Op verkenning in de digitale frontlinie: de mogelijke toepassingen van kunstmatige intelligentie bij de Koninklijke Marechaussee

Jorrit Bootsma en Mariel van Staveren*

Artificial intelligence (AI) wordt steeds meer een onderdeel van het dagelijks leven. Of je nu naar je werk gaat met de auto of met de trein: Google Maps bepaalt de route en voorspelt de verkeersdrukte, allemaal middels AI.[1] Dit is slechts één van de vele voorbeelden van AI in ons dagelijks leven. Daarnaast dringt de technologie ook door in het veiligheidsdomein. Het veiligheidsdomein staat momenteel voor grote uitdagingen, zowel internationaal als nationaal. Conflicten zoals de oorlog in Oekraïne en het Israëlisch-Palestijns conflict zorgen voor grootschalige internationale disruptie. De veiligheid en stabiliteit van Europa, en daarmee Nederland, komen hierdoor in het geding. Daarnaast moet Nederland binnen zijn eigen grenzen reageren op toenemende dreiging en ondermijnende activiteiten vanuit de georganiseerde misdaad.[2]

Gelijktijdig met de veranderende dynamiek in het veiligheidsdomein bevindt de Defensie-organisatie zich op een kantelpunt wat betreft de visie op AI. Een groot deel van de discussie rondom AI gaat over de risico's die zij met zich meebrengt. Op hoog internationaal niveau vinden er bijeenkomsten plaats over dit onderwerp, zoals de REAIM Summit begin 2023 in Den Haag. Dit was de eerste wereldwijde conferentie over *Responsible AI in the Military Domain* (REAIM). Het betrof een tweedaagse bijeenkomst met zo'n 2.000 deelnemers vanuit de hele

* J.H. Bootsma MSc is freelance Data Scientist gespecialiseerd in Computer Vision. Majoor M.J. van Staveren MSc is stafadviseur informatievoorziening bij de Koninklijke Marechaussee. De auteurs schrijven deze bijdrage op persoonlijke titel.

1 Zie https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/.

2 Zie https://nos.nl/artikel/2324283-veel-meer-dreigingen-van-criminelen-politie-kan-beveiliging-nauwelijks-nog-aan.

doi: 10.5553/JV/016758502024050001004

wereld, met als doel om het verantwoorde gebruik van AI in het militaire domein hoger op de politieke agenda te zetten.[3] Internationele bijeenkomsten over AI zoals de REAIM Summit zijn bepalend voor de ontwikkelingen en inzet van AI in het Defensie-domein. Ze vormen echter slechts het begin. De huidige discussies worden vooral nog op een theoretisch niveau gevoerd over AI in het algemeen: 'Hoe kunnen we AI verantwoord en effectief inzetten?' Het antwoord op deze vraag is enorm divers vanwege de grote diversiteit in AI-technologieën en de talloze toepassingsgebieden. De toegevoegde waarde en de (veiligheids)implicaties van AI worden daarom pas echt duidelijk als er wordt gesproken over concrete toepassingen. Het helpt dan ook om in te zoomen op AI-toepassingen binnen, bijvoorbeeld, één krijgsmachtonderdeel.

Dit artikel beoogt inzicht te geven in mogelijke toepassingen van AI binnen de Koninklijke Marechaussee. Hiervoor is het belangrijk om de taken van de Marechaussee voor ogen te hebben. De Koninklijke Marechaussee waakt over de veiligheid van het Koninkrijk der Nederlanden.[4] Zij heeft drie hoofdtaken. De eerste is de grenspolitietaak: het bestrijden van grensoverschrijdende criminaliteit zoals terrorisme, drugssmokkel, identiteitsfraude en mensensmokkel. De tweede is bewaken en beveiligen: het beveiligen van objecten, personen en diensten die van cruciaal belang zijn voor de Nederlandse Staat. Ondanks dat de Marechaussee onderdeel is van Defensie voert zij de eerste twee taken uit onder gezag van het Ministerie van Justitie en Veiligheid. De derde taak betreft internationale en militaire politietaken: de Marechaussee treedt op als politie voor alle Nederlandse Defensie-onderdelen, ook in het buitenland en in conflictgebieden en in Nederland gevestigde internationale militaire hoofdkwartieren. Het takenpakket van de Marechaussee is zeer divers, zo ook de mogelijke toepassingen van AI. Om de mogelijke toepassingen te onderzoeken voert de Marechaussee diverse experimenten uit. Drie experimenten worden in dit artikel behandeld. Tot slot komen de uitdagingen vanuit het perspectief van de mogelijke toepassingen aan bod.

3 Zie www.defensie.nl/actueel/nieuws/2023/02/16/call-to-action-verantwoord-gebruik-ai-in-het-militaire-domein.
4 Zie www.marechaussee.nl/over-de-marechaussee.

doi: 10.5553/JV/016758502024050001004

Kansen van AI voor de Marechaussee

Net als de rest van de Defensie-organisatie moet de Koninklijke Marechaussee reageren op de veranderende dynamiek in het veiligheidsdomein. De Marechaussee zet groots in op informatiegestuurd en datagedreven optreden om slimmer te kunnen acteren op uiteenlopende dreigingen. Daarnaast wordt de Marechaussee door krapte op de arbeidsmarkt gedwongen om arbeidsextensiever te werken. Automatisering middels AI is hierin een veelbelovend stukje van de puzzel. Nieuwe technologie, waaronder AI, krijgt daarom bij de Marechaussee veel aandacht. Deze paragraaf geeft drie voorbeelden van hoe AI de Marechaussee zou kunnen helpen.

De virtuele grenswachter

De eerste mogelijke toepassing van AI bij de Marechaussee betreft automatisering van arbeidsintensieve processen binnen de grenspolitietaak op Schiphol. Een voorbeeld uit de praktijk is de selfservice-paspoortcontrole, waarbij het systeem op basis van gezichtsherkenningstechnologie controleert of het paspoort bij de reiziger hoort. Zo kunnen reizigers de grens passeren zonder tussenkomst van een menselijke grenswachter.[5] Voor bepaalde reizigers is alleen een paspoortcontrole niet voldoende en onderwerpt de Marechaussee hen aan een aantal vragen. Deze vragen gaan over het doel van de reis, de duur van de reis en de middelen die de reiziger heeft om de reis te bekostigen. Op basis van de antwoorden van de reiziger wordt besloten of diegene de grens mag passeren. Dit is een tijdrovend proces. De Marechaussee doet daarom onderzoek naar de inzet van taalmodellen (oftewel Large Language Models zoals ChatGPT) om gesprekken met reizigers te automatiseren. Het idee is dat een virtuele grenswachter in gesprek gaat met de reiziger om de benodigde informatie te verzamelen, en vervolgens de uitkomsten van het gesprek doorgeeft aan de 'echte' grenswachter. De grenswachter bepaalt vervolgens of de reiziger het land in mag of dat er een aanvullende controle nodig is. Een samenwerking tussen virtuele en menselijke grenswachters heeft een aantal voordelen. Ten eerste is er minder personeel nodig om alle gesprekken te voeren, terwijl de uiteindelijke toegang nog altijd door een mens

5 Zie www.marechaussee.nl/onderwerpen/selfservice-paspoortcontrole.

doi: 10.5553/JV/016758502024050001004

wordt goedgekeurd. Dat betekent dat de virtuele grenswachter meer tijd aan één reiziger kan besteden dan de menselijke grenswachter, zodat alle relevante informatie verzameld kan worden. Ten tweede komen alleen de opvallende reizigers terecht bij de menselijke grenswachter voor aanvullende controle, wat de taak voor de grenswachters diverser en uitdagender maakt. Ten derde spreekt de digitale grenswachter alle talen, waardoor taalbarrières geen negatieve invloed hebben op de kwaliteit van het gesprek.
Het onderzoek naar de virtuele grenswachter richt zich niet alleen op de achterliggende taalmodellen, maar ook op de manier waarop de virtuele grenswachter aan de reiziger gepresenteerd wordt. Het is belangrijk dat reizigers begrijpen dat ze met een controle van de Marechaussee te maken hebben, en dat ze zich voldoende op hun gemak voelen om het gesprek aan te gaan met de virtuele grenswachter. Het onderzoek bestaat daarom ook uit de ontwikkeling van een herkenbare holografische avatar van een grenswachter van de Marechaussee. Naast de visuele presentatie middels de holografische avatar is het belangrijk dat de virtuele grenswachter het werk op een professionele en correcte manier uitvoert. Het taalmodel moet een bepaalde rol spelen en kennis hebben over alle wet- en regelgeving en procedures die de Marechaussee hanteert aan de grens. Het taalmodel moet dus worden gevoed met de juiste informatie en het moet deze informatie op de juiste momenten inzetten. Daarnaast moet het taalmodel, net als de menselijke grenswachter, een inschatting kunnen maken van de betrouwbaarheid van de informatie die de reiziger verstrekt. Dit zijn de punten waar het meeste werk in zit. Het is duidelijk dat moderne taalmodellen heel krachtig zijn, maar de correcte en verantwoorde vertaling naar de werkvloer is uiteindelijk bepalend.

Slimme analyse van sensordata

De tweede mogelijke toepassing van AI betreft slimme sensoranalyse ter ondersteuning van de tweede hoofdtaak: bewaken en beveiligen (B&B). Het domein B&B staat onder druk vanwege de verharding in de georganiseerde criminaliteit en de lagere drempels voor het uiten van geweld. Het gevolg is dat meer mensen beveiligd moeten worden dan

doi: 10.5553/JV/016758502024050001004

voorheen.[6] Om met deze ontwikkelingen om te kunnen gaan moet het stelsel B&B een flinke transformatie ondergaan. Er zal efficiënter moeten worden omgegaan met personele en materiële capaciteit. Het maken van de juiste strategische en operationele keuzes kan alleen als er een sterke informatiepositie is over de dreiging. Sensordata, bijvoorbeeld videobeelden van bewakingscamera's, zijn een belangrijke bron voor de opbouw van een dreigingsbeeld. De analyse van die sensordata gebeurt echter grotendeels nog handmatig, waardoor slechts een beperkte hoeveelheid data geanalyseerd kan worden. Daarom worden beelden van beveiligingscamera's doorgaans vooral gebruikt om een incident na afloop te analyseren. Een deel van de beelden wordt wel realtime uitgekeken, maar de praktijk laat zien dat mensen weinig geschikt zijn voor dergelijke langdurige visuele taken. Bij het uitkijken van bagagescans op Schiphol, bijvoorbeeld, worden de medewerkers daarom elke 20 minuten gewisseld. Bovendien kunnen er met geautomatiseerde analyse een stuk meer beelden worden gemonitord, waardoor meer mensen (op afstand) beveiligd kunnen worden.

De Marechaussee doet onderzoek naar de inzet van AI voor geautomatiseerde analyse van sensordata. In dit onderzoek ligt de focus op de analyse van camerabeelden. Dat type AI heet Computer Vision. Er wordt onderzocht hoe Computer Vision kan helpen bij de realtime-analyse van camerabeelden, bijvoorbeeld het detecteren van objecten of patronen die van betekenis zijn in een beveiligingscontext. Verschillende soorten AI worden onderzocht: van de meer volwassen technologie voor objectdetectie tot meer experimentele technologie voor het detecteren van afwijkingen op basis van een normaalbeeld. Op basis van tests in operationele setting wordt bepaald welke vormen van AI de meeste waarde hebben voor de Marechaussee.

Het idee van het detecteren van patronen op camerabeelden middels AI is overigens niet nieuw. Enkele jaren geleden is op Schiphol onderzoek gedaan, door de Marechaussee en TNO, naar de inzet van AI voor het detecteren van flauwvallende mensen en achtergelaten koffers.[7] Het bleek dat de AI-technologie nog niet goed genoeg was en te veel valse meldingen genereerde. Een systeem dat veel valse meldingen

6 Zie het rapport van de adviescommissie toekomstbestendig stelsel bewaken en beveiligen: www.rijksoverheid.nl/documenten/rapporten/2021/10/27/tk-bijlage-rapport-adviescommissie-toekomstbestendig-stelsel-definitief.

7 Zie https://fd.nl/bedrijfsleven/1507865/experts-zien-praktische-hobbels-bij-inzet-ai-tegen-winkeldiefstal?gift=SRM34.

doi: 10.5553/JV/016758502024050001004

genereert, is in de praktijk onbruikbaar. De uitdaging is om de juiste balans te vinden: aan de ene kant wil je dat alle relevante patronen gedetecteerd worden (de zogenaamde sensitiviteit van een AI-model), aan de andere kant wil je valse meldingen voorkomen (de zogenaamde specificiteit). Deze balans zal ook per locatie verschillen: in het centrum van Amsterdam zien de relevante patronen er anders uit dan op het platteland in Groningen. Per locatie zal deze balans dus opnieuw moeten worden opgemaakt, in nauwe samenwerking met de operatie.

Autonome robotica

De derde mogelijke toepassing van AI is het (gedeeltelijk) autonoom laten optreden van fysieke robots. De Marechaussee bewaakt door het hele land belangrijke locaties zoals koninklijke paleizen, overheidsgebouwen en bepaalde maatschappelijke instellingen. Het continu bewaken van een groot gebied vergt veel personele capaciteit. De Marechaussee onderzoekt daarom hoe robots kunnen worden ingezet om menselijke beveiligers te ondersteunen. Mensen en robots hebben andere talenten, en voor een succesvolle man-machinesamenwerking is het van belang om die talenten een aanvulling op elkaar te laten zijn. Robots zijn bijvoorbeeld, in vergelijking met de mens, heel goed in het uitvoeren van repetitieve taken en kunnen uitgerust worden met 'zintuigen' die de mens niet heeft (bijvoorbeeld een infraroodsensor). Een ander voordeel is dat robots ingezet kunnen worden in gevaarlijke situaties, zodat mensen minder risico hoeven te lopen. Een voorbeeld daarvan is het betreden van een gedeeltelijk verwoest gebouw bij een reddingsactie of het betreden van een ruimte waar mogelijk giftige of explosieve stoffen aanwezig zijn.
Op dit moment functioneren de meeste roboticatoepassingen alleen goed in een voorspelbare en statische omgeving. In een beveiligingscontext is de omgeving per definitie onvoorspelbaar en dynamisch. Om dit probleem op te lossen is de Marechaussee samen met de Technische Universiteit Delft, de Universiteit van Amsterdam en TNO een vierjarig Open Technology Programme (OTP) gestart vanuit de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). Dit langdurige wetenschappelijke programma onderzoekt hoe robots voldoende autonoom kunnen worden gemaakt zodat ze kunnen opereren in de onvoorspelbare beveiligingscontext. AI-technologie speelt

doi: 10.5553/JV/016758502024050001004

een rol in alle facetten van dit onderzoek: perceptie (het waarnemen van de omgeving waar de robot zich bevindt), redenatie (het begrijpen van de omgeving en eventueel het probleem dat opgelost moet worden), planning (het bepalen van de juiste acties) en uitvoering.[8]
Het OTP houdt zich bezig met fundamenteel wetenschappelijk onderzoek. Dit programma zal dan ook niet binnen een jaar leiden tot een operationeel inzetbare robot voor de Marechaussee. De behoefte van de Marechaussee aan arbeidsextensief werken is echter groot, daarom doet zij tegelijkertijd onderzoek naar eenvoudigere roboticaoplossingen die op de korte termijn (namelijk binnen een jaar) ingezet kunnen worden. In dat onderzoek ligt de focus niet op AI, maar op de inbedding van volwassen roboticaoplossingen in operationele processen, en op het definiëren van operationele scenario's waar robotica ondersteuning kan bieden. Deze lessen uit de praktijk zullen ook worden ingebracht in het OTP om de onderzoeksvragen aan te scherpen.

De virtuele grenswachter, de slimme analyse van sensordata en de autonome robotica illustreren de mogelijke toepassingen van AI voor de Marechaussee. Verschillende soorten AI kunnen worden ingezet voor zeer uiteenlopende taken en processen. Per toepassing zullen de risico's en potentiële problemen moeten worden geïdentificeerd. De volgende paragraaf gaat dieper in op deze risico's en potentiële problemen.

Potentiële problemen rondom de inzet van AI

De inzet van AI-toepassingen in het publieke domein is niet zonder risico's. In deze paragraaf behandelen we vier potentiële problemen die komen kijken bij de inzet van AI in het publieke domein. Deze potentiële problemen illustreren we aan de hand van voorbeelden uit de praktijk. Vervolgens wordt per potentieel probleem aangegeven hoe dit terugkomt in AI-toepassingen die verkend worden door de Marechaussee en hoe dit gemitigeerd kan worden.
De vier potentiële problemen die we behandelen, zijn:
1. een gebrek aan effectiviteit van de toepassing in het bereiken van het beoogde doel;

8 Zie www.nwo.nl/nieuws/zeven-toepassingsgerichte-projecten-van-start-via-het-open-technologieprogramma.

doi: 10.5553/JV/016758502024050001004

2. een risico op bevooroordeeldheid van de toepassing jegens verschillende bevolkingsgroepen;
3. een gebrek aan transparantie over hoe voorspellingen tot stand komen; en
4. een gebrek aan juridische kaders op het gebied van AI in het veiligheidsdomein.

Gebrek aan effectiviteit

Een potentieel probleem bij de inzet van AI-toepassingen is het gebrek aan effectiviteit in het bijdragen aan het bereiken van het beoogde doel. Wanneer een toepassing niet bijdraagt aan het bereiken van het beoogde doel heeft een toepassing geen nut. Een voorbeeld van een dergelijke toepassing in de praktijk is het COMPAS-algoritme dat wordt toegepast in het rechtssysteem in de Verenigde Staten. Dit algoritme adviseert rechters door te voorspellen hoe waarschijnlijk het is dat een gedaagde recidiveert. Onderzoekers vergeleken de voorspellingen van het algoritme met voorspellingen van een groep mensen (Dressel & Farid 2018). Deze groep mensen hadden de auteurs via een *crowdsource platform* gerekruteerd en zij hadden geen specifieke juridische kennis. Er bleek geen significant verschil te zitten in de kwaliteit van de voorspellingen van het algoritme en de kwaliteit van de voorspellingen van deze groep mensen. Kortom, het advies van een willekeurige groep mensen heeft evenveel waarde als de uitkomst van het algoritme. Zie de linker staafdiagram in figuur 1 (p. 57) voor de resultaten van de vergelijking. De effectiviteitsvraag is hier: in hoeverre draagt het advies van dit algoritme bij aan het bepalen van de juiste strafmaat? Dit wordt mede bepaald door hoe rechters dit advies meewegen in hun vonnis. Het is voorstelbaar dat rechters advies afkomstig van een willekeurige groep mensen anders zouden wegen dan advies van een zogenaamd intelligent algoritme.
Dit voorbeeld laat zien hoe belangrijk het is dat de effectiviteit van een AI-systeem meetbaar wordt gemaakt. In het geval van de slimme sensoranalyse is de vraag: in welke mate vergroot de slimme sensoranalyse de veiligheid rondom een gebouw of persoon? Het antwoord op deze effectiviteitsvraag bepaalt mede of de inbreuk van het middel op de privacy van mensen proportioneel is.
Het meten van de effectiviteit van een systeem kan bijvoorbeeld worden gedaan door de toepassing in een schaduwomgeving in te zetten.

doi: 10.5553/JV/016758502024050001004

Parallel aan de huidige systemen en werkwijze wordt het systeem dan toegepast op operationele data, zonder acties te verbinden aan de voorspellingen van het systeem. Door vooraf criteria te definiëren die de veiligheid van een situatie representeren, en door de volledige werking van het systeem vast te leggen kan worden bepaald in welke mate het systeem bijdraagt aan de veiligheid ten opzichte van de huidige methoden. Een uitdaging, in dit geval, is dat onveilige situaties zich sporadisch voordoen en daarnaast verschillende verschijningsvormen hebben. Hierdoor zal het lastig zijn om statistisch valide resultaten te behalen. Naast een dergelijke kwantitatieve analyse kan er daarom ook een kwalitatieve analyse worden gedaan. Om een kwalitatief beeld te krijgen van de effectiviteit van een systeem moet er nauw worden samengewerkt met de mensen die het systeem gebruiken. Hoe ervaren zij het werken met het systeem? Welke impact heeft het op hun werkwijze? Denk hierbij bijvoorbeeld aan het signaleren van de overdaad aan valse meldingen van het systeem op Schiphol.

Risico op bevooroordeeldheid van een systeem

De Nederlandse Grondwet (Gw) waarborgt onze democratische rechtsstaat[9] en artikel 1 Gw verbiedt discriminatie.[10] Het belang van non-discriminatie van AI-systemen is daarmee niet te overschatten. Desalniettemin zijn AI-systemen vaker wel dan niet bevooroordeeld. Dit komt doordat de vooroordelen ingebakken zitten in de data waarmee de AI-systemen getraind worden.
Het potentiële probleem van bevooroordeeldheid is, net als het gebrek aan effectiviteit, te illustreren aan de hand van het COMPAS-algoritme dat het risico op recidive bij gedaagden in de Verenigde Staten voorspelt. ProPublica legde bloot dat dit algoritme bevooroordeeld is als het gaat om etniciteit van gedaagden.[11] Wanneer het een zwárte gedaagde betrof, dan gaf het algoritme váker ónterecht aan dat er een hoog risico was dat de gedaagde zou recidiveren dan wanneer het een wítte gedaagde betrof. Met andere woorden: foutpositieve voorspellingen kwamen vaker voor bij zwarte gedaagden dan bij witte gedaagden. Daarnaast gaf het algoritme váker ónterecht aan dat er een laag risico was dat een gedaagde zou recidiveren wanneer het een wítte gedaagde

9 Zie www.denederlandsegrondwet.nl/id/vlxuq54is4ik/algemene_bepaling.
10 Zie www.denederlandsegrondwet.nl/id/vlxups19rfoe/hoofdstuk_1_grondrechten.
11 Zie www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

doi: 10.5553/JV/016758502024050001004

betrof dan wanneer het een zwárte gedaagde betrof. Met andere woorden: foutnegatieve voorspellingen kwamen vaker voor bij witte gedaagden dan bij zwarte gedaagden. Zie het middelste en het rechter staafdiagram in figuur 1 voor een visuele weergave van het verschil tussen deze twee groepen.

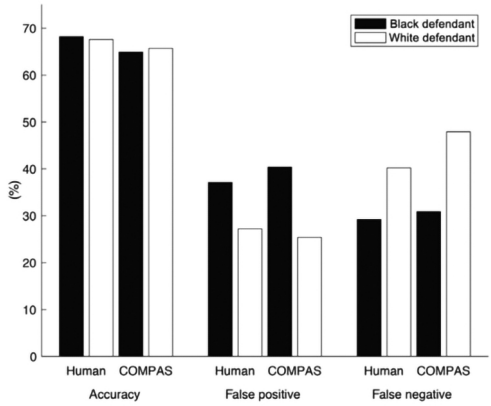
Het risico op een bevooroordeeld systeem speelt onder andere bij de analyse van camerabeelden ten behoeve van B&B. De gebruikte AI-modellen zijn voorgetraind ('pre-trained') op data die van het internet zijn geschraapt en bevatten daarmee allerlei menselijke stereotypen, vooroordelen en over- en ondervertegenwoordigingen. Na het 'pre-trainen' volgt een stap van 'finetunen'. Dit is het verder trainen van het model op taakspecifieke data zodat voorspellingen toegespitst zijn op de beoogde taak. Het verkrijgen van een goed gebalanceerde, taakspecifieke dataset is een grote uitdaging. Een mogelijke oplossing is het gebruik van synthetische data. Dit zijn computergegeneerde (nep)data waarbij op het moment van generatie uiterlijke kenmerken kunnen worden gerandomiseerd. Hierbij kunnen zelfs uiterlijke kenmerken worden gebruikt die niet bestaan, neem bijvoorbeeld een knalblauwe huidskleur. Dit zal ertoe leiden dat het AI-model leert dat uiterlijke kenmerken geen informatie bevatten over, in dit geval, de veiligheid van een situatie. Recent onderzoek (Wang & Russakovsky 2023, p. 3934) toont aan dat door een dergelijke vooroordeelvrije dataset te gebruiken bij het 'finetunen' van het model vooroordelen afkomstig uit de 'pre-training' zelfs overschreven zouden kunnen worden.

Naast het tegengaan van vooroordelen bij het ontwikkelen van een systeem is het nodig om blijvend te monitoren hoe uitkomsten van het systeem verdeeld zijn over verschillende groepen. Alleen zo kun je er zeker van zijn dat het systeem, ook in de praktijk, niet discrimineert. Voor inspiratie hoe de governance rondom de monitoring van modellen ingericht kan worden, kan worden gekeken naar de bankensector. Sinds de bankencrisis van 2007 dienen banken aan de Europese Centrale Bank (ECB) en De Nederlandsche Bank (DNB) aan te tonen dat de inzet van bijvoorbeeld een kredietrisicomodel gepast is.[12]

12 Zie www.bankingsupervision.europa.eu/legalframework/publiccons/pdf/ssm.pubcon230622_guide.en.pdf.

doi: 10.5553/JV/016758502024050001004

Figuur 1 **Onderzoeksresultaten recidivismevoorspellingen van COMPAS**



Bron: Dressel & Farid 2018

Figuur 1 toont drie staafdiagrammen met de uitkomsten van het COMPAS-systeem, dat de kans op recidive voorspelt voor gedaagden in de Verenigde Staten. Het linker staafdiagram geeft de overeenkomst in accuraatheid weer tussen de voorspellingen gedaan door een groep mensen zonder specifieke juridische kennis en het COMPAS-systeem. Het middelste staafdiagram geeft het verschil weer in foutpositieve voorspellingen tussen zwarte en witte gedaagden. Het rechter staafdiagram laat ditzelfde zien voor foutnegatieve voorspellingen. 'False positive' betekent hier: er is voorspeld dat een gedaagde zal recidiveren, maar dat gebeurt niet, en 'False negative': er is voorspeld dat een gedaagde níét zal recidiveren, maar het gebeurt wel.

Gebrek aan transparantie

Een derde potentieel probleem bij de inzet van AI-toepassingen is een gebrek aan transparantie in hoe beslissingen tot stand komen. Wanneer de werking van een systeem niet duidelijk is voor de gebruikers of degenen die erdoor beïnvloed worden, kan dit leiden tot wantrouwen en bemoeilijkt dit het maken van bezwaar tegen de beslissingen (Kulk & Van Deursen 2020, p. 138, Rudin e.a. (2020), p. 27). Wederom is dit te

doi: 10.5553/JV/016758502024050001004

illustreren aan de hand van het algoritme dat de kans op recidive voorspelt in de Verenigde Staten. Dit algoritme is ontwikkeld door een privaat bedrijf dat de werking van het algoritme beschouwt als bedrijfsgeheim. Dit beperkt gedaagden om inhoudelijk te reageren op hun risicoscore. Naast een vorm van gebrek aan transparantie als gevolg van bedrijfsgeheim en intellectueel eigendom, bestaat het risico op gebrek aan transparantie als gevolg van de complexiteit van de modellen. Deze complexiteit zorgt ervoor dat het lastig uit te leggen is hoe de uitkomst van een dergelijk model tot stand komt. Ook wanneer het model openbaar is.
Wat betreft de toepassingen die de Marechaussee verkent, speelt het belang van transparantie een belangrijke rol bij de virtuele grenswachter. Wanneer een persoon de toegang tot het land ontzegd wordt, moet door de 'echte' grenswachter uit te leggen zijn waarom dit het geval is. Een simpel algoritme is beter uitlegbaar dan een complex algoritme, maar simplificeren kan ten koste gaan van de accuraatheid. Dit is een afweging die gemaakt moet worden.
In het geval van de inzet van autonome robotica om een terrein te inspecteren is het belang van transparantie richting burgers minder groot. Een dergelijke toepassing heeft namelijk geen directe invloed op het leven van burgers. Ditzelfde geldt voor de toepassing van slimme sensoranalyse ten behoeve van de opbouw van een dreigingsbeeld. Een zekere mate van inzicht in hoe uitkomsten tot stand komen is desalniettemin vereist om een toepassing effectief in te kunnen zetten. Hiervoor kan worden gekeken naar de ontwikkelingen op het gebied van Explainable AI. Dit zijn technische mogelijkheden om complexe algoritmen uitlegbaar te maken (zie bijvoorbeeld Ras e.a. (2022) voor een uitgebreid overzicht). Deze technieken zijn echter (nog) niet van een dergelijk niveau dat zij de uitdagingen rondom uitlegbaarheid van algoritmen richting burgers het hoofd kunnen bieden.

Gebrek aan juridische kaders op het gebied van AI in het veiligheidsdomein

Kulk en Van Deursen (2020) deden onder andere onderzoek naar de bestendigheid van de juridische kaders om kansen te verwezenlijken en risico's te mitigeren bij algoritmen die besluiten nemen. Zij concluderen dat 'de algemene juridisch kaders (…) voldoende in staat zijn om publieke waarden en belangen te borgen' en dat de breed en open

doi: 10.5553/JV/016758502024050001004

geformuleerde normen het mogelijk maken om de juridische kaders geleidelijk en flexibel vorm te geven. Maar, stellen zij, 'daarvoor is wel vereist dat de nodige rechtsontwikkeling plaatsheeft, door bijvoorbeeld rechters en toezichthouders, waarmee een op algoritmen toegesneden uitleg of interpretatie wordt gegeven aan deze algemene kaders'. Deze toegesneden uitleg of interpretatie ontbreekt en dat is in de hedendaagse maatschappij een uitdaging waar meerdere organisaties voor staan. Zie bijvoorbeeld de uitzending van *Nieuwsuur* van 5 januari 2024.[13] Hierin legt Theo van der Plas (portefeuillehouder digitalisering Politie) uit dat voor de rechter niet aan te tonen is dat datgene wat gebruikt wordt uit het Wetboek van Strafvordering of het Wetboek van Strafrecht voldoet aan de criteria, omdat deze criteria ontbreken. Het gevolg is dat organisaties opereren aan de hand van zelf opstelde richtlijnen en kaders.
Hierop aanvullend: AI-toepassingen kunnen gebruik maken van gevoelige gegevens en daarmee is de vergelijking te maken met het gebruik van andersoortige gevoelige persoonsgegevens. Denk aan vingerafdrukken en DNA. Wanneer en hoe deze persoonsgegevens gebruikt mogen worden, is vastgelegd in wetgeving. Kortom, dat is inmiddels goed geregeld, aldus Aleid Wolfsen, voorzitter van de Autoriteit Persoonsgegevens. Voor gezichtsherkenning, bijvoorbeeld, is dit echter niet het geval.
De AI Act is (op het moment van schrijven) in aantocht en is een stap in de goede richting. Echter, dit zal geen heilige graal blijken die alle onduidelijkheid weet weg te poetsen. Hiervoor zal jurisprudentie nodig zijn. Zoals Kulk & van Deursen (2020, p. 195) ook concluderen: 'daar waar knelpunten worden ervaren, […] deze knelpunten zo veel mogelijk domeinspecifiek aan te pakken'. De AI Act richt zich op hoofdlijnen en daarmee is deze uitdaging er eentje waar het laatste woord nog niet over geschreven is.

Naast de vier potentiële problemen van een gebrek aan effectiviteit, een risico op bevooroordeeldheid, een gebrek aan transparantie en een gebrek aan juridische kaders zijn er meer potentiële problemen te noemen bij de inzet van AI-systemen. Denk bijvoorbeeld aan afhankelijkheid van private partijen wanneer deze de AI-systemen leveren,

13 Zie https://nos.nl/nieuwsuur/artikel/2503831-politie-experimenteert-met-gezichtsherkenning-maar-wetgeving-ontbreekt.

doi: 10.5553/JV/016758502024050001004

maar ook aan het verlies van de menselijke maat[14] en het risico op onvoorspelbare uitkomsten.[15] Hier zal dit artikel echter niet verder op ingaan.

Conclusie

Het moge duidelijk zijn dat AI niet meer weg te denken is uit onze samenleving, en ook niet uit het veiligheidsdomein. Dit artikel bespreekt de mogelijke toepassingen van AI voor de Koninklijke Marechaussee, en de uitdagingen die daaraan verbonden zijn. AI is een veelbelovende oplossing voor de uitdagingen waar de Marechaussee voor staat en er wordt dan ook actief onderzoek naar gedaan. Drie voorbeelden passeren de revue: de inzet van taalmodellen om grenscontroles op Schiphol te automatiseren, de inzet van Computer Vision voor geautomatiseerde monitoring van beveiligingsbeelden en de inzet van verschillende vormen van AI voor het ontwikkelen van autonome robotica ten behoeve van beveiligingstaken.
Potentiële problemen zoals gebrek aan effectiviteit van een AI-systeem, het risico op bevooroordeeldheid van een AI-systeem en gebrek aan transparantie van een AI-systeem moeten worden gemitigeerd. Elk van deze problemen kent zijn eigen oplossingsrichtingen, maar overkoepelend geldt dat vastlegging en monitoring van een systeem en zijn voorspellingen essentieel zijn. Daarnaast moeten beslissingen waarbij AI-systemen een rol spelen zo veel mogelijk uitlegbaar zijn voor betrokkenen, zodat men zich, waar nodig, inhoudelijk kan verweren.
Een andere uitdaging is dat de Koninklijke Marechaussee, net als de rest van het publieke domein, te maken heeft met de impasse van (a) technologie niet willen inzetten vanwege het gebrek aan jurisprudentie en (b) ertegen aanlopen dat jurisprudentie pas komt wanneer technologie daadwerkelijk wordt ingezet. Deze impasse kan alleen worden doorbroken door gecontroleerd te experimenteren en de bestaande kaders actief ter discussie te stellen.

14 Zie het rapport *Blind voor mens en recht*: www.tweedekamer.nl/kamerleden_en_commissies/commissies/pefd.
15 Zie www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know.

doi: 10.5553/JV/016758502024050001004

Literatuur

Dressel & Farid 2018
J. Dressel & H. Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* (4) 2018, afl. 1, DOI: 10.1126/sciadv.aao5580.

Kulk & Van Deursen 2020
S. Kulk & S. van Deursen, *Juridische aspecten van algoritmen die besluiten nemen. Een verkennend onderzoek*, Den Haag: WODC 2020.

Ras e.a. 2022
G. Ras, N. Xie, M. van Gerven & D. Doran, 'Explainable deep learning: a field guide for the uninitiated', *Journal for Artificial Intelligence Research* 2022, p. 329-396.

Rudin e.a. 2020
C. Rudin, C. Wang & B. Coker, 'The age of secrecy and unfairness in recidivism prediction', *Harvard Data Science Review* (2) 2020, afl. 1, DOI: 10.1162/99608f92.6ed64b30.

Wang & Russakovsky 2023
A. Wang & O. Russakovsky, 'Overwriting pretrained bias with finetuning data', *International Conference on Computer Vision* 2023, p. 3934-3945.

doi: 10.5553/JV/016758502024050001004