliance*). Om H2 te toetsen wordt een Cochran-Mantel-Haenszel toets gebruikt. Dit houdt in dat een tweede toets wordt uitgevoerd die kijkt of de relatie tussen *appropriate reliance* en het wel of niet afleggen van verantwoording wordt beïnvloed door de indeling naar verschillende mate van ervaring. De ervaringscategorieën die worden aangehouden zijn: 5 jaar of minder en 20 jaar of meer. De keuze voor deze categorieën hangt samen met het feit dat bij meer categorieën het ordinale data is. In dat geval kan de volgorde van de categorieën niet meegenomen worden binnen de Cochran-Mantel-Haenszel toets. 5 jaar en 20 jaar liggen ver uit elkaar zodat men voldoende ervaring heeft kunnen opbouwen. De scenario's zijn niet apart getoetst om een te kleine steekproef te voorkomen.

Het resultaat uit Cochran-Mantel-Haenszel toets is $\chi 2(1, n = 162) = 1{,}532$, $p = 0{,}216$. Dat houdt dat H2 kan worden verworpen. Werkervaring binnen de politie beïnvloedt daarmee de relatie tussen verantwoording en *appropriate reliance* niet. De toets heeft een power van 0,72. Deze uitkomst is niet verrassend op basis van de resultaten uit 4.2. Hierin wordt geconstateerd dat er geen relatie is tussen verantwoording afleggen en *appropriate reliance*. Daarmee is het logisch dat ervaring deze relatie amper beïnvloedt, aangezien een niet bestaande relatie over het algemeen weinig wordt beïnvloedt.

Naast het modererende effect kan gekeken worden naar een effect tussen ervaring en *appropriate reliance*. Hiervoor is noch een eerdere hypothesis opgesteld noch is hiervoor een manipulatie voor uitgevoerd binnen het experiment. Een eventueel effect is daarom geen causaal effect, maar een correlerend effect. Het resultaat uit de chi-kwadraat toets tussen deze variabele is: $\chi 2(1, n = 162) = 3{,}247$, $p = 0{,}072$. Met een betrouwbaarheidsinterval van 95% houdt dit in dat er geen significant effect is tussen ervaring en *appropriate reliance*. Het valt op dat het resultaat bijna significant is. Nog opvallender is het feit dat de 5 jaar en onder

categorie bijna twee keer zoveel *underreliant* respondenten heeft als *appropriate reliant* respondenten. Dit is terug te zien in bijlage D.

4.4 Exploratieve analyses

In dit deelhoofdstuk worden naast de hoofdvragen nog een aantal exploratieve analyses uitgevoerd. Met de beschikbare data is het interessant om deze exploratieve analyses uit te voeren, zodat toekomstige onderzoeken hier verder op kunnen voortbouwen. Voor deze analyses zijn geen hypotheses opgezet, daarom blijven de analyses exploratief en worden er geen harde conclusies voor dit onderzoek uitgetrokken.

4.4.1 Tijd

De eerste exploratieve analyse is de invloed van tijd op *appropriate reliance*. De reden waarom naar tijd wordt gekeken, is omdat een van de effecten van verantwoording afleggen, meer overwogen keuzes zijn (Aleksovska et al., 2019; Ličen & Slapničar, 2022; Tetlock, 1983). Een van de manieren om dit te meten, is de tijd die participanten nemen om een keuze te maken. Daarom is het interessant om te kijken of agenten die langer de tijd nemen om hun keuze te maken 'beter overwogen' keuzes maken en dus meer *appropriate reliant* zijn.

De variabele tijd is de gespendeerde tijd van een respondent bij een vragengroep. In dit geval de vragengroep waarbij de deelnemer een locatie moest kiezen en verantwoording moest afleggen. De data voor alleen de locatie keuze was niet beschikbaar, daarom is de gehele groep, inclusief het afleggen van verantwoording meegenomen. Bovendien zijn deelnemers die langer dan tien minuten over de vraag hebben gedaan weggelaten uit de analyse. Dit voorkomt uitschieters die de gemiddelden vergelijkingen onbruikbaar maken (Field, 2018).

Tabel 7 toont aan dat er geen significante verschillen zijn tussen de groepen van agenten die *appropriate reliant* zijn en die *underreliant* zijn. Het lijkt er daarmee op dat in dit onderzoek tijd gespendeerd bij een vragengroep geen invloed heeft gehad op *appropriate reliance*. Daarbij valt op dat de gemiddelde tijd in de UR groep hoger is dan in de AR groep. Dit is verassend, want in overeenkomst met de literatuur wordt verwacht dat de agenten die *appropriate reliant* zijn meer

overwogen keuzes maken en dus langer de tijd nemen. Hier zal verder op ingegaan worden in de discussie.

Tabel 7.
Verschil in *tijd* tussen *Appropriate reliant* en *Underreliant* groepen

Scenario's	Gemiddelde in seconde (SD tussen haakjes)		t-toetsen
	AR	**UR**	
Bergen op Zoom	174,209 (112,630)	186,632 (112,972)	t(127, n = 129) = -0,625, p = 0,267
Sassenheim	149,940 (90,901)	164,134 (104,945)	t(126, n = 128) = -0,802, p = 0,212
Beide scenario's	163,142 (103,601)	183,165 (129,531)	t(257, n = 259) = -1,368, p = 0,086

Noot 1. *p < 0.025; **p < 0.005; ***p < 0.0005 (tweezijdig)

4.4.2 Vertrouwen in technologie

De tweede exploratieve analyse is de invloed van vertrouwen in technologie op *appropriate reliance*. Hiervoor wordt de eerder opgezette vertrouwen in technologie schaal gebruikt. Deze exploratieve analyse wordt uitgevoerd, omdat algemeen vertrouwen in technologie bij kan dragen aan meer *appropriate reliant* zijn (Alon-Barkat & Busuioc, 2023; Selten et al., 2023). De kans bestaat dat een agent die technologie meer vertrouwt binnen dit experiment sneller *appropriate reliant* is in plaats van *underreliant*.

Deze verwachting wordt niet ondersteunt door de resultaten in tabel 8 De tabel laat zien dat er geen significant verschil is op het hebben van meer vertrouwen op *appropriate reliance*. In het Bergen op Zoom scenario, scoort de *underreliance* groep zelfs beter op vertrouwen dan de *appropriate reliance* groep. Dit draait om in het Sassenheim scenario. Wanneer beide scenario's samen worden genomen verschillen de uiteindelijk gemiddelde vertrouwens nauwelijks van elkaar.

Tabel 8.
Verschil in *vertrouwen in technologie* tussen Appropriate Reliant en Underreliant groepen

Scenario's	Gemiddelde op 7 puntschaal (SD tussen haakjes)		t-toetsen
	AR	**UR**	

Bergen op Zoom	4,868 (1,201)	5,064 (1,185)	t(129, n = 131) = -0,938, p = 0,175
Sassenheim	5,132 (1,131)	4,827 (1,231)	t(129, n = 131) = 1,436, p = 0,073
Beide scenario's	4,989 (1,172)	4,936 (1,211)	t(260, n = 262) = 0,360, p = 0,359

Noot 1. *p < 0.025; **p < 0.005; ***p < 0.0005 (tweezijdig)

4.4.3 De gegeven verantwoording

De laatste exploratieve analyse die is uitgevoerd, is een codering van de verantwoording die de respondenten hebben gegeven bij hun locatiekeuze. Deze codering geeft diepere inzicht, in waarom de agenten de keuzes gemaakt hebben die zij gemaakt hebben.

Het eerste dat opvalt is dat de gegeven verantwoording veelal de redenering van de HOV volgt. Figuur 9 geeft dit weer, de rest van de figuren zijn te vinden in Bijlage E. In de codering is onderscheidt gemaakt tussen de verschillende aarden van de verantwoordingen. Deze aarden zijn: Het opschrijven van redenen die de HOV in de casus aangeeft, een eigen andere ingeving of *"het is de beste keuze"* ofwel geen gegeven reden. Tussen dit onderscheidt valt op dat veel agenten wel expliciet de redeneerlijn volgen van de HOV in hun antwoorden. Ook als ze uiteindelijk niet de aanbevolen locatie kiezen. De HOV suggereert twee locaties, maar doet een expliciete aanbeveling voor één locatie. In de suggesties staan wel een uitleg waarom de locaties worden gesuggereerd. Ter illustratie, deze agenten hebben het advies van de HOV niet opgevolgd, maar wel dezelfde redeneerlijn gevolgd voor de gesuggereerde locatie:

"*Doordat je samen het gebied afsluit heb je een grotere pak kans. Als iedereen zich aan zijn/haar taak houdt is de pak kans dus het grootst.*" (Scenario Bergen op Zoom)

"*Omdat locaties B en C gedekt werd door collega's en ik denk dat een N weg ook een serieuze vluchtroute is.*" (Scenario Bergen op Zoom)

In dit scenario suggereert de HOV ook een N-weg die verder verspreid is van andere agenten, maar doet uiteindelijk een aanbeveling voor de andere locatie, omdat jouw eenheid sneller op de belangrijkere locatie is. Ondanks dat de agenten

niet de expliciet aanbevolen locatie hebben gekozen, lijkt het alsof de uitleg achter de suggesties serieus wordt meegenomen in de afweging.

Tegelijkertijd lijken veel agenten keuzes te maken op basis van hun eigen kennis. Hiervoor worden per scenario diverse redenen gegeven, maar er wordt frequent beroep gedaan op hun eigen kennis voor de verantwoording. Er wordt dan gerefereerd naar plaats bekendheid of er wordt een generalisatie gedaan die toepasbaar is op deze situatie:

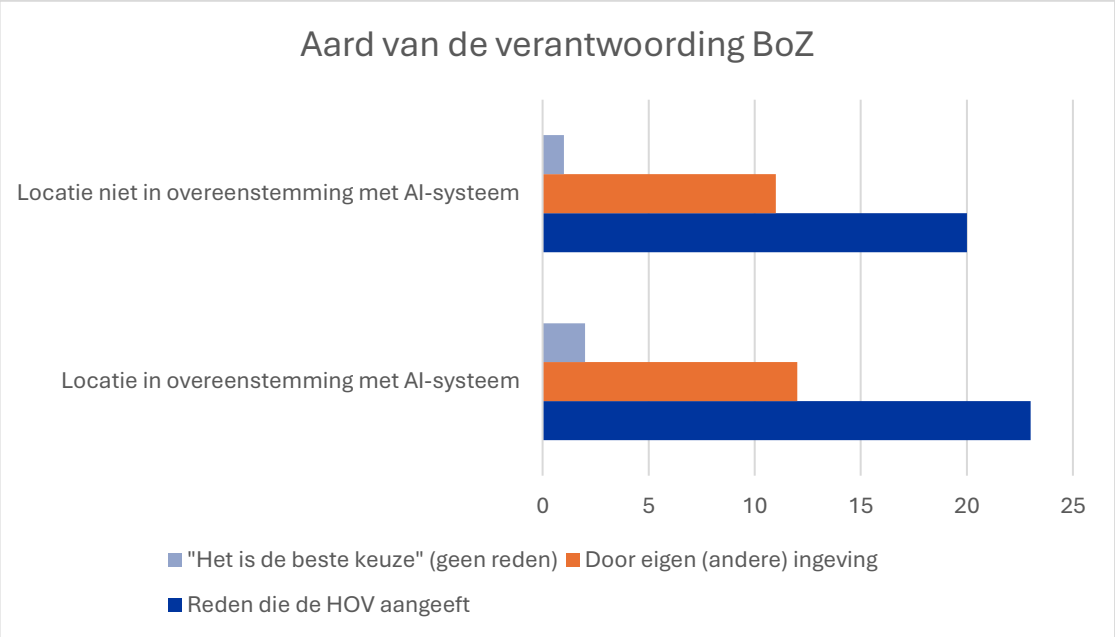
"Mijn ervaring is dat verdachten het liefst zo snel mogelijk weg willen komen van plaats incident. Een snelweg is daarin de snelste optie. Dat de brug open staat weten verdachten vaak niet. Indien de verdachten dit ontdekken, kunnen ze altijd terugkomen over dezelfde route en ga ik ze treffen." (Sassenheim)

*"**Ik zou kiezen voor** een minder opvallende weg, **mij lijkt** dat de verdachte niet wilt opvallen."* (Bergen op Zoom)

Agenten die de HOV opvolgen geven weer dat zij of eens zijn met het systeem, veelal ook voor redenen die de HOV aanhaalt. Een groep agenten noemt de HOV hierbij expliciet en geven soms zelfs aan dat zij vertrouwen op het systeem:

"Ik zie geen reden om af te wijken van de HOV. Lijkt ook logisch om zo dicht mogelijk bij de snelweg positie in te nemen." (Scenario Bergen op Zoom)

"Ik vertrouw op de HOV." (Scenario Bergen Op Zoom)



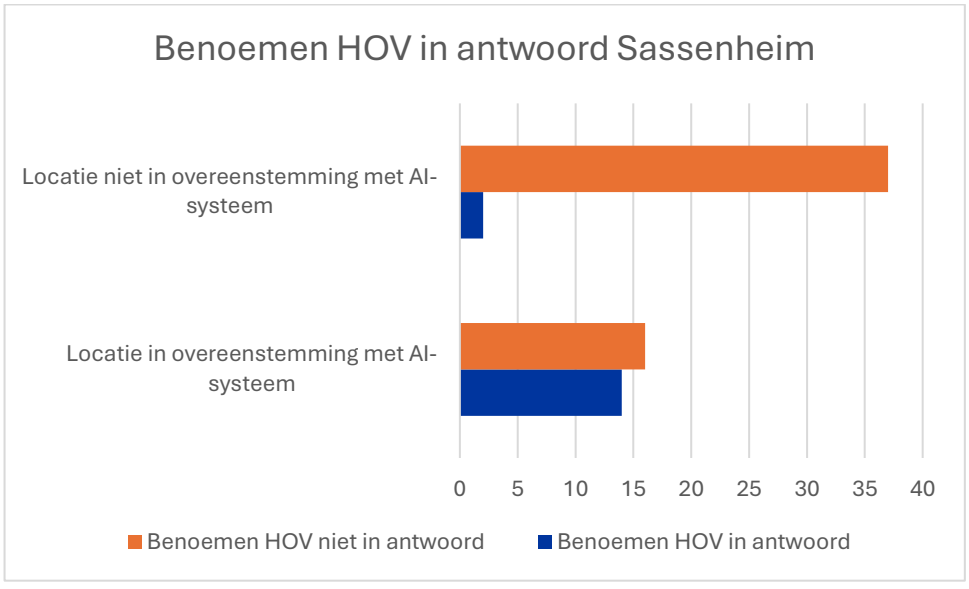
Figuur 9. Aard van verantwoording in scenario Bergen op Zoom.

Het tweede dat opvalt aan de reacties is dat weinig agenten de HOV, het systeem of het algoritme expliciet benoemen in hun antwoord. Dit geldt voor agenten die in overeenstemming zijn met de HOV, maar nog meer voor agenten die niet in overeenstemming zijn met de HOV. Zoals eerder geconcludeerd volgend veel agenten dezelfde redeneerlijn als de HOV, maar ze benoemen het systeem niet expliciet in hun antwoord. De agenten die niet in overeenstemming zijn met de HOV, focussen vooral op hun redenering en waarom zij hebben gekozen voor een locatie. Echter, zij zetten zich hiermee niet expliciet af van de HOV aanbevolen locatie:

"Ik heb gekozen voor locatie B omdat deze locatie ook de route naar locatie A en de snelweg afdekt." (Scenario Sassenheim)

"*locatie B wordt door andere auto gedekt volgens meldkamer.*" (Scenario Bergen op Zoom)

Een aanzienlijk deel van de respondenten die in overeenstemming met de HOV voor een locatie heeft gekozen, verwijzen ook niet expliciet naar de HOV. Simpel gezegd, het dialoog in de verantwoording mist. De agenten beginnen niet met: *"De HOV heeft gelijk/ongelijk, want … "*. De antwoorden draaien om waarom de agenten iets potentieels goed hebben, in plaats van waarom de aanbeveling van de HOV goed of fout is.



Figuur 10. Benoemen van HOV/algoritme/computersysteem binnen gegeven verantwoording

4.5 Samenvattende conclusie resultaten

In dit deelhoofdstuk wordt stilgestaan bij de behaalde resultaten binnen dit onderzoek en de opgestelde hypotheses. De beschrijvende statistiek en resultaten uit de toetsen geven patronen weer. Deze patronen en de conclusies rondom de hypotheses worden hieronder samengevat. De (onderzoekers)interpretatie van de resultaten en de consequenties voor de hoofdvraag worden behandeld in het discussie hoofdstuk.

Ten eerste geven de resultaten weer dat het afleggen van verantwoording geen effect heeft op het voorkomen van *underreliance*. Hiermee is H1 afgewezen. Uit de beschrijvende statistiek viel op dat de groepen een relatief gelijke verdeling hadden tussen agenten die het algoritmische advies op hebben gevolgd en agenten die het advies naast zich neer hebben gelegd. Tegelijkertijd tonen de agenten uit de verantwoordingsgroep meer vertrouwen in hun eigen antwoord dan de agenten die geen verantwoording af hebben gelegd. De uitzondering op deze regel zijn de agenten die het advies niet hebben nageleefd in het Bergen op Zoom scenario. Sterker nog, ondanks dat het algoritme hun keuze bevestigd, hebben agenten die het advies van het algoritme opvolgen over het algemeen minder vertrouwen in hun eigen antwoorden (weer m.u.v. het *underreliant* BoZ scenario).

Ten tweede, bleek ervaring geen modererend effect te hebben op de relatie tussen het afleggen van verantwoording en *appropriate reliance*. Daarmee is H2 afgewezen. Na verdere toetsing viel op dat ervaring bijna correleerde met *appropriate reliance*. Namelijk minder ervaring leidt tot meer *underreliance*. Ook dit was geen significant resultaat, maar het was opmerkelijk dat agenten met minder ervaring, vaker besloten het advies van het algoritme naast zich neer te leggen dan agenten met meer ervaring.

Als laatste zijn een aantal exploratieve analyses uitgevoerd. Deze analyses geven extra context bij de hoofdresultaten en bieden aanknopingspunten voor toekomstig onderzoek. Opvallend is dat zowel de gespendeerde tijd bij de blokkade locatievraag als vertrouwen in technologie geen effect hebben op *appropriate reliance*. Daarbij zijn in Tabel 9 enkele bevindingen genoteerd die opvielen uit de resultaten van de codering van de gegeven verantwoording.

Tabel 9.
Samenvatting van resultaten

Concept	Hypothese	Bevindingen
Appropriate reliance & Verantwoording afleggen	H1. Het afleggen van verantwoording door politieagenten aan een prominent publiek (zoals een teamleider) draagt bij aan het voorkomen van *underreliance* op AI-advies.	Afgewezen
Appropriate reliance, Verantwoording afleggen en werkervaring	H2. Ervaring binnen de politie heeft een negatief effect op de relatie tussen verantwoording en *underreliance*.	Afgewezen
Exploratieve analyses	Geen opgestelde hypotheses	
Tijd	Besteden tijd aan de locatiekeuze heeft in dit onderzoek geen effect op *appropriate reliance*.	
Vertrouwen in technologie	Vertrouwen in technologie heeft in dit onderzoek geen effect op *appropriate reliance*.	
Codering verantwoording	De agenten besteedde aandacht aan de gegeven uitleg van de HOV en namen het mee in hun antwoorden. Sommige agenten leken zelfs te ‘vertrouwen op de HOV’. Een aanzienlijk deel van de agenten gaven eigen unieke redenen om wel of niet in overeenstemming te zijn met het algoritme. De agenten benoemde de HOV weinig in hun antwoorden en gingen in hun verantwoording niet in ‘dialoog’ met de HOV.	

5. Discussie

AI-systemen krijgen een steeds prominentere rol binnen de samenleving. Vanuit die rol beïnvloeden algoritmen, onder andere door adviezen vanuit AI-systemen, de complexe besluitvormingsmechanismen van contactambtenaren (Bovens & Zouridis, 2002; J. Bullock et al., 2020; de Boer & Raaphorst, 2023; Meijer et al., 2021; Peeters, 2020; Young et al., 2019). Het is van belang dat de contactambtenaren die deze AI-adviezen ontvangen, goede besluiten kunnen nemen op basis van deze adviezen (Benda et al., 2022; Bruijnes et al., 2024; Young et al., 2019). Dit houdt in dat ze niet blindelings vertrouwen op AI-advies, maar de adviezen ook niet altijd naast zich neerleggen. Men spreekt dan van *appropriate reliance* (Benda et al., 2022; Bruijnes et al., 2024; Lee & See, 2004; Schemmer et al., 2022). Het doel van dit onderzoek is daarmee een bijdrage te leveren aan betere besluitvorming met AI-adviezen door het creëren van randvoorwaarden waaraan voldaan kan worden aan *appropriate reliance*. De randvoorwaarde die centraal staat in dit onderzoek is het afleggen van verantwoording over de besluiten die contactambtenaren nemen op basis van AI-adviezen. De verwachting is dat verantwoording afleggen beter overwogen keuzes teweegbrengt (Aleksovska et al., 2019; Overman et al., 2021; Tetlock, 1983). Hiermee wordt gezocht naar een antwoord op de volgende hoofdvraag: *Wat is het effect van het afleggen van verantwoording door politieagenten op gepast vertrouwen in AI-systemen bij besluitvormingsprocessen met AI-adviezen?*

De hoofdvraag is onderzocht middels een vragenlijstexperiment waarbij gebruik is gemaakt van vignetten. De vragenlijst is door straatagenten ingevuld. De agenten moesten blokkades plaatsen om vluchtroutes af te sluiten met behulp van algoritmisch advies van een fictief systeem (de HOV). Binnen de vragenlijst zijn twee verschillende groepen gedefinieerd: een groep die verantwoording moest afleggen en een controlegroep die geen verantwoording hoefde af te leggen. Het experiment leverde de volgende resultaten op: (1) Het afleggen van verantwoording draagt niet bij aan het voorkomen van *underreliance* op AI-advies. (2) Meer werkervaring binnen de politie beïnvloedt de relatie tussen het afleggen van verantwoording en *appropriate reliance* niet. Daarnaast zijn exploratieve analyses uitgevoerd die

verdere resultaten hebben opgeleverd. Deze bevindingen hebben implicaties voor besluitvorming op basis van AI-adviezen binnen contactambtenaarberoepen.

De eerste implicatie is dat het consistent bereiken van *appropriate reliance* bij contactambtenaren een uitdaging blijft (Bruijnes et al., 2024) Op basis van de literatuur was de verwachting dat het afleggen van verantwoording zou resulteren in meer weloverwogen besluiten, waarbij eventuele bias zou worden heroverwogen (Aleksovska et al., 2019; Tetlock, 1983). Het afleggen van verantwoording moest daarom bijdragen aan het voorkomen van underreliance op een AI-systeem dat een waarschijnlijk juiste aanbeveling geeft. De resultaten van het uitgevoerde experiment toonde dit effect niet aan. Tegelijkertijd, het falen van de manipulatiecheck liet zien dat er geen verschil was in verantwoordingsgevoel tussen de verantwoordings- en controlegroep (de verdere beperking hiervan wordt besproken in de beperkingen). Hierdoor kan geen uitsluitsel worden gegeven over de rol van verantwoording die wel of niet bijdraagt aan het bereiken van *appropriate reliance*. Desalniettemin kan op basis van de beschrijvende statistiek, ondersteunt door literatuur, worden geconcludeerd dat *underreliance* voorkomen lastig blijft (Bruijnes et al., 2024; Schemmer et al., 2022; Selten et al., 2023).

De tweede implicatie is dat werkervaring geen rol speelt in het versterken of verzwakken van de relatie tussen verantwoording afleggen en *underreliance*, maar dat *underreliance* voor kan komen bij werknemers van lage en hoge ervaringsniveaus. Contactambtenaren worden gevormd door de kennis, normen en waarden van hun omgeving (Maynard-Moody & Musheno, 2000; Noordegraaf, 2015). Ervaring bouwt dit verder op en kan vervolgens opvattingen en bias versterken (Alon-Barkat & Busuioc, 2023; Selten et al., 2023). Op basis van deze achtergrond is in het theoretisch kader voorspeld dat het voor de verantwoording-variabele moeilijker wordt om bij te dragen aan het opvolgen van AI-adviezen in plaats van eigen ingevingen. In tegenstelling tot de hypothese, blijkt de relatie tussen verantwoording afleggen en *appropriate reliance* niet aangetast te worden door ervaring. Ook bij deze implicatie is het belangrijk om te onthouden dat de manipulatiecheck niet geslaagd was en de verantwoordingsvariabele minder sterk was dan verwacht. Bovendien, uit de beschrijvende statistiek in de resultaten blijkt

dat een groter aandeel minder ervaren agenten het algoritmische advies naast zich neerlegt. Hoewel dit verschil niet significant is, suggereert het dat *underreliance* op algoritmische adviezen niet afhankelijk hoeft te zijn van het ervaringsniveau.

Een derde implicatie betreft de manier waarop agenten omgaan met het geven van verantwoording bij besluitvorming op basis van AI-adviezen. De respondenten die verantwoording moesten afleggen, vertoonden geen significante verschillen in de tijd die zij besteedden aan de locatiekeuzevraag; het zijn van *appropriate reliant* of *underreliant* had hierop geen effect. Het idee dat *appropriate reliant* agenten bedachtzamer zijn, bleek dus niet uit de bestede tijd bij de vragen (Aleksovska et al., 2019; Ličen & Slapničar, 2022; Skitka et al., 2000; Tetlock, 1983). In de verantwoording die agenten gaven, leken zij niet in 'dialoog' te treden met het AI-systeem. Ondanks dat Nederlandse agenten AI-advies meer zien als een 'collegiaal' advies, richtten de antwoorden van de agenten zich voornamelijk op waarom hun eigen idee juist was (Meijer et al., 2021). Het AI-systeem werd in hun verantwoording weinig benoemd of expliciet weerlegd. Dit blijkt uit de exploratieve analyses. Voor deze analyses zijn geen hypotheses opgesteld, maar zijn bedoeld om 'aanknopingspunten te bieden voor toekomstig onderzoek.

In de volgende subhoofdstukken wordt besproken hoe deze implicaties doorwerken in de theorie en de praktijk. Eerst wordt gekeken naar de beperkingen van dit onderzoek. Vervolgens wordt stilgestaan bij de implicaties voor de literatuur. Daarna worden de implicaties voor het werkveld besproken. Als laatste wordt een conclusie aan het onderzoek gebonden.

5.1 Beperkingen

Het uitgevoerde onderzoek heeft een aantal beperkingen. De beperkingen betreffen de steekproef, manipulatie check en aard van *appropriate reliance*. Elk van deze beperkingen wordt hieronder besproken.

De eerste beperkingen betreffen de steekproef. De steekproef heeft namelijk een beperkte omvang en is niet volledig gerandomiseerd. De grootte van de steekproef was 131 politieagenten. Voor de hoofdvraag is dit aantal verdubbeld tot 262, omdat elke agent twee scenario's heeft ingevuld. Deze hoeveelheid is

voldoende groot, dit blijkt uit de power van 0,9. Voor het vinden van het modererende effect is het aantal gereduceerd tot 161, door de uitsluiting van bepaalde leeftijdscategorieën, met als resultaat een power van 0,72. Deze power is lager dan de gewenste 0,8, waardoor de kans op een Type-II fout (aangezien geen effect is gevonden) groter wordt.

Daarnaast is de steekproef niet volledig gerandomiseerd, zowel in de selectie van participanten als door het falen van één van de randomisaties. Dit vergroot de kans dat de gevonden of niet-gevonden effecten te wijten zijn aan derde variabelen. Desondanks bestaat de steekproef uit 131 politieagenten wiens ervaring en kennis direct uit het veld komen. Een volledig willekeurige steekproef van contactambtenaren is buitengewoon moeilijk en zeldzaam in dit onderzoeksveld. Voor de beperkte scope van dit onderzoek is de praktijkervaring van de agenten en de hoeveelheid al van grote waarde. De randomisatie van de groepen niet is eveneens niet volledig geslaagd, want de variabele leeftijd in de groep Sassenheim is niet gelijkmatig verdeeld. Toch was leeftijd wel gelijkmatig verdeeld over de steekproef. Aangezien ervaring en leeftijd vaak samenhangen, wordt verwacht dat de niet volledig geslaagde randomisatie geen impact heeft op het onderzoek.

De tweede en een belangrijke beperking binnen het onderzoek is het falen van de manipulatiecheck. Dit houdt in dat de manipulatie van verantwoording afleggen binnen het experiment niet goed heeft gewerkt. De controlegroep ondervond zelfs een klein hoeveelheid meer verantwoording dan de verantwoordingsgroep. Daardoor zijn vraagtekens te zetten bij de uiteindelijke uitvoering van de manipulatie. Mogelijke verklaringen hiervoor zijn de omgekeerde werking van berichtgeving of de algemene ervaren verantwoording door agenten. Wellicht gaf de berichtgeving: ***“Je teamleider vraagt in dergelijke situaties nooit om je keuzes achteraf uit te leggen.”*** een omgekeerde werking, waarbij de agent dacht, normaal moet ik dit wel. Terwijl de berichtgeving: ***“Bij gebruik van de HOV vindt jouw teamleider het belangrijk dat jij je keuze achteraf verantwoord door uit te leggen waarom jij wel of niet hebt geluisterd naar het algoritme.”*** leidde tot de gedachte dat de agent dit normaal niet zo streng hoeft na te leven. Bovendien blijkt uit recent onderzoek in Bejing dat agenten die met AI-adviezen werken, zich

minder verantwoordelijk voelen voor de besluiten in het algemeen (Deng & Sun, 2024). Wellicht heeft dit effect het algehele effect van de manipulatie gedempt. Beide mogelijkheden zijn speculatie, maar kunnen mogelijke verklaringen geven.

Het gevolg van de gefaalde manipulatiecheck is dat de resultaten van de toetsen op de hoofdhypotheses onzeker zijn. De hypotheses gingen ervan uit de controlegroep minder verantwoording voelde dan de verantwoordingsgroep, dit verschil is weggevallen. Het resultaat hiervan is dat goed naar de beschrijvende statistiek gekeken moet worden en dat interpretaties van resultaten voorzichtig moeten blijven. Voor toekomstig onderzoek raad ik een sterkere manipulatie aan dan een dikgedrukt tekstblok. Denk hierbij aan een korte video of wellicht zelfs een labexperiment waarbij de participanten worden overzien door hun leidinggevende.

De laatste beperking binnen dit onderzoek is de definitie van *appropriate reliance*. *Appropriate reliance* is een lastig te definiëren en operationaliseren begrip. In dit onderzoek is gekozen voor een binaire definitie, met de toevoeging van ‘waarschijnlijk juist’ of ‘waarschijnlijk onjuist’. Het voordeel hiervan is, is het erkennen dat niet in alle situaties een 100% juist antwoord bestaat, terwijl het in de binairiteit te operationaliseren is. Het nadeel ervan is dat *appropriate reliance* weergeven als een spectrum met gradaties van ‘juistheid’ realistischer is (Bruijnes et al., 2024). Tegelijkertijd is het moeilijker te operationaliseren in een experiment. Aan de andere kant is ervan uitgaan dat er een juist en een onjuist antwoord is, maakt de operationalisatie makkelijker (Benda et al., 2022; Schemmer et al., 2023). Echter, het is minder realistisch en niet toepasbaar op een aanzienlijk deel van besluitvormingsprocessen in de publieke sector (Bruijnes et al., 2024). Daarom heeft dit experiment voor de middenweg gekozen, waarin een antwoord, ‘waarschijnlijk juist’ of ‘waarschijnlijk onjuist’ is. Om kritiek te ondervangen zijn de aannames achter de juistheid bij de opzet van het experiment gecontroleerd door agenten uit het veld. Al met al maakt het de keuze voor een definitie beperkend, maar wel de meest toepasselijke.

5.2 Literatuur implicaties

De eerder getrokken conclusies hebben implicaties voor de literatuur rondom AI-besluitvorming en verantwoording binnen publiek management.

5.2.1 AI-besluitvorming

AI-besluitvorming is een complex onderwerp, met diverse invalshoeken en discussiepunten. Dit onderzoek heeft een bijdrage geleverd aan deze literatuur. Een centrale discussie binnen de algoritmische besluitvorming literatuur is de eventuele beperking of toename van discretionaire ruimte bij contactambtenaren door de implementatie van algoritmische systemen (Bovens & Zouridis, 2002; Buffat, 2015; J. B. Bullock, 2019; de Boer & Raaphorst, 2023; Eck et al., 2018; Evans & Hupe, 2019). Deze discussie geeft weer dat aan de ene kant AI-systemen beperkend kunnen zijn voor de keuzevrijheid van contactambtenaren (Bovens & Zouridis, 2002; Buffat, 2015). Aan de andere kant, kunnen diezelfde contactambtenaren AI-systemen gebruiken als 'tools' om hun keuzevrijheid te vergroten (J. B. Bullock, 2019; Young et al., 2019). Dit onderzoek draagt bij aan die discussie door te laten zien dat contactambtenaren algoritmische adviezen kunnen opvolgen of naast zich neerleggen.

Ondanks dat de AI-systemen een waarschijnlijk juist advies gaven, besloten veel agenten de adviezen naast zich neer te leggen en hun eigen ingevingen te volgen. Dit bevestigt onderzoeken zoals die van Peeters (2020), Alon-Barkat et al (2023) en Selten et al. (2023) die weergeven dat contactambtenaren bias vertonen richting hun eigen ingevingen en niet zomaar algoritmische adviezen opvolgen. Dit onderzoek voegt hieraan toe dat dit ook gebeurt in contexten waarin algoritmes een waarschijnlijk juist antwoord geven.

Algoritmes zijn niet compleet sturend, maar worden ook niet genegeerd. Veel van de verklaringen die worden gegeven in de verantwoordinggroep als de agenten het algoritmisch advies opvolgen, refereren ernaar dat de HOV heeft bijgedragen aan hun keuze. Tegelijkertijd worden hier veelal redenen aan toegevoegd zoals *"het volgen van mijn intuïtie"*, *"locatie bekendheid"* of andere bijkomende redenen. De focus van de verantwoording lag op het bewijzen van eigen gelijk, niet op het

bijvoorbeeld weerleggen van de AI-suggestie. Het AI-advies lijkt te dienen als iets dat de Boer & Raaphorst (2023) omschrijven als een algoritmische nudge of zoals Meijer et al. (2021) het omschrijft, een collega die advies geeft. Een duwtje in een bepaalde richting, maar niet het discretie een beperkende systeem dat wordt omschreven door Bovens & Zouridis (2002). Tegelijkertijd ligt dit ook aan de vorm van het experiment, de suggestie van het AI-systeem is ook duidelijk gepresenteerd als advies.

Verder onderzoek naar de toepassing van AI-adviezen en de effecten op besluitvorming is nodig. Zoals Peeters (2020) aangeeft is een samenwerking wenselijk waarbij een mens de mogelijkheid heeft om AI-besluiten te controleren en te verbeteren. Door technologische beperkingen en menselijke controle problemen is deze samenwerking moeilijk te realiseren (Bruijnes et al., 2024; Sun & Medaglia, 2019; Tsamados et al., 2022; Young et al., 2019). Hopelijk brengen toekomstige onderzoeken de beste vorm van samenwerking tussen systeem en mens beter in kaart, zodat de toekomst hierop voort kan bouwen.

Voor dergelijke onderzoeken heb ik twee suggesties. (1) Een exploratief of literatuur onderzoek achter de beweegredenen van contactambtenaren om AI-adviezen wel of niet op te volgen. Veel verschillende redenen en contextfactoren lijken van belang te zijn, maar het gehele beeld is nog niet volledig in kaar gebracht. Denk hierbij aan de aard van het systeem, aard van de contactambtenaar, organisatiecontext, mogelijk morele kreukelzones en meer (Alon-Barkat & Busuioc, 2023; J. B. Bullock, 2019; Elish, 2019; Giest & Grimmelikhuijsen, 2020; Peeters, 2020; Selten et al., 2023). Zowel een literatuur als exploratief kwalitatief onderzoek die de achterliggende beweegredenen van contactambtenaren in kaart kan brengen kan een waardevolle toevoeging zijn aan de literatuur. Soares et al. (2023) heeft recent een etnografisch onderzoek uitgebracht die een aanzet geeft voor een preciezere in kaart brenging van de contextfactoren die wel of niet bijdragen aan het opvolgen van AI-advies.

(2) Een onderzoek dat nogmaals de rol van verantwoording afleggen onderzoekt, eventueel in een andere context. Het afleggen van verantwoording kan afschrikwekkende effecten hebben tegenover AI-systemen, zoals de morele

kreukelzonetheorie weergeeft (Elish, 2019), maar kan ook bijdragen aan het overkomen van bias (Aleksovska et al., 2019; Tetlock, 1983). Tegelijkertijd blijkt het dat agenten die AI-advies opvolgen zich minder verantwoordelijk voelen (Deng & Sun, 2024). De niet geslaagde manipulatiecheck zorgt ervoor dat dit onderzoek geen uitsluitsel kan geven over de rol van verantwoording bij het opvolgen van AI-adviezen. Daarom hoop ik nog een onderzoek te zien waarin een groep die zich echt verantwoordelijk voelt gehouden, tegenover een groep wordt gezet die geen verantwoording voelt. Hierdoor kan de rol van verantwoording in kaart worden gebracht bij besluitvorming met AI-advies.

5.2.2 Appropriate reliance en verantwoording

Appropriate reliance is een belangrijk concept om de 'juiste' hoeveelheid vertrouwen in AI-adviezen te bereiken en contactambtenaren in staat te stellen te handelen op deze 'juiste' hoeveelheid vertrouwen (Benda et al., 2022; Lee & See, 2004; Schemmer et al., 2022, 2023). De hoeveelheid onderzoek naar het concept is nog summier. Dit onderzoek draagt bij aan het in kaart brengen van contextfactoren die *appropriate reliance* op AI-advies vergroten. Specifiek is dit onderzoek gedaan naar de rol van het afleggen van verantwoording. De resultaten toonde aan dat verantwoording afleggen niet bijdraagt bij het voorkomen van *underreliance*. Tegelijkertijd is hier geen uitsluitsel over te geven, dankzij het falen van de manipulatiecheck.

Alsnog is het resultaat kenmerkend voor het feit dat *appropriate reliance*, vooral bij contactambtenaren, niet makkelijk te bereiken is. De literatuur suggereert dat automatiseringsbias, een hiërarchische organisatiecultuur of begrensde rationaliteit bijdragen aan *overreliance* (Meijer et al., 2021; Peeters, 2020). Terwijl ervaring of een grote professionele toewijding bij kan dragen aan *underreliance* (Bruijnes et al., 2024; Selten et al., 2023). De beschrijvende statistiek in dit onderzoek geeft iets anders weer. Daarin is te zien dat agenten met minder ervaring in verhouding meer *underreliant* zijn dan agenten met meer ervaring. Een verklaring hiervoor is dat het scholing en socialisatieproces recenter heeft plaatsgevonden bij minder ervaren agenten en dat dit bijdraagt aan de vertrouwen op deze kennis (Noordegraaf, 2015). Echter, het opbouwen van ervaring zal hier dan eveneens aan

bij horen te dragen, ondanks dat dit niet bleek uit de resultaten (Aleksovska et al., 2019; Alon-Barkat & Busuioc, 2023; Noordegraaf, 2015).

Het afleggen van verantwoording heeft uiteindelijk niet bijgedragen aan het *appropriate reliant* zijn van politieagenten. Dit geeft nieuwe nuances in zowel verantwoording als de *appropriate reliance* literatuur. Volgens de klassieke verantwoording literatuur hoort de kennis dat verantwoording afgelegd moet worden te helpen bij het voorkomen van eerste indruk bias en bias in het algemeen (Aleksovska et al., 2019; Ličen & Slapničar, 2022; Skitka et al., 2000; Tetlock, 1983). Het experiment was zodanig ontworpen dat een eerste indruk de klassieke professionele bias aanspreekt. Indien een respondent zorgvuldiger leest en zijn keuze heroverweegt, blijkt dat het AI-systeem wellicht zelfs een betere keuze maakt. Dit effect had kunnen optreden terwijl de respondent zijn verantwoording moest typen. Echter, dit effect leek niet voor te komen in het experiment.

Een mogelijke verklaring hiervoor is de morele kreukelzonetheorie van Elish (2019), ondersteund door de recente bevindingen van Deng & Sun (2024). De morele kreukelzonetheorie stelt dat ambtenaren afstand nemen van AI-adviezen omdat zij verwachten verantwoordelijk te worden gehouden voor eventuele fouten. Dit komt vooral doordat zij niet verantwoordelijk willen worden gehouden voor beslissingen waarvan de redenering onduidelijk is, dit is vaak een probleem is bij AI-systemen (Burrell, 2016; Giest & Grimmelikhuijsen, 2020; Tsamados et al., 2022). Deng & Sun (2024) tonen een soortgelijk effect aan, waarbij agenten zich minder verantwoordelijk voelen wanneer zij een AI-advies hebben opgevolgd. Met andere woorden, de agenten lijken afstand te doen van de eigenaarschap van de beslissing. Tegelijkertijd als dit het geval is, zou verantwoording moeten leiden tot *underreliance*. Dit is echter niet gebleken uit de resultaten. In de verantwoording die de agenten gaven, noemden een groot deel van de agenten de HOV niet. Zij ontweken de adviezen niet expliciet; sterker nog, de groep agenten die de adviezen van de HOV niet opvolgde, gebruikte alsnog de uitleg achter de waarschijnlijk onjuiste locatie in hun verantwoording. De morele kreukelzonetheorie biedt daarom geen sterke verklaring voor de resultaten van dit onderzoek (Elish, 2019).

Het voorgaande samenvattend maakt duidelijk dat verder onderzoek nodig is naar *appropriate reliance*. Hiervoor doe ik drie suggesties. (1) Onderzoek naar *appropriate reliance* binnen andere contexten in de publieke sector. Elk beroep van contactambtenaar is uniek en kent eigen denkwijzen (Noordegraaf, 2015). Dit onderzoek richtte zich specifiek op de politie, maar het zou zeer waardevol zijn om te onderzoeken of *appropriate reliance* andere implicaties heeft in andere beroepen. (2) Een onderzoek naar *appropriate reliance* dat uitgaat van een strikt binaire definitie. Dit onderzoek ging uit van een 'waarschijnlijk juist' en een 'waarschijnlijk onjuist' antwoord binnen *appropriate reliance* (Schemmer et al., 2023). Er bestaan echter ook taken waarbij besluiten zijn te categoriseren als feitelijk juist of feitelijk onjuist. Denk hierbij aan een onderzoek waarbij een AI-systeem consequent juiste antwoorden geeft, totdat het plotseling een onjuist antwoord geeft. Zal een contactambtenaar dat verschil kunnen opmerken, of zal de persoon stuiten op de problemen met menselijke controle zoals beschreven door Peeters (2020)? Een dergelijk onderzoek kan ook beter de *overreliance* zijde beter belichten. (3) Ten slotte, een onderzoek dat de organisatorische context goed in kaart brengt. Eerder is aangetoond dat de organisatorische context bepalend is voor de mate waarin politieagenten AI-besluiten naleven(Meijer et al., 2021). In het onderzoek van Meijer et al. (2021) naar de Duitse en Nederlandse politiecontexten wordt het concept *appropriate reliance* niet gebruikt, maar een onderzoek die dat wel doet kan eveneens interessante verschillen opleveren tussen organisationele contexten.

5.3 Implicaties Werkveld

De maatschappelijke relevantie in de inleiding heeft weergegeven dat dit onderzoek bij kan dragen aan een verbetering van AI-besluitvorming door contactambtenaren. AI-systemen bieden namelijk veel mogelijke uitdagingen en kansen binnen de Rijksoverheid (Digitale Overheid, 2023a; Rijksoverheid, 2022). Het goed in kaart brengen op welke manier een 'juiste' hoeveelheid vertrouwen in AI-systemen teweeg gebracht kan worden draagt bij aan een verbetering van AI-besluitvorming. Dit onderzoek legt een klein, beginnend deel van deze puzzel neer. Deze puzzel heeft nog werk nodig. Dit onderzoek probeert één contextfactor in kaart

te brengen, maar er zullen ook andere contextfactoren zijn die in de praktijk helpen om *appropriate reliance* te bereiken. Met andere woorden, er is meer onderzoek nodig, maar de puzzel begint ergens op te lijken.

Een belangrijke bevinding bijvoorbeeld, is dat het afleggen van verantwoording in dit onderzoek niet heeft geleid tot een verbetering van *appropriate reliance*. Wie verantwoordelijk wordt gehouden voor besluiten die met behulp van AI-systemen worden genomen, is een belangrijk onderzoeksthema. Wieringa (2020) pleit er bijvoorbeeld voor dat ook ambtelijke besluitnemers verantwoordelijk worden gehouden. Echter, het creëren van een expliciete verantwoordingsverwachting bij de contactambtenaar, verbetert de samenwerking met het AI-systeem niet direct. Indien verantwoordingsstructuren voor contactambtenaren worden ingericht op het werken met AI-besluitvormingssystemen, is het van belang dat niet wordt verwacht dat hierdoor de kwaliteit van AI-besluitvorming verbetert. Daarvoor is nog meer onderzoek nodig.

Daarbij is de bevinding van waarde dat werknemers die minder ervaren zijn binnen een vakgebied ook *underreliance* kunnen vertonen. Jonge of onervaren medewerkers kunnen eveneens bias vertonen of AI-adviezen naast zich neerleggen. Alleen omdat jongere medewerkers zijn ‘opgegroeid met technologie’ of nog niet volledig gesocialiseerd zijn moet de verwachting niet gewekt worden dat zij beter overweg kunnen met technologieën (Noordegraaf, 2015). Net zoals het stereotype rondom oudere medewerkers kunnen zij sceptisch zijn rondom AI-adviezen.

Voor verdere praktische aanbevelingen rondom AI-besluitvorming verwijs ik door naar onderzoeken van onder andere: (Bruijnes et al., 2024; J. Bullock et al., 2020; de Boer & Raaphorst, 2023; Deng & Sun, 2024; Grimmelikhuijsen & Meijer, 2022; Soares et al., 2023; Wieringa, 2020; Young et al., 2019). Deze onderzoekers brengen in kaart welke praktische implicaties de toevoeging van AI-systemen hebben binnen het werkveld.

Naast de aanbevelingen voor het werkveld is er een belangrijke vraag die onderzocht moet worden. Namelijk de bruikbaarheid van *appropriate reliance* in de publieke sector. Het concept staat of valt bij ‘juistheid’ en ‘onjuistheid’ of verschillende gradaties daarvan. Tegelijkertijd is binnen de publieke sector datgene

dat juist is, lastig te definiëren. Kijk bijvoorbeeld naar het concept van publieke waarde. Moore (2013) heeft zijn leven gespendeerd om te definiëren wat publiek waardevol is, maar de precieze definiëring van het concept blijft lastig samen te vatten. Eveneens met *appropriate reliance*, wellicht kijken wij als samenleving over tien jaar compleet anders naar iets dat nu gezien wordt als 'juist'. Indien de definiëring van 'juistheid' te complex wordt gemaakt, dan is de term niet meer praktisch bruikbaar. Het uitdiepen van deze vraag en potentiële bezwaren tegen het concept zullen waardevol zijn voor de eventuele praktische toepassingen.

5.4 Afsluiting

Hopelijk blijven de ontwikkelingen zich in het werkveld en de wetenschap zich voortzetten zodat AI-besluitvorming beter benut kan worden. AI-systemen kunnen de publieke sector veel blijven bieden, maar het blijft belangrijk om zorgvuldig onderzoek te doen naar de implicaties in het werkveld. Hierbij is het van belang om te luisteren naar de ervaringen van de professionals die zich bevinden in de frontlinie van deze grotere veranderingen. Verandering teweeg brengen binnen deze frontlinies is veelal een van de grootste uitdagingen van digitalisering, maar het zorgvuldig en goed oppakken van deze veranderingen mede door te luisteren naar wetenschap en professionals, is veelal van grote publieke waarde. In alle gevallen van besluitvorming, op welke wijze dan ook binnen de publieke sector, heeft dat namelijk rechtstreekse gevolgen voor de maatschappij en haar burgers.

Bronnen

Aleksovska, M., Schillemans, T., & Grimmelikhuijsen, S. (2019). Lessons from five decades of experimental and behavioral research on accountability: A systematic literature review. *Journal of Behavioral Public Administration*, *2*(2). https://doi.org/10.30636/jbpa.22.66

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, *33*(1). https://doi.org/10.1093/jopart/muac007

Anderson-Cook, C. M. (2005). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. *Journal of the American Statistical Association*, *100*(470). https://doi.org/10.1198/jasa.2005.s22

Baekgaard, M., Baethge, C., Blom-Hansen, J., Dunlop, C. A., Esteve, M., Jakobsen, M., Kisida, B., Marvel, J., Moseley, A., Serritzlew, S., Stewart, P., Thomsen, M. K., & Wolf, P. J. (2015). Conducting Experiments in Public Management Research: A Practical Guide. *International Public Management Journal*, *18*(2). https://doi.org/10.1080/10967494.2015.1024905

Bannister, F., & Connolly, R. (2020). Administration by algorithm: A risk management framework. *Information Polity*, *25*(4). https://doi.org/10.3233/IP-200249

Benda, N. C., Novak, L. L., Reale, C., & Ancker, J. S. (2022). Trust in AI: why we should be designing for APPROPRIATE reliance. *Journal of the American Medical Informatics Association*, *29*(1). https://doi.org/10.1093/jamia/ocab238

Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation and Governance*, *16*(1). https://doi.org/10.1111/rego.12358

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review*, *62*(2). https://doi.org/10.1111/0033-3352.00168

Brandler, S., & Roman, C. P. (2020). Identifying Dimensions and Types in Public Administration Research: Introduction to Principal Components Analysis, Factor Analysis, and Cluster Analysis. In *Handbook of Research Methods in Public Administration*. https://doi.org/10.1201/9781420013276-41

Bruijnes, M., Robeer, M., & Grimmelikhuijsen, S. (2024). Explainable AI is No Silver Bullet: Towards a Contextual Understanding of Appropriate Reliance on AI in Law Enforcement. In J. Goossens, E. Keymolen, & A. Stanojević (Eds.), *Public Governance and Emerging Technologies: Values, Trust, and Compliance by Design*. Springer.

Bryman, A. (2015). Social Research Methods (4th Edition) by Alan Bryman. *Abhigyan VO - 32*, *4*.

Buffat, A. (2015). Street-Level Bureaucracy and E-Government. *Public Management Review*, *17*(1). https://doi.org/10.1080/14719037.2013.771699

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *American Review of Public Administration*, *49*(7). https://doi.org/10.1177/0275074019856123

Bullock, J., Young, M. M., & Wang, Y. F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity*, *25*(4). https://doi.org/10.3233/IP-200223

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Chavannes, M. (2020, November 27). De overheid werd een bedrijf dat mensen onverdiend wantrouwt. *De Correspondent*. https://decorrespondent.nl/11858/de-overheid-werd-een-bedrijf-dat-mensen-onverdiend-wantrouwt-alleen-kamer-en-kabinet-kunnen-die-denkfout-herstellen/bdca3c9c-4140-01a5-14a1-eabed68709e0

Chi, O. H., Jia, S., Li, Y., & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, *118*. https://doi.org/10.1016/j.chb.2021.106700

Cohen, J. (1992). Quantitative Methods in Psychology: A Power Primer. *Psychological Bulletin*, *112*(1).

de Boer, N., & Raaphorst, N. (2023). Automation and discretion: explaining the effect of automation on how street-level bureaucrats enforce. *Public Management Review*, *25*(1). https://doi.org/10.1080/14719037.2021.1937684

Den Hartog, T. (2016, February 18). *Agent op straat te oud om werk goed te doen*. AD. https://www.ad.nl/binnenland/agent-op-straat-te-oud-om-werk-goed-te-doen~a39f3206/?referrer=https%3A%2F%2Fwww.google.com%2F#:~:text=Uit%20het%20jaarverslag%20van%20de,gestegen%20naar%2041%2C7%20jaar.

Deng, Y., & Sun, Y. (2024). How does the usage of artificial intelligence affect felt administrative accountability of street-level bureaucrats? The mediating effect of perceived discretion. *Public Management Review*. https://doi.org/10.1080/14719037.2024.2370982

Digitale Overheid. (2023a). *Naar verantwoorde inzet van algoritmes*. Rijksoverheid. https://www.digitaleoverheid.nl/overzicht-van-alle-

onderwerpen/algoritmes/#:~:text=Een%20algoritme%20is%20een%20set,(bron%3A%20Algemene%20Rekenkamer).

Digitale Overheid. (2023b, October 13). *AI maakt politiewerk makkelijker én moeilijker*. Digitale Overheid. https://www.digitaleoverheid.nl/achtergrondartikelen/artifical-intelligence-maakt-politiewerk-makkelijker-en-moeilijker/#:~:text=AI%20kan%20helpen%20bij%20het,verzamelen%20en%20verdachten%20eerder%20opsporen.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (2011). Experiments: An Introduction to Core Concepts. In *Cambridge Handbook of Experimental Political Science*.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human Computer Studies*, *58*(6). https://doi.org/10.1016/S1071-5819(03)00038-7

Eck, B., Bovens, M., & Zouridis, S. (2018). Algoritmische rechtstoepassing in de democratische rechtsstaat. *Nederlands Juristenblad*, *93*(40).

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, *5*. https://doi.org/10.17351/ests2019.260

Evans, T. (2011). Professionals, managers and discretion: Critiquing street-level bureaucracy. *British Journal of Social Work*, *41*(2). https://doi.org/10.1093/bjsw/bcq074

Evans, T., & Hupe, P. (2019). Discretion and the Quest for Controlled Freedom. In *Discretion and the Quest for Controlled Freedom* (1st ed.). Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-030-19566-3

Field, A. P. (2018). Discovering statistics using IBM SPSS statistics: 5th edition. In *SAGE Publications, Inc.* (Vol. 4, Issue 1).

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, *3*(AUG). https://doi.org/10.3389/fpsyg.2012.00325

Giest, S., & Grimmelikhuijsen, S. (2020). Introduction to special issue algorithmic transparency in government: Towards a multi-level perspective. In *Information Polity* (Vol. 25, Issue 4). https://doi.org/10.3233/IP-200010

Grimmelikhuijsen, S. (2023). Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making. *Public Administration Review*, *83*(2). https://doi.org/10.1111/puar.13483

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, *5*(3). https://doi.org/10.1093/ppmgov/gvac008

Hannah-Moffat, K. (2013). Actuarial Sentencing: An "Unsettled" Proposition. *Justice Quarterly*, *30*(2). https://doi.org/10.1080/07418825.2012.682603

Hauser, D. J., & Schwarz, N. (2015). It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015584617

Hochwarter, W. A., Ferris, G. R., Gavin, M. B., Perrewé, P. L., Hall, A. T., & Frink, D. D. (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). https://doi.org/10.1016/j.obhdp.2006.09.003

Keddell, E. (2019). Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice. *Social Sciences*, *8*(10). https://doi.org/10.3390/socsci8100281

Kennis van de Overheid. (2022). *Omvang en samenstelling personeelsbestand*. Ministerie van Binnenlandse Zaken En Koninkrijksrelaties. https://www.kennisvandeoverheid.nl/cijfers-overheidspersoneel/omvang-en-samenstelling-personeelsbestand

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. In *Human Factors* (Vol. 46, Issue 1). https://doi.org/10.1518/hfes.46.1.50_30392

Ličen, M., & Slapničar, S. (2022). Can process accountability mitigate myopic biases? An experimental analysis. *Journal of Management Control*, *33*(1). https://doi.org/10.1007/s00187-021-00330-7

Lipsky, M. (2010). Street-level bureaucracy: Dilemmas of the individual in public services. In *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. https://doi.org/10.2307/2392554

Maynard-Moody, S., & Musheno, M. (2000). State Agent or Citizen Agent: Two Narratives of Discretion. *Journal of Public Administration Research and Theory*, *10*(2). https://doi.org/10.1093/oxfordjournals.jpart.a024272

McDermott, R. (2002). Experimental methods in political science. In *Annual Review of Political Science* (Vol. 5). https://doi.org/10.1146/annurev.polisci.5.091001.170657

Meijer, A., Lorenz, L., & Wessels, M. (2021). Algorithmization of Bureaucratic Organizations: Using a Practice Lens to Study How Context Shapes Predictive

Policing Systems. *Public Administration Review*, *81*(5). https://doi.org/10.1111/puar.13391

Miller, T. (2019a). "But why?" Understanding explainable artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, *25*(3). https://doi.org/10.1145/3313107

Miller, T. (2019b). Explanation in artificial intelligence: Insights from the social sciences. In *Artificial Intelligence* (Vol. 267). https://doi.org/10.1016/j.artint.2018.07.007

Moore, M. H. (2013). Recognizing Public Value. In *Recognizing Public Value*. https://doi.org/10.4159/harvard.9780674067820

Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, *113*(2). https://doi.org/10.1017/S0003055418000837

Mutz, D. C., & Pemantle, R. (2015). Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods. *Journal of Experimental Political Science*, *2*(2). https://doi.org/10.1017/XPS.2015.4

Nederlandse Politie Bond. (2023, September 9). *Professionele Ruimte*. NPB.

Noordegraaf, M. (2015). *Public Management. Performance, Professionalism, Politics*. Palgrave McMillan.

Overman, S., & Schillemans, T. (2022). Toward a Public Administration Theory of Felt Accountability. *Public Administration Review*, *82*(1). https://doi.org/10.1111/puar.13417

Overman, S., Schillemans, T., & Grimmelikhuijsen, S. (2021). A validated measurement for felt relational accountability in the public sector: gauging the account holder's legitimacy and expertise. *Public Management Review*, *23*(12). https://doi.org/10.1080/14719037.2020.1751254

Peeters, R. (2020). The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making. *Information Polity*, *25*(4). https://doi.org/10.3233/IP-200253

Peeters, R., & Widlak, A. C. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review*, *83*(4). https://doi.org/10.1111/puar.13615

Peters, B. G., & Guedes-Neto, J. V. (2020). Experimental methods A: Survey experiments in public administration. In *Handbook of Research Methods in Public Administration, Management and Policy*. https://doi.org/10.4337/9781789903485.00020

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., & Goel, S. (2020). A

large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, *4*(7). https://doi.org/10.1038/s41562-020-0858-1

Politie. (n.d.). *Mandaten en Regelingen*. Z.d. Retrieved May 28, 2024, from https://www.politie.nl/onderwerpen/mandaten-en-regelingen.html

Politie. (2017). *Arbeidsmarkt en politie: Een verkenning van de wervingskansen*.

Rijksoverheid. (2022, January 10). *Coalitieakkoord "Omzien naar elkaar, vooruitkijken naar de toekomst": Digitalisering*.

Rijksoverheid. (2024, January 18). *Kabinet presenteert visie op generatieve AI*. Rijksoverheid.

Samuels, P. (2016). Advice on Exploratory Factor Analysis. *Centre for Academic Success, Birmingham City University*, *June*.

Scharowski, N., Perrig, S. A. C., Von Felten, N., & Brühlmann, F. (2022). Trust and Reliance in XAI - Distinguishing Between Aitudinal and Behavioral Measures. *CHI 2022: Workshop on Trust and Reliance in AI-Human Teams*.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *In CHI Conference on Human Factors in Computing Systems (CHI '22), Workshop on Trust and Reliance in AI-Human Teams (TrAIt)*.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. *International Conference on Intelligent User Interfaces, Proceedings IUI*. https://doi.org/10.1145/3581641.3584066

Schillemans, T., Overman, S., Fawcett, P., Flinders, M., Fredriksson, M., Laegreid, P., Maggetti, M., Papadopoulos, Y., Rubecksen, K., Rykkja, L. H., Salomonsen, H. H., Smullen, A., & Wood, M. (2021). Understanding Felt Accountability: The institutional antecedents of the felt accountability of agency-CEO's to central government. *Governance*, *34*(3). https://doi.org/10.1111/gove.12547

Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, *83*(2). https://doi.org/10.1111/puar.13602

Silber, H., Roßmann, J., & Gummer, T. (2022). The Issue of Noncompliance in Attention Check Questions: False Positives in Instructed Response Items. *Field Methods*, *34*(4). https://doi.org/10.1177/1525822X221115830

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human Computer Studies*, *52*(4). https://doi.org/10.1006/ijhc.1999.0349

Soares, C., Grimmelikhuijsen, S., & Meijer, A. (2023). Screen-level bureaucrats in the age of algorithms. An ethnographic study of algorithmically supported public service workers in the Netherlands Police. *Information Polity*. https://doi.org/10.3233/ip-220070

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*(2). https://doi.org/10.1016/j.giq.2018.09.008

Tetlock, P. E. (1983). Accountability and the Perseverance of First Impressions. *Social Psychology Quarterly*, *46*(4). https://doi.org/10.2307/3033716

Thomann, E., van Engen, N., & Tummers, L. (2018). The Necessity of Discretion: A Behavioral Evaluation of Bottom-Up Implementation Theory. *Journal of Public Administration Research and Theory*, *28*(4). https://doi.org/10.1093/jopart/muy024

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: key problems and solutions. *AI and Society*, *37*(1). https://doi.org/10.1007/s00146-021-01154-8

Tummers, L., & Bekkers, V. (2014). Policy Implementation, Street-level Bureaucracy, and the Importance of Discretion. *Public Management Review*, *16*(4). https://doi.org/10.1080/14719037.2013.841978

Van Eijk, G. C. N.-H. . A. 2020. (2020). Algorithmic reasoning: The production of subjectivity through data. *The Algorithmic Society: Technology, Power, and Knowledge*.

Van Koetsveld, R., Hartmans, R., & De Man, A. P. (2016). *Politie cultuur in beweging*.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372833

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration. *Perspectives on Public Management and Governance*, *2*(4). https://doi.org/10.1093/ppmgov/gvz014

Bijlage

Bijlage A. Factor analyses

Tabel A1.
Meetschalen Technologie vertrouwen via exploratieve factoranalyses

Construct	Factor lading
Technologie vertrouwen (KMO = 0,69; α = 0,778)	
Factor 1. Vertrouwen in technologie	
Technologie_Q1 Ik vertrouw nieuwe technologieën totdat ze mij bewijzen dat ik dat niet zou moeten doen.	0,790
Technologie_Q2 Over het algemeen geef ik een technologie het voordeel van de twijfel wanneer ik het voor het eerst gebruik.	0,642
Technologie_Q3 Meestal vertrouw ik in technologie totdat het mij een reden geeft om er niet in te vertrouwen.	0,774

Noot 1. α = Cronbach's alpha; Sig <0,001; Schalen via Maximum Likelihood extractie en Direct Oblimin rotatie

Tabel A2.
Meetschalen verantwoordingsgevoel via exploratieve factoranalyses

Construct	Factor lading
Verantwoordingsgevoel (KMO = 0,627; α = 0,76)	
Factor 1. Verantwoordingsgevoel	
Verantwoording_Q1 Ik word verantwoordelijk gehouden voor besluiten die ik neem op basis van algoritmische adviezen zoals het HOV.	0,571
Verantwoording_Q2 Ik moet vaak uitleggen waarom ik een bepaalde keuzes maak op werk na een algoritmisch advies.	0,648
Verantwoording_Q3 Mijn teamleider houdt mij verantwoordelijk voor mijn besluiten die ik neem op basis van algoritmische advies.	0,969

Noot. 1 α = Cronbach's alpha; Sig <0,001; Schalen via Maximum Likelihood extractie en Direct Oblimin rotatie

Bijlage B. Representativiteit toets

Om te toetsen of de populatiedata significant verschilt van de steekproefdata wordt voor de variabele geslacht een tweezijdige Chi-kwadraat-goodness-of-fit (GOF) toets uitgevoerd met een criterium van een significantie van $p < 0{,}025$. De GOF toets laat zien of een frequentie significant verschilt van een verwachte frequentie. Ofwel verschilt de daadwerkelijke hoeveelheid mannen en vrouwen in de steekproef met de verwachte hoeveelheid als gekeken wordt naar de populatie (Cohen, 1992). Uit de resultaten van de toets is te zien dat de steekproef qua geslacht niet representatief is ($\chi 2(1, n = 129) = 11{,}867$, $p = 0{,}001$). Bij de toets is uitgegaan van de volgende H0. *De steekproef en populatiedata verschillen niet significant van elkaar qua geslacht.* Hierbij is H1. *De steekproef en populatiedata verschillen wel significant van elkaar qua geslacht.* Aangezien het resultaat niet significant is, is H0 dus verworpen.

De leeftijdscategorieën van de populatie en steekproefdata staan weergegeven in Tabel B1. Met behulp van een betrouwbaarheidsinterval, die de categoriegemiddelden gebruikt, is berekend dat de leeftijdsverdeling representatief is (Field, 2018). De gemiddelde leeftijd van de populatie ligt op 41,7 (Den Hartog, 2016). Het betrouwbaarheidsinterval van een 95% CI [39,77, 43,81] laat zien dat het populatiegemiddelde binnen de lower- en upperbounds van het steekproefgemiddelde ligt. Qua leeftijd is de steekproef dus representatief.

Tabel B1.
Steekproef kenmerken en representativiteit

	Populatie	Steekproef
Geslacht		
Man	61,8%	74,8%
Vrouw	39,2%	23,6%
Anders	...%	1,6%
Leeftijdscategorieën		
Jonger dan 25 jaar	3%	5,3%
25 tot 35 jaar	22%	26,7%
35 tot 45 jaar	23%	31,3%
45 tot 55 jaar	27%	19,1%
Ouder dan 55 jaar	25%	17,6%

Werkervaring		
5 jaar of minder	...%	18,3%
6 tot en met 10 jaar	...%	11,5%
11 tot en met 15 jaar	...%	16,8%
16 jaar of langer	...%	53,4%
Executieve status		
Ja	...%	92,3%
Nee	...%	7,7%

(Den Hartog, 2016; Kennis van de Overheid, 2022; Politie, 2017)
Noot 1. Om anonimiteit te behouden is leeftijd en ervaring gemeten in groepen. De interval gemiddelden zijn gebaseerd op groepsgemiddelden.

Bijlage C. Crosstabs hoofdvraag

Tabel C1.
Crosstabs vergelijking controle- en verantwoordinggroep *appropriate reliance Bergen op Zoom*

		Scenario Bergen op Zoom		
		AR	**UR**	**Totaal**
Controlegroep	Telling	34	35	69
	Verwachte telling	35,8	33,2	
Verantwoordingsgroep	Telling	34	28	62
	Verwachte telling	35,2	29,8	
Totaal	Telling	68	63	N = 131

Tabel C2.
Crosstabs vergelijking controle- en verantwoordinggroep *appropriate reliance Sassenheim*

		Scenario Sassenheim		
		AR	**UR**	**Totaal**
Controlegroep	Telling	29	35	64
	Verwachte telling	28,3	35,7	
Verantwoordingsgroep	Telling	29	38	67
	Verwachte telling	29,7	37,3	
Totaal	Telling	58	73	N = 131

Tabel C3.
Crosstabs vergelijking controle- en verantwoordinggroep *appropriate reliance totaal*

		Beide scenario's		
		AR	**UR**	**Totaal**
Controlegroep	Telling	63	70	133
	Verwachte telling	64	69	
Verantwoordingsgroep	Telling	63	66	129
	Verwachte telling	62	67	
Totaal	Telling	126	136	N = 262

Bijlage D. Crosstabs moderator

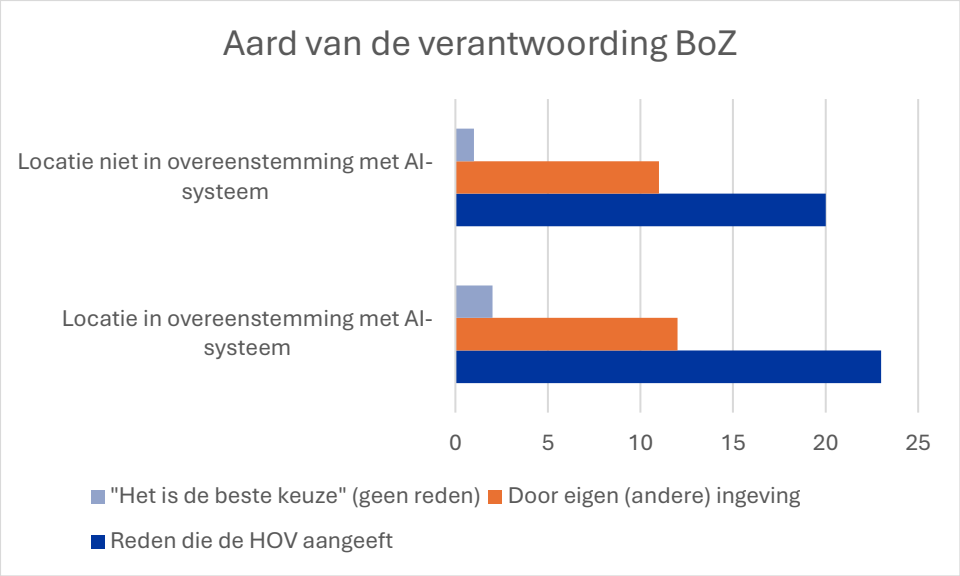
Tabel D1.
Crosstabs vergelijking controle- en verantwoordinggroep *appropriate reliance met ervaring als moderator*

Ervaringsgroepen	**Controle of verantwoordingsgroep**				
			AR	**UR**	**Totaal**
5 jaar en onder	Controlegroep	Telling	10	18	28
		Verwachte telling	9,9	18,1	
	Verantwoordingsgroep	Telling	7	13	20
		Verwachte telling	7,1	12,9	
Totaal		Telling	17	31	48
20 jaar en boven	Controlegroep	Telling	20	27	47
		Verwachte telling	23,9	23,1	
	Verantwoordingsgroep	Telling	38	29	67
		Verwachte telling	34,1	32,9	
Totaal		Telling	75	87	N = 162

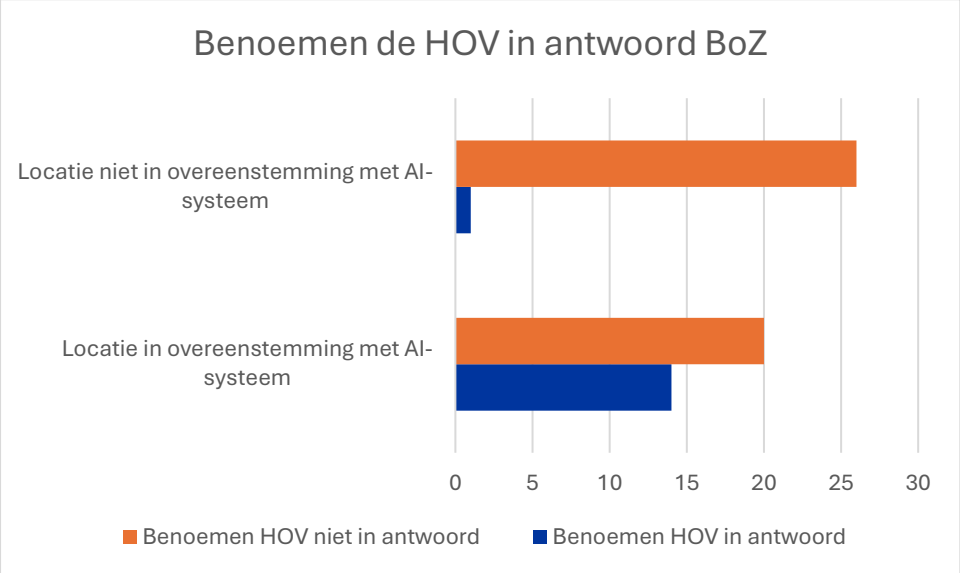
Tabel D2.
Crosstabs vergelijking werkervaringgroepen en *appropriate reliance*

	Beide scenario's			
		AR	**UR**	**Totaal**
Ervaring 5 jaar en onder	Telling	17	31	48
	Verwachte telling	22,2	25,8	
Ervaring boven 5 jaar	Telling	58	56	114
	Verwachte telling	52,8	61,2	
Totaal	Telling	75	87	N = 162

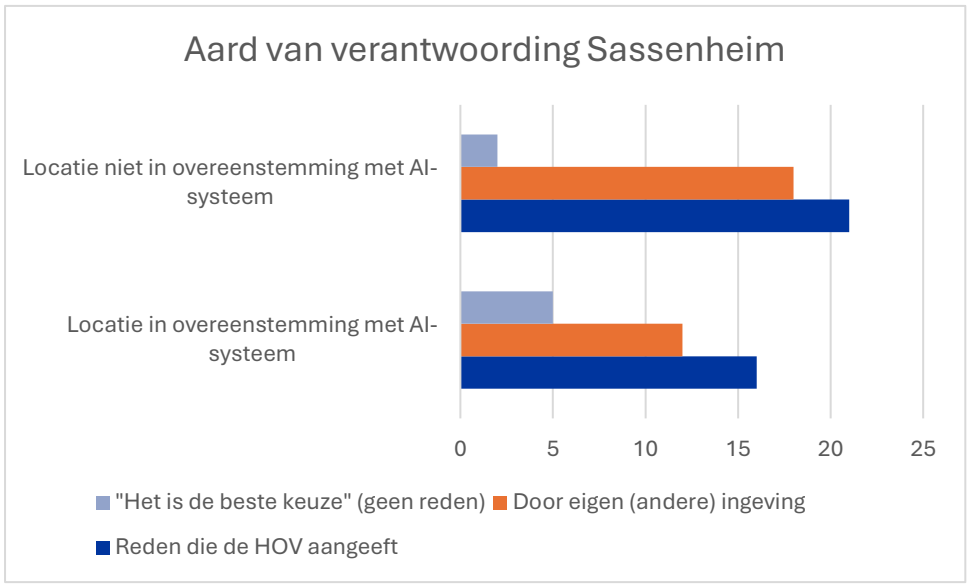
Bijlage E. Grafieken codering



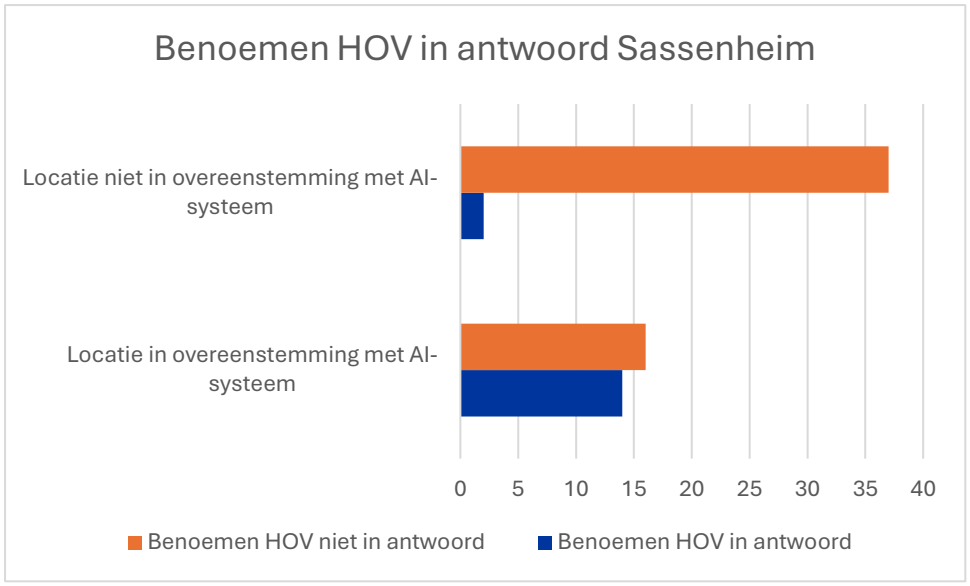
Figuur E1. Aard van verantwoording in scenario Bergen op Zoom.



Figuur E2. HOV benoemt in antwoord BoZ.



Figuur E3. Aard van verantwoording in scenario Sassenheim.



Figuur E4. HOV benoemt in antwoord Sassenheim.

Bijlage F. Enquête

Welkom bij dit onderzoek!

Om te beginnen, hartelijk dank dat je de tijd neemt om deze vragenlijst in te vullen. Wij zullen uw tijd respecteren en daarom duurt het formulier een kleine 7 minuten om in te vullen. Dit onderzoek wordt uitgevoerd door Mitch Opperhuizen in het kader van een masterscriptie bij de Universiteit Utrecht in samenwerking met het Nationaal Politielab AI.

Privacy verklaring

Het doel van dit onderzoek is om te leren hoe algoritmes het werk van politieagenten beïnvloeden. Het invullen van de vragenlijst draagt bij aan de ontwikkeling van deze algoritmes binnen de politie en het opdoen van wetenschappelijke kennis.

Het formulier presenteert een aantal scenario's waarin een algoritme wordt ingezet. Je wordt vervolgens gevraagd een aantal vragen te beantwoorden over deze scenario's.

Van belang is dat je het volgende weet:

- Je deelname is vrijwillig en je hebt het recht om je deelname zonder opgaaf van redenen te beëindigen.
- De verzamelde data zal niet direct terug te leiden zijn naar personen.
- Dit betekent dat het bewaren van persoonlijk identificeerbare informatie vervalt en niemand zal kunnen identificeren welke antwoorden / gegevens van jou zijn.
- Dit onderzoek maakt deel uit van een afstudeerproject. De resultaten worden gepubliceerd in een masterthesis en eventuele wetenschappelijke vervolgpublicaties. De gepubliceerde informatie zal op geen enkele wijze te herleiden zijn tot jou als persoon.

Voor vragen over dit onderzoek kun je contact opnemen met de uitvoerder van dit onderzoek Mitch Opperhuizen (m.e.opperhuizen@students.uu.nl). Daarbij vervullen dr. Stephan Grimmelikhuijsen vanuit de Universiteit Utrecht en Bas Testerink vanuit de politie een begeleidende rol. Indien er privacy zorgen zijn, dan kan contact opgenomen worden met de privacy officer van de Universiteit Utrecht (privacy@uu.nl).

1. *Ik ga akkoord aan deelname van het onderzoek*

◯ Ja, ik heb het bovenstaande gelezen en begrepen, en ik geef toestemming om mijn antwoorden te gebruiken voor wetenschappelijk onderzoek, zoals hierboven beschreven.

Introductie

Toelichting onderzoek

- Deze vragenlijst wordt uitgezet onder politiemedewerkers die werkzaam zijn in verschillende functies verspreid over het land.
- In deze vragenlijst worden twee verschillende bedachte situaties aan je voorgelegd. Na het lezen van een kort verhaal bij de bedacht situatie worden enkele vragen gesteld over wat jij zou doen.
- De situaties hebben niets met elkaar te maken. Het is daarom belangrijk dat je je antwoorden alleen baseert op de informatie die je in dat specifieke scenario kunt lezen.

Wat ga je doen

De eerste 10 minuten na de melding van een strafbaar feit zijn van cruciaal belang voor het

aanhouden van een verdachte. Een centralist bij de meldkamer kan in deze eerste tien minuten niet alle eenheden op een ideale locatie positioneren. **Daarom moet jij als agent eerst een eigen inschatting maken op welke plek je het best positie kunt innemen.**

Om je te ondersteunen bij het maken van deze inschatting heeft de politie een computersysteem

ontwikkeld: **de Hulp Onderschepping Verdachten (HOV)**. De HOV maakt een **inschatting van de**

vluchtroutes die verdachten kunnen gebruiken en adviseert waar jij het beste positie kan innemen

om hen te onderscheppen.

Je gaat zo meedoen aan een experiment waarin je wordt gevraagd om te werken met de HOV. Over de werking van de HOV moet je weten dat:

- Tests hebben uitgewezen dat in de meeste gevallen de HOV een betere inschatting maakt over de vluchtroute van verdachten dan een mens. We vragen je daarom het advies van de HOV serieus te nemen.
- Jouw eigen kennis, ervaring en intuïtie blijven echter belangrijk. De HOV geeft je een gefundeerd advies, maar je kunt ervoor kiezen dit advies niet te volgen.

Let op: De HOV bestaat niet echt, maar is wel gebaseerd op een systeem dat wordt ontwikkeld door de politie.

Achtergrond vragen

Er volgen nu zes korte vragen over jouw achtergrond en werkzaamheden bij de politie. Daarna zal je de scenario's te zien krijgen.

1. Wat is je geslacht?

◯ Vrouw
◯ Man
◯ Anders
◯ Wil ik niet zeggen

2. In welke leeftijdscategorie val je?

◯ Jonger dan 25
◯ 25 tot 35 jaar
◯ 35 tot 45 jaar
◯ 45 tot 55 jaar
◯ 55 jaar en ouder

3. In welke politie regio ben je werkzaam?

◯ Noord-Nederland
◯ Oost-Nederland
◯ Midden-Nederland
◯ Noord-Holland
◯ Amsterdam
◯ Den Haag
◯ Rotterdam
◯ Zeeland-West Brabant
◯ Oost-Brabant
◯ Limburg
◯ Landelijke eenheid
◯ Wil ik niet zeggen

4. Hoeveel jaar ben je al bij de politie werkzaam? (je mag afronden naar boven)

Getal 0-60 invullen

5. Heb je een executieve status?

◯ Ja

◯ Nee
◯ Nee, maar ik heb een executieve status gehad
◯ Nee, maar ik ben in opleiding voor een executieve status
◯ Wil ik niet zeggen

6. **Hoe kijk je over het algemeen tegen technologie aan?**

   a. **Ik vertrouw nieuwe technologieën totdat ze mij bewijzen dat ik dat niet zou moeten doen.**

Helemaal oneens Helemaal eens
1. 2. 3. 4. 5. 6. 7.

   b. **Over het algemeen geef ik een technologie het voordeel van de twijfel wanneer ik het voor het eerst gebruik.**

Helemaal oneens Helemaal eens
1. 2. 3. 4. 5. 6. 7.

   c. **Meestal vertrouw ik in technologie totdat het mij een reden geeft om er niet in te vertrouwen.**

Helemaal oneens Helemaal eens
1. 2. 3. 4. 5. 6. 7.

Scenario BoZ: Groep 1 (Controle)

Beeld je het volgende in: Jij bent een surveillance aan het uitvoeren in de Regio Zeeland – West Brabant. Rond 20:30 krijg je een melding via de portofoon dat er een liquidatie heeft plaatsgevonden in het theater. Via de portofoon hoor je dat er direct 112 is gebeld. De getuige heeft gezien dat er twee verdachten zijn, die zijn gevlucht in **zuidelijke richting** in een grijze sportwagen.

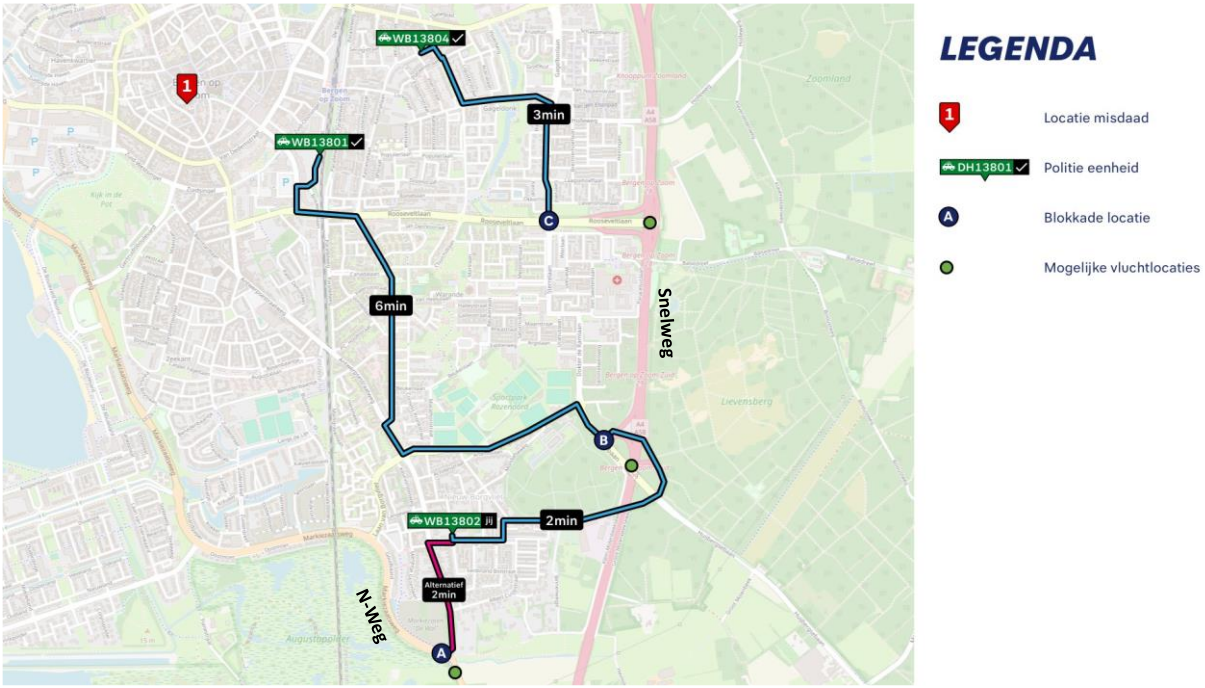
Jij bevindt je in de omgeving van de liquidatie en wordt gevraagd om te helpen met het dichtzetten van vluchtroutes. Je hebt van de meldkamer doorgekregen dat er een paar andere collega's ook onderweg zijn. **Jij bent de auto WB13802 (de onderste).**

Om de kans te vergroten dat de verdachten worden aangehouden wordt de Hulp Onderschepping Verdachten (HOV) ingeschakeld. **Je team leider vraagt in dergelijke situaties nooit om je keuzes achteraf uit te leggen.**

De HOV geeft het volgende in de app weer:

- **Locatie A, locatie B** zijn mogelijke vluchtroutes die jij kan dekken**. Locatie B en C zullen gedekt worden door collega auto's.**
- De HOV heeft **Locatie A** overwogen zodat alle vluchtroutes zijn afgezet.
- De HOV heeft **Locatie B** overwogen door de nabijheid van de snelweg.
- **De HOV beveelt op basis van deze informatie locatie B aan.**

Onderstaand wordt een kaart getoond die de verschillende locaties van dit scenario weergeven. Wij vragen je deze kaart goed te bekijken.



Scenario BoZ: Groep 2 (Verantwoording)

Beeld je het volgende in: Jij bent een surveillance aan het uitvoeren in de Regio Zeeland – West Brabant. Rond 20:30 krijg je een melding via de portofoon dat er een liquidatie heeft plaatsgevonden in het theater. Via de portofoon hoor je dat er direct 112 is gebeld. De getuige heeft gezien dat er twee verdachten zijn, die zijn gevlucht in **zuidelijke richting** in een grijze sportwagen.

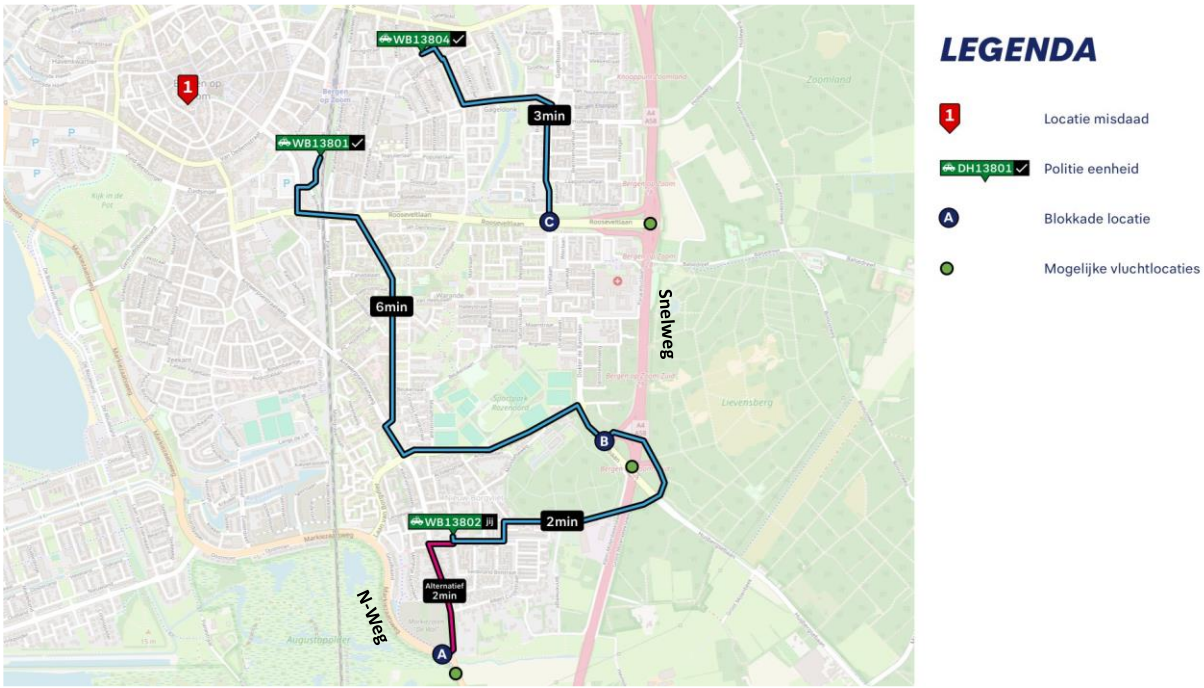
Jij bevindt je in de omgeving van de liquidatie en wordt gevraagd om te helpen met het dichtzetten van vluchtroutes. Je hebt van de meldkamer doorgekregen dat er een paar andere collega's ook onderweg zijn. **Jij bent de auto WB13802 (de onderste).**

Om de kans te vergroten dat de verdachten worden aangehouden wordt de Hulp Onderschepping Verdachten (HOV) ingeschakeld. **Bij gebruik van de HOV vindt jouw teamleider het belangrijk dat jij je keuze achteraf verantwoord door uit te leggen waarom jij wel of niet hebt geluisterd naar het algoritme.**

De HOV geeft het volgende in de app weer:

- **Locatie A, locatie B** zijn mogelijke vluchtroutes die jij kan dekken**. Locatie B en C zullen gedekt worden door collega's.**
- De HOV heeft **Locatie A** overwogen zodat alle vluchtroutes zijn afgezet.
- De HOV heeft **Locatie B** overwogen door de nabijheid van de snelweg.
- **De HOV beveelt op basis van deze informatie locatie B aan.**

Onderstaand wordt een kaart getoond die de verschillende locaties van dit scenario weergeven. Wij vragen je deze kaart goed te bekijken.



1. **Op basis van het advies van de HOV en jouw eigen inschatting van de situatie, welke locatie kies je dan om jezelf op te stellen om vluchtroutes van de verdachten af te zetten?**

◯ Locatie A

◯ Locatie B

2. **Na afloop van het incident vraagt je teamleider waarom je hebt gekozen voor jouw gekozen locatie. Geef hieronder kort je uitleg aan die persoon. [*Deze vraag komt alleen voor bij groep 2]***

[Vul hier je uitleg in]

3. **Je hebt de keuze gemaakt om naar [V1] te gaan. Hoe zeker ben je er van dat de verdachten**

daadwerkelijk langs deze route gevlucht zijn?

Slider van 0 naar 100

4. **Geef aan in welke mate je het eens bent met de volgende stellingen**

   a. **Ik word verantwoordelijk gehouden voor besluiten die ik neem op basis van algoritmische adviezen zoals het HOV**

Helemaal oneens Helemaal eens

1. 2. 3. 4. 5. 6. 7.

   b. **Ik moet vaak uitleggen waarom ik een bepaalde keuzes maak op werk na een algoritmisch advies**

Helemaal oneens Helemaal eens

1. 2. 3. 4. 5. 6. 7.

   c. **Mijn teamleider houdt mij verantwoordelijk voor mijn besluiten die ik neem op basis van algoritmische advies.**

Helemaal oneens Helemaal eens

1. 2. 3. 4. 5. 6. 7.

Scenario Sassenheim: Groep 1 (Controle)

Beeld je het volgende in: jij bent aan het surveilleren in Sassenheim. Om 20:57 krijg je een melding via de portofoon dat er een inbraak heterdaad is geweest. De bewoner van een woonhuis is thuisgekomen en heeft twee inbrekers betrapt. De verdachten zijn in een **Noord Oostelijke** richting gevlucht in donkerblauwe auto.

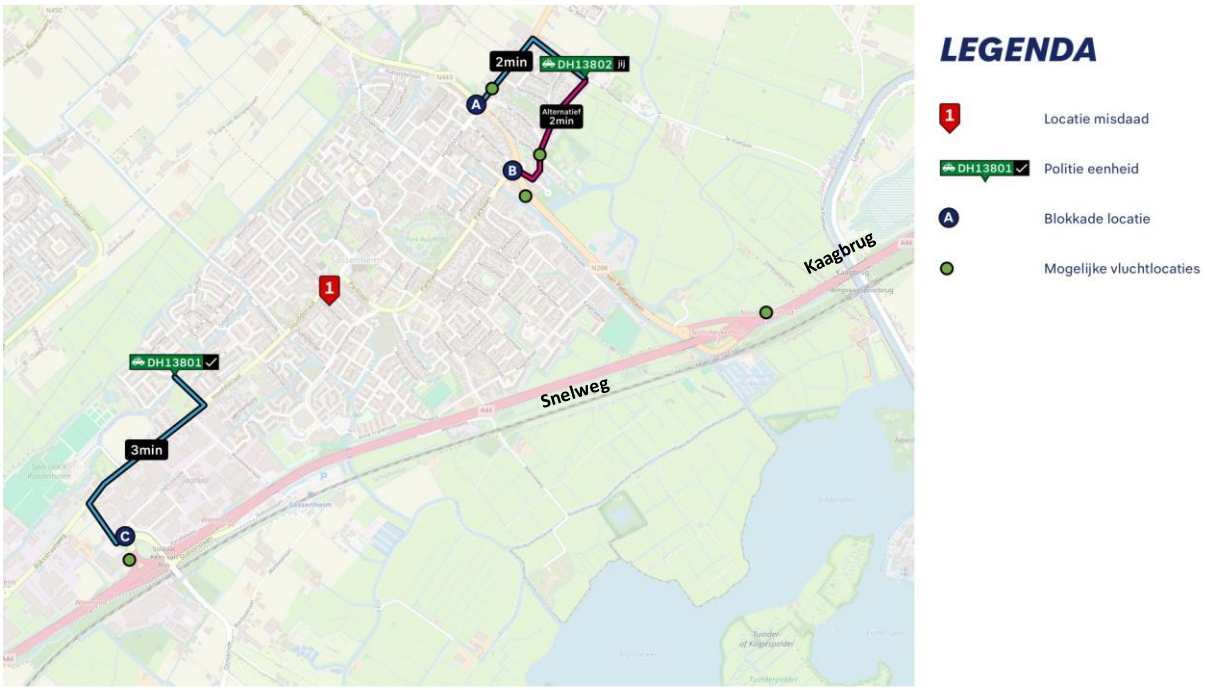
Jij bevindt je in de omgeving van deze inbraak en wordt gevraagd om te helpen met het dichtzetten van vluchtroutes. Je hebt van de meldkamer doorgekregen dat er maar één andere eenheid in de buurt van het incident is. **Jij bent de auto DH13802 (bovenste).**

Om de kans om de verdachten aan te houden te vergroten wordt de Hulp Onderschepping Verdachten (HOV) ingeschakeld. **Je team leider vraagt in dergelijke situaties nooit om je keuzes achteraf uit te leggen.**

De HOV geeft het volgende in de app weer:

- **Locatie A, locatie B** zijn mogelijke vluchtroutes die jij kan dekken**. Locatie C zal gedekt worden door één collega.**
- De HOV heeft **Locatie A** overwogen omdat dit de landroutes afsluit en de Kaagbrug op de snelweg momenteel openstaat.
- De HOV heeft **Locatie B** overwogen omdat dit de snelweg met verdere vluchtopties afsluit.
- **De HOV beveelt op basis van deze informatie locatie A aan.**

Onderstaand wordt een kaart getoond die de verschillende locaties van dit scenario weergeven. Wij vragen je deze kaart goed te bekijken.



Scenario Sassenheim: Groep 2 (Verantwoording)

Beeld je het volgende in: jij bent aan het surveilleren in Sassenheim. Om 20:57 krijg je een melding via de portofoon dat er een inbraak heterdaad is geweest. De bewoner van een woonhuis is thuisgekomen en heeft twee inbrekers betrapt. De verdachten zijn in een **Noord Oostelijke** richting gevlucht in donkerblauwe auto.

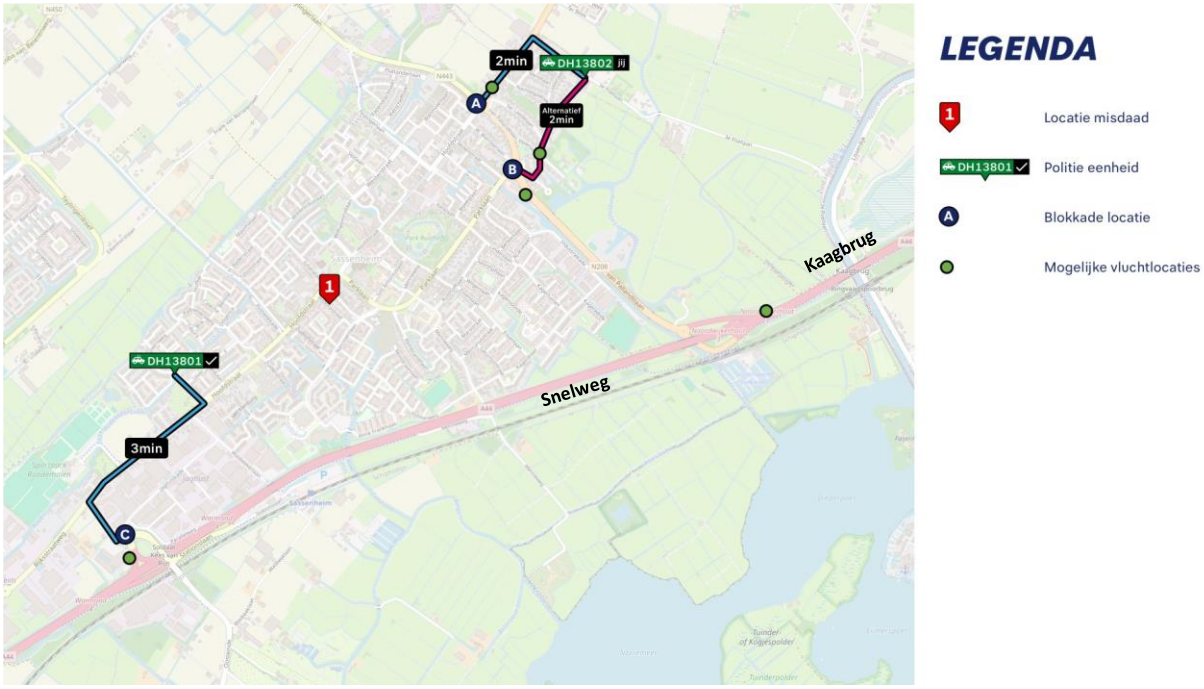
Jij bevindt je in de omgeving van deze inbraak en wordt gevraagd om te helpen met het dichtzetten van vluchtroutes. Je hebt van de meldkamer doorgekregen dat er maar één andere eenheid in de buurt van het incident is. **Jij bent de auto DH13802 (bovenste).**

Om de kans om de verdachten aan te houden te vergroten wordt de Hulp Onderschepping Verdachten (HOV) ingeschakeld. **Bij gebruik van de HOV vindt jouw teamleider het belangrijk dat jij je keuze achteraf verantwoord door uit te leggen waarom jij wel of niet hebt geluisterd naar het algoritme.**

De HOV geeft het volgende in de app weer:

- **Locatie A, locatie B** zijn mogelijke vluchtroutes die jij kan dekken**. Locatie C zal gedekt worden door één collega.**
- De HOV heeft **Locatie A** overwogen omdat dit de landroutes afsluit en de Kaagbrug op de snelweg momenteel openstaat.
- De HOV heeft **Locatie B** overwogen omdat dit de snelweg met verdere vluchtopties afsluit.
- **De HOV beveelt op basis van deze informatie locatie A aan.**

Onderstaand wordt een kaart getoond die de verschillende locaties van dit scenario weergeven. Wij vragen je deze kaart goed te bekijken.



1. **Op basis van het advies van de HOV en jouw eigen inschatting van de situatie, welke locatie kies je dan om jezelf op te stellen om vluchtroutes van de verdachten af te zetten?**

◯ Locatie A

◯ Locatie B

2. **Na afloop van het incident vraagt je teamleider waarom je hebt gekozen voor jouw gekozen locatie. Geef hieronder kort je uitleg aan die persoon [*Deze vraag komt alleen voor bij groep 2]*.**

[Vul hier je uitleg in]

3. **Je hebt de keuze gemaakt om naar [V1] te gaan. Hoe zeker ben je er van dat de verdachten daadwerkelijk langs deze route gevlucht zijn?**

Slider van 0 naar 100

Dank voor je deelname aan het onderzoek!

In dit onderzoek is getoetst wat het effect is van verantwoording afleggen op het vertrouwen in algoritmes. Van elk scenario (de liquidatie en de inbraak) zijn twee varianten, een waarin verantwoording afgelegd moet worden en een waarin dat niet nodig is. De aanbevelingen die het algoritme deed waren in de scenario's de 'juiste' antwoorden. Bij de liquidatie omdat je collega te laat aan kan komen bij de snelweg oprit en bij de inbraak omdat de open brug de snelweg al afzet.

De situaties die zijn gebruikt in dit onderzoek zijn fictief en ook de HOV is een fictief systeem. Ondanks het fictieve karakter van de situaties in deze enquête draagt jouw deelname direct bij aan de ontwikkeling van slimme algoritmen binnen de politie en wetenschappelijke kennis over dit onderwerp. Meer informatie over de ontwikkeling van slimme algoritme binnen de politie kun je vinden op de website van het Nationaal Politie lab AI (https://www.uu.nl/onderzoek/ai-labs/nationaal-politielab-ai).

Dit onderzoek wordt uitgevoerd door Mitch Opperhuizen in het kader van een masterscriptie bij de Universiteit Utrecht in samenwerking met het Nationaal Politie AI-Lab. Voor vragen over dit onderzoek kun je contact opnemen met de uitvoerder van dit onderzoek Mitch Opperhuizen (m.e.opperhuizen@students.uu.nl). Daarbij vervullen dr. Stephan Grimmelikhuijsen vanuit de Universiteit Utrecht en Bas Testerink vanuit de politie een begeleidende rol.