Dit is de geaccepteerde versie van het artikel:

Testerink, B., Nieuwenhuizen, E. N., & Bex, F. J. (2023). Wat doet het ertoe dat je een mens bent? Autonome AI systemen voor de politie. In T. Snaphaan, W. Hardyns, A. J. van Dijk, R. Spithoven & R. Van Brakel (Eds.), Big data policing (pp. 121–134). Gompel&Svacina.

Wat doet het ertoe dat je een mens bent? Autonome AI systemen voor de politie

Bas Testerink, Esther Nieuwenhuizen, Floris Bex[1]

De groeiende digitale voetafdruk van de samenleving levert de politie grote uitdagingen op. Grote hoeveelheden data doorzoeken op opsporingsindicaties, belastend en ontlastend bewijs en intelligence vereist een enorme hoeveelheid mentale arbeid, en de komende jaren gaan duizenden medewerkers van de politie met pensioen. Mede hierom is er bij de politie veel aandacht voor Artificiële Intelligentie (AI) als onderdeel van de oplossing. Door mentale arbeid uit te besteden aan AI zal er al snel sprake zijn van autonome functionaliteiten. Machine autonomie is echter een gevoelig onderwerp. In dit artikel stellen we de vraag centraal hoe de Nederlandse politie omgaat met AI-autonomie en hoe de politie kan zorgen dat AI-autonomie geen afbreuk doet aan het vertrouwen in de politie.

Introductie

De Nederlandse politie is een voorloper op het gebied van AI.[2] Naast het gebruik van slimme algoritmen om, bijvoorbeeld, burgers te informeren[3], zet de politie ook steeds vaker *data mining* en AI in om in grote datasets te zoeken naar, bijvoorbeeld, opsporingsindicaties en bewijzen.[4] Hoewel AI en slimme technologie regelmatig als hulpmiddel wordt genoemd[5], zijn er ook veel zorgen rondom de inzet van AI, en meer in het bijzonder de inzet van AI-systemen die enige autonomie vertonen. Overheidsrichtlijnen stellen dat *"Hoe groter de impact en autonomie van het algoritme, hoe stringenter de richtlijnen zullen moeten worden toegepast"*[6], en zowel de AVG als de aankomende EU AI-wetgeving stellen strenge eisen aan systemen welke zonder enige menselijke tussenkomst beslissingen nemen. In het defensiedomein zien

[1] Medewerkers Nationaal Politielab AI, Nationale Politie.

[2] De Nederlandse politie had wereldwijd het eerste "AI Politielab" (https://icai.ai/police-lab-ai/) en is naast Australië (https://ailecs.org/) nog steeds één van de weinige korpsen met een speciaal lab toegewijd aan AI.

[3] Zoals bijvoorbeeld via een chatbot of een keuzehulp. Zie: https://www.politie.nl/nieuws/2019/oktober/9/cybercrime-melden-aan-virtuele-agent-wout.html en https://www.politie.nl/nieuws/2019/september/26/sneller-duidelijkheid-bij-aangifte-internetoplichting.html.

[4] We bedoelen met de term AI zowel *machine learning* algoritmen als (meer klassieke) regel gebaseerde algoritmen.

[5] In de eindrapportage van de Commissie Schneiders (2022) wordt 'datagedreven werken' als essentieel benoemd.

[6] Pag. 14, Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses (Ministerie van Justitie en Veiligheid, 2021); Proposal for a Regulation laying down harmonised rules on artificial intelligence (European Commission, 2021)

we ook terughoudendheid wat betreft autonome technologie, met zelfs een roep om een verbod op autonome wapensystemen.[7] Dit onbehagen rondom autonomie van machines leeft ook onder experts. Zo hebben ruim 170 Nederlandse academici met specialisaties in AI en robotica een brief ondertekend specifiek tegen autonome wapens, maar niet tegen wapens met AI in het algemeen.[8]

Naast autonomie wordt vaak bias genoemd als risicofactor van AI.[9] Daarbij is het risico concreet: bias kan bijvoorbeeld leiden tot ongewenste discriminatie. Ook is er veel geschreven over de verschijningsvormen van bias, de detectie van ongewenste bias en mitigerende maatregelen tegen de ongewenste effecten van bias (Mehrabi et al, 2021). Voor het concept autonomie is dit veel minder het geval.[10] Dat een zogenaamde *human-in-the-loop* wenselijk is, ook bij gebruik van AI door de politie, is meermaals geopperd.[11] Echter, precies *waarom* we deze human-in-the-loop willen blijft vaak onduidelijk. Ook vindt men het lastig om te verwoorden waarom, bijvoorbeeld, een (autonome) politiehond inzetten niet inherent problematisch is, maar een autonome robothond wel.[12] Blijkbaar heerst er een ongemak over de autonomie van systemen, maar kunnen we zelden concrete problemen van autonome AI benoemen die specifiek door autonomie veroorzaakt worden en niet door, bijvoorbeeld, een bias, onnauwkeurigheid of programmeerfout in het algoritme. Er zijn veel minder handvatten beschikbaar dan bij bias om na te denken over de verschijningsvormen van autonomie, de detectie van ongewenste autonomie en mitigerende maatregelen tegen de ongewenste effecten van autonomie.

Als samenleving moeten we bepalen wat uiteindelijk gewenste en ongewenste autonomie in onze systemen is. Als er situaties zijn waar we mensen over machines prefereren, dan doet het er blijkbaar toe dat we mensen zijn. Maar waarom dan? Waarom wantrouwen we een systeem meer dan een mens? Hoewel wij niet de illusie hebben in dit artikel een sluitend antwoord op deze fundamentele vragen te geven, willen we een constructieve dialoog aanmoedigen door inzicht te geven in autonome systemen bij de politie, en hoe hier bij de politie mee wordt omgegaan. Onze meer specifieke vragen zijn dan ook: hoe gaat de Nederlandse politie om met AI-autonomie, en hoe kan de politie zorgen dat AI-autonomie geen afbreuk doet aan het vertrouwen in de politie?

De opbouw van dit artikel is als volgt. Eerst bespreken we de voorkomens van autonome AI bij de Nederlandse politie en geven we handvatten om over verschillende soorten autonomie na te denken. Daarna gaan we in op de risico's die autonomie met zich meeneemt en voorbeelden van maatregelen bij de politie om deze risico's te verkleinen. Als laatste bespreken

[7] Zie bijvoorbeeld Kamerstukken/2021, 35848, nr. 2 ("Autonome Wapensystemen").
[8] Zie Open letter from scientists to the Dutch government on autonomous systems. https://drive.google.com/file/d/1ubAli1csfahmAuzLHkgXoA9VthPbZ2xA/view.
[9] Bias is de vooringenomenheid van een systeem. Bijvoorbeeld bias in gezichtsherkenning kan ertoe leiden dat mensen met een lichte huidskleur beter herkend worden dan mensen met een donkere huidskleur. Voor een uitgebreid overzicht van de precieze technische definities van bias in AI, zie (Mehrabi et al, 2021). Bias als risico van AI wordt in algemene zin genoemd door o.a. Osoba & Welser (2019), en specifiek in de context van de (Nederlandse) politie door Dechesne e.a. (2019). Verder wordt bias als een mogelijk risico van AI genoemd door zowel de Nederlandse overheid (zie richtlijnen van Ministerie van Justitie en Veiligheid, 2021) als de Europese Unie (zie o.a. Council of the European Union, 2020, en European Commission, 2021).
[10] Dat wil niet zeggen dat er niets over geschreven is. Zie bijvoorbeeld (Dignum 2017) en (Chesterman, 2021).
[11] Zie bijvoorbeeld Enarsson et al. (2022) en Dechense et al. 2019.
[12] Zie bijvoorbeeld *Politie in New York stopt met robothond na golf van kritiek, Nederlandse politie blijft apparaat inzetten*, Volkskrant 30-4-2021. Er zijn uiteraard bezwaren tegen de inzet van politiehonden, maar deze richten zich op bijvoorbeeld de proportionaliteit en het welzijn van de dieren, niet op de autonomie van de hond zelf.

we hoe het gebruik van (autonome) AI-systemen het vertrouwen van burgers in de politie kan beïnvloeden, en hoe transparantie over AI gebruik van de politie een centrale rol kan spelen in het vergroten van dit burgervertrouwen.

Autonome AI-systemen bij de Nederlandse politie

Het maken van keuzes kost mentale arbeid. Hiermee bedoelen we niet alleen besluiten in werkprocessen, zoals of iemand gearresteerd moet worden, maar ook allerlei kleine keuzes, zoals of een foto op een gegevensdrager relevant is voor de politie. De data op bijvoorbeeld in beslag genomen servers of telefoons neemt toe en de politie moet voor de beschikbare datapunten keuzes maken over of ze wel of niet bijdragen aan een onderzoek. De hoeveelheid keuzes groeit mee met de data en het aantal mensen bij de politie groeit niet evenredig mee. De realiteit is daarom dat bepaalde keuzes niet (meer) gemaakt worden (door bijvoorbeeld bepaalde databronnen niet mee te nemen) en dat de 'keuzeproductiviteit' per medewerker moet groeien. Het automatiseren van keuzes is een belangrijk middel om de productiviteit te verhogen.[13] Een toepassing die een keuze automatiseert zal sowieso in meer of mindere mate "intelligent" moeten zijn. Artificiële Intelligentie (AI) wordt vandaag de dag gebruikt om een breed scala aan algoritmen aan te duiden. De Europese Commissie definieert AI als volgt: *"Systemen die intelligent gedrag vertonen door hun omgeving te analyseren en – met een zekere mate van zelfstandigheid – actie te ondernemen om specifieke doelstellingen te bereiken"*.[14] Deze definitie roept wellicht beelden op van geavanceerde, autonoom acterende 'robots' die algemeen intelligent gedrag vertonen en direct de wereld om ons heen beïnvloeden. Vanuit onze ervaring in het Nationaal Politielab AI zien we dat AI-systemen vaak simpelere (combinaties van) algoritmen welke ter ondersteuning van menselijke beslissers kleinere deeltaken uitvoeren, zoals bijvoorbeeld objecten herkennen op foto's, documenten classificeren en informatie verstrekken over een specifiek onderwerp.

De bij het grote publiek meest bekende AI-toepassing van de politie is wellicht het *predictive policing* systeem CAS, dat gegeven informatie uit politiesystemen en bepaalde sociaaleconomische indicatoren het risico op criminaliteit in bepaalde gebieden voorspelt (Waardenburg 2021, hfstk 3). Bij de Nederlandse Politie wordt echter aan veel meer AI gewerkt ter ondersteuning van een groot aantal werkprocessen. Zo is er AI voor spraakherkenning, welke automatisch verhoren kan transcriberen.[15] Ook is er een groot aantal classificatiealgoritmen. Deze algoritmen nemen een bepaalde input, bijvoorbeeld een chatbericht, website of een foto, en delen deze in een categorie in. Zo kunnen bijvoorbeeld chatberichten worden geclassificeerd als doodsbedreiging of niet, websites van webshops als malafide of bonafide (Odekerken en Bex, 2020) en foto's van autobestuurder naargelang ze wel of geen telefoon in de hand hebben.[16] Verder zijn er AI-chatbots en -adviessystemen, welke bijvoorbeeld in geval van een online aangifte verdere informatie aan de burger kunnen vragen om zo een gedetailleerd individueel advies te kunnen geven (Testerink et al, 2019, Odekerken et al, 2020).

De bovengenoemde systemen maken gebruik van verschillende AI-technieken, zoals datagedreven *machine learning*, zoek- en optimalisatietechnieken en logica- en kennisgebaseerde technieken. Sommige systemen maken gebruik van meerdere technieken:

[13] Naast productiviteit kan automatisering ook bevorderend zijn voor onder anderen consistentie en snelheid in een werkproces. Verder zijn sommige taken ook simpelweg niet uitvoerbaar voor mensen.
[14] Europese Commissie Doc. 15641/18. Gecoördineerd plan inzake kunstmatige intelligentie.
[15] Zie: https://www.innoveermeemetjenv.nl/documenten/videos/2021/03/09/ai_experiment_nationale_politie
[16] Zie: https://www.politie.nl/nieuws/2021/juli/1/00-monocam-ingezet-tegen-afleiding-in-verkeer.html

bepaalde versies van het aangifte-adviessysteem voor online handelsfraude gebruiken bijvoorbeeld *machine learning* om uit de tekstvelden van de aangifte basisfeiten te extraheren, argumentatielogica om hier conclusies uit te trekken wat betreft de mogelijkheid dat het een geval van fraude betreft, en optimalisatietechnieken om alleen de strikt noodzakelijke extra vragen aan de aangever te stellen.

Alle genoemde systemen hebben een bepaalde mate van autonomie, maar deze is niet altijd van dezelfde aard. Om beter grip te krijgen op het concept van autonomie kunnen we deze opsplitsen in vier categorieën: bronautonomie, modelautonomie, keuzeautonomie en doelautonomie (Walsh, Mahesh, & Trumbach, 2018) (Figuur 1). Laten we deze bespreken aan de hand van het voorbeeld van een AI-toepassing die chatberichten met doodsbedreigingen uit cryptotelefoondata filtert[17].

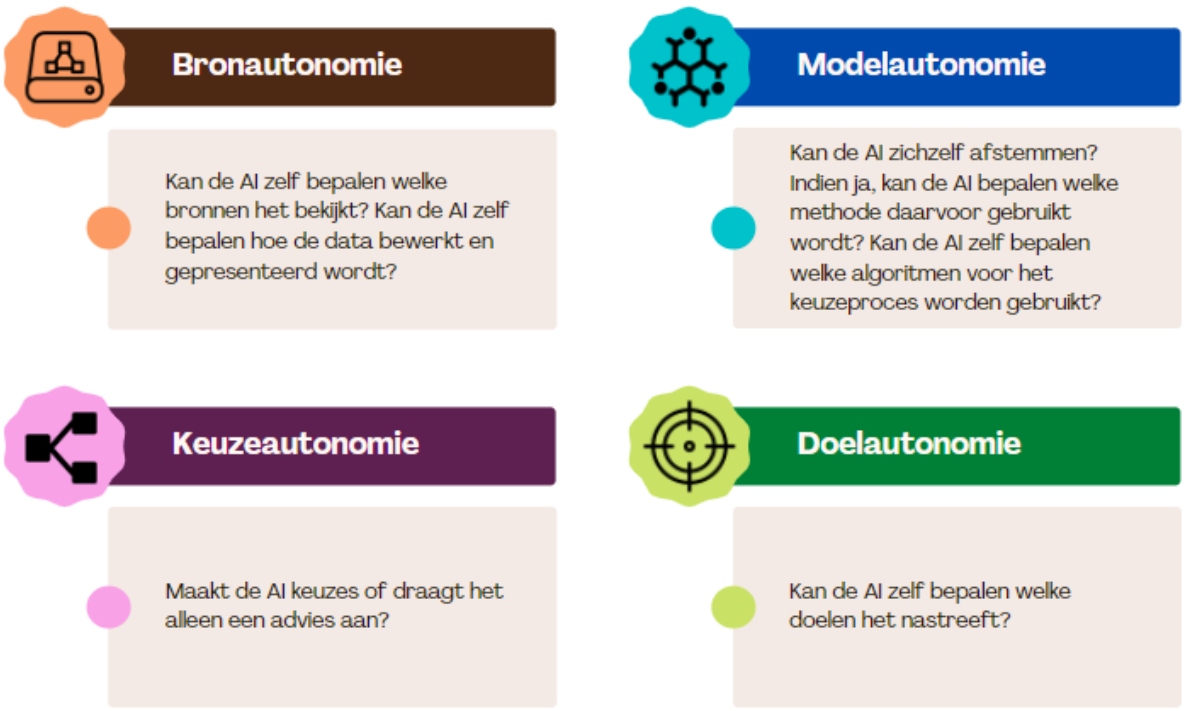
Bronautonomie betreft de mogelijkheid van de AI om zelf te kiezen welke databronnen het hanteert. Voor het AI voorbeeld zou dit betekenen dat de AI zelf kan bepalen welk (gedeelte van) bepaalde datasets met chatberichten het bekijkt, maar ook hoe bijvoorbeeld deze berichten voorbewerkt worden. Kan de AI bijvoorbeeld zelf ervoor kiezen om alle berichten eerst naar het Engels te vertalen? In de echte toepassing wordt door de mens bepaald welke dataset aan het algoritme wordt aangeboden, en hoe deze wordt voorbewerkt. Wel kan het algoritme er vervolgens voor kiezen om een deel van de berichten uit de inputdata te negeren, zoals bepaalde woorden, bij het bepalen of een bericht een doodsbedreiging is of niet.

Modelautonomie stelt een AI in staat om zelf te leren, om de methode hiervoor te kiezen en/of om te bepalen welke type algoritmen het gebruikt voor keuzes. Deze vorm van autonomie volgt altijd uit de toepassing van machine learning. Ook bij de filtering van chatberichten heeft de AI zichzelf geleerd hoe het doodsbedreigingen kan herkennen aan de hand van voorbeelddata. Daarbij hebben overigens mensen wel bepaald wélk algoritme de AI gebruikt (neurale netwerken) en de leermethode (*deep learning*).

Keuzeautonomie stelt een AI-toepassing in staat om een keuze ook uit te voeren. Dit is een vrij subtiele vorm van autonomie die vooral als een spectrum gezien moet worden. Bij het filteren van chatberichten kan men namelijk zeggen dat de AI niet acteert op doodsbedreigingen maar alleen relevante berichten aandraagt, en dus geen keuzeautonomie heeft. Echter, de keuze voor wat relevant is wordt wel degelijk door de AI autonoom uitgevoerd, en deze keuze beïnvloed in meerdere of mindere mate de mens die actie onderneemt.

Doelautonomie stelt de AI in staat om zelf te bepalen welke doelen het nastreeft. Dit komt in de praktijk niet voor bij de politie. In het geval van het berichtenfilter hebben wederom mensen bepaald dat het doel van de AI is om doodsbedreigingen uit de data te filteren. Een vorm van doelautonomie had kunnen zijn dat de AI zelf aan de hand van het gedrag van analisten had kunnen bepalen dat het aandragen van berichten van bepaalde personen nuttiger was geweest dan het aandragen van mogelijke doodsbedreigingen.

[17] De auteurs van dit artikel waren niet direct betrokken bij de ontwikkeling van dit systeem of de zaken waarin deze is toegepast. Wel hebben ze inzicht in hoe het ontwikkelproces verloopt van deze en soortgelijke tekstclassificatie modules. Voor meer informatie over dergelijke systemen zie https://www.forensischinstituut.nl/actueel/nieuws/2021/05/05/nfi-leert-computers-om-berichten-met-doodsbedreiging-uit-grote-hoeveelheden-data-te-filteren



Figuur 1. Categorieën van autonomie.

Technologische valkuilen en risico's

In deze sectie concentreren we ons in bijzonder op de valkuilen en risico's wanneer we keuzes toevertrouwen aan een AI-toepassing en geven we praktijkvoorbeelden van ontwikkelingen bij het Nationaal Politielab AI (NPAI) die daaraan gerelateerd zijn.[18]

Weloverwogen keuzes

De politie heeft strenge regels te volgen met betrekking tot het gebruik van data. Dit heeft als gevolg dat de bron- en doelautonomie van een systeem in de regel zeer beperkt zal zijn bij AI-toepassingen voor de politie – voor elke databron moet immers zorgvuldig door mensen bepaald worden of en voor welke specifieke doelen deze gebruikt mag worden. Een aandachtspunt hier is echter wel dat AI een zekere mate van bronautonomie heeft in het bepalen welk gedeelte van de aangeboden data het in beschouwing neemt. Zo kan het dus gebeuren dat een AI-toepassing bronautonomie uitoefent door bepaalde factoren te negeren die voor de politie wel relevant zijn. Neem als voorbeeld een triageproces over welke aangiftes handelsfraude als volgende worden opgepakt, waar de AI relevante aangiftes aandraagt op basis van, bijvoorbeeld, hoeveel geld er met de mogelijke fraude gemoeid is. Voor de politie is de kwetsbaarheid van het slachtoffer ook een belangrijke factor, maar deze kan overkomen op het leeralgoritme als negeerbare ruis in de data, zonder dat het algoritme herkent dat het een wezenlijk onderdeel is van de besluitvorming.

Het Nationaal Politielab AI richt zich onder andere op het ontwikkelen van *transparante* AI-toepassingen. Bij zulke toepassingen kan een datawetenschapper of eindgebruiker meer inzicht krijgen in het keuzeproces van een AI toepassing. Zo kan bijvoorbeeld gekeken worden of alle relevante factoren voor een keuze ook daadwerkelijk door een AI-toepassing meegewogen

[18] Het Nationaal Politielab AI is een samenwerkingsverband tussen de Landelijke Eenheid van de Politie en verschillende Nederlandse universiteiten – zie https://icai.ai/police-lab-ai/.

worden. Uiteraard hoeft niet voor elke keuze door een AI-toepassing een specifieke uitleg opgevraagd te worden. Men kan bijvoorbeeld na de ontwikkelfase controleren of de AI toepassing in het algemeen relevante soorten informatie negeert. Ook is het mogelijk om alleen de uitleg bekijken wanneer besloten wordt te acteren op de output van het systeem.

Eén manier om de transparantie van AI-toepassingen te verhogen is de inzet van tools waarmee van nature minder transparante, *black-box* machine learning modules geanalyseerd kunnen worden. Een dergelijke tool, de Explabox, wordt door het NPAI ontwikkeld. [19] De Explabox maakt het mogelijk om te kijken door welke factoren de output van een machine learning algoritme beïnvloed wordt. Stel dat bijvoorbeeld de eerdergenoemde toepassing die chatberichten met doodsbedreigingen uit cryptotelefoondata filtert, anders beslist als een bericht in het Engels of het Nederlands is, terwijl dit voor de politie geen factor van betekenis is. Met de Explabox kan de invloed van zulke factoren in het algemeen en bij individuele beslissingen bepaald worden.[20]

Naast het transparant maken van black-box machine learning modules, valt onder transparantieverhogende technieken ook werk rondom AI-technieken die inherent niet als black-box beschouwd worden. Een voorbeeld hiervan zijn de regelgebaseerde argumentatielogica's (Testerink et al, 2019, Odekerken et al, 2020), welke onder andere worden ingezet in eerdergenoemde adviessystemen voor online aangiften. Dergelijke argumentatielogica's worden gebruikt om te bepalen of bepaalde basisobservaties uit een formulier een juridische conclusie onderbouwen. De regels die in deze logica gebruikt worden zijn direct afgeleid van het wetsartikel 326 wetboek van strafrecht over fraude. De juridische duiding op basis van argumentatielogica's kan dus stap-voor-stap gevolgd worden door een persoon.

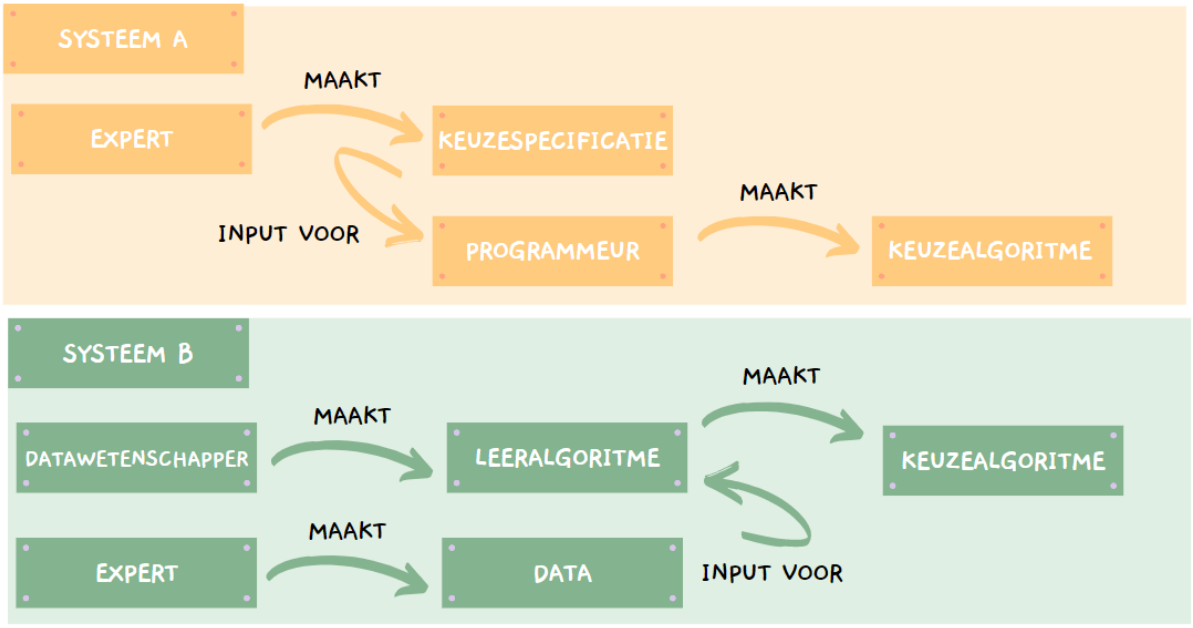
Verantwoordelijkheidsvacuüm

Een van de ongewenste effecten van autonome AI is dat er mogelijk een verantwoordelijkheidsvacuüm ontstaat (Matthias, 2004). Daarbij refereren we aan verantwoordelijkheid in de fundamentele zin en niet in de juridische zin. Laten we twee typen benaderingen voor het bepalen van risicoprofielen van, bijvoorbeeld, verdachten vergelijken (Figuur 2). Bij systeem A schrijft een expert op hoe combinaties van factoren relateren aan een risico-indicatie. Deze specificatie wordt 1-op-1 geprogrammeerd in systeem A, waarbij de enige vorm van autonomie keuzeautonomie is. Bij systeem B geven we een systeem een dataset met voorbeelden van combinaties van factoren en de in het verleden toegekende risico-indicatie. Systeem B krijgt bronautonomie in de zin dat Systeem B zelf mag kiezen of factoren wel of niet meegewogen worden, modelautonomie in de zin dat het zichzelf mag afstemmen (bijvoorbeeld hoe zwaar een factor meetelt) en keuzeautonomie omdat een triage proces direct gekoppeld wordt aan de risico-indicatie.

Bij Systeem A is de expert verantwoordelijk voor hoe een risico-indicatie wordt opgesteld en de programmeur voor de technische correctheid van het systeem, gegeven de input van de expert. Gezamenlijk dragen zij zo verantwoordelijkheid voor de risico-indicaties. Bij Systeem B is de expertise verscholen in de historische data. Hier is niet iemand aan te wijzen als verantwoordelijke hoe combinaties van factoren in het verleden hebben geleid tot een specifieke risico-indicatie. Ook wisten de experts uit het verleden niet dat hun keuzes later ter

[19] https://github.com/MarcelRobeer/explabox.
[20] Hiermee past de Explabox binnen het gebied van *Explainable AI*, waarin methoden om machine learning algoritmen te evalueren en uit te leggen ontwikkeld worden, zie Gunning et al. 2019.

input van Systeem B gebruikt zou worden. De datawetenschapper van Systeem B is verantwoordelijk voor een degelijke implementatie van de autonome technologie (zoals het leeralgoritme), maar ziet mogelijk de data zelf niet. Dit is niet heel ongebruikelijk omdat ontwikkelaars regelmatig maar beperkt operationele data mogen inzien. Wanneer de datawetenschapper haar werk gedaan heeft kan ze zien dat Systeem B op de data goed presteert, maar kan niet vanuit het operationele domein uitleggen waarom de keuzes van Systeem B operationeel gezien redelijk en gewenst zijn. De gebruikers van Systeem B zijn mogelijk zelf geen experts en experts die er wel zijn weten waarschijnlijk niet hoe Systeem B intern werkt. Het gevolg is een verantwoordelijkheidsvacuüm waar niemand verantwoordelijkheid kan nemen voor de keuzes van Systeem B. In fundamentele zin heeft Systeem B zelf de verplichting van deze verantwoordelijkheid, maar is ontheven hiervan als het niet in staat is gesteld om deze verantwoordelijkheid te nemen. Het vacuüm ontstaat in dit geval door modelautonomie. Een verantwoordelijkheidsvacuüm kan daarom een ongewenst effect van modelautonomie zijn. Ook doelautonomie kan hiertoe leiden, maar is bij ons weten niet iets dat speelt bij de politie.



Figuur 2. Totstandkoming van het keuzealgoritme in systeem A en B.

In de afgelopen jaren heeft het NPAI onderzoek gedaan naar AI-toepassingen die model- en keuzeautonomie hebben, maar beperkt worden in hun gedragingen, om zo basisprincipes voor de politie te waarborgen. Een voorbeeldtoepassing met een dergelijk ontwerp is wederom het online-adviessysteem dat aangevers helpt bij het doen van aangiftes van internetoplichting (Testerink et al, 2019, Odekerken et al, 2020). Deze AI toepassing kan autonoom informatie inwinnen (keuzeautonomie). Een principe dat daarbij komt kijken is dat politietoepassingen alleen informatie zouden moeten inwinnen die relevant zijn voor de taak voorhanden. Het redeneermechanisme onderliggend aan de argumentatielogica zorgt ervoor dat de AI alleen maar informatie kan inwinnen wanneer expliciet beargumenteerd kan worden dat deze informatie nodig is gegeven de wetteksten rondom online handelsfraude en de al afgeven informatie van de aangever. Het ontwerp achter de keuzehulp staat ook toe om te optimaliseren middels modelautonomie op bijvoorbeeld het gebruik van open data versus gesloten data, niet persoonsgebonden data versus persoonsgegevens, etc. Hierbij kan de introductie van een dergelijke AI-toepassing juist privacy verhogend werken ten opzichte van de huidige manier van werken. De beperkingen van gedragingen mitigeren het risico dat relevante factoren

onvoldoende worden meegewogen en verkleinen het verantwoordelijkheidsvacuüm omdat het systeem niet onverwachte keuzes kan maken.

Technologische limieten m.b.t. redeneren over normen en waarden

De kernmissie van de politie is om waakzaam en dienstbaar te zijn aan de waarden van de rechtsstaat. Wanneer de politie een deel van deze keuzes overhevelt naar autonome AI-toepassingen, dan moet er ook zorg zijn voor het behoud van ethische principes. Op het integriteitsvlak is het belangrijk dat politiemensen normen en waarden meenemen in hun keuzes en vanuit zichzelf "het juiste" willen doen (Gehem et al. 2021). Wanneer we autonomie toekennen aan een AI-toepassing is er geen alom erkende techniek voor het redeneren met normen en waarden en bestaat er geen technische notie van "het juiste doen". Deze afwegingen moeten dus in het ontwerp van het systeem en de ingebruikname zorgvuldig door de mens meegegeven worden. In het kader van empathie is het gewenst dat de politie keuzes maakt die in lijn zijn met de samenleving. Het is op empathisch niveau gewenst dat de politie rekening houdt met daders, slachtoffers en omstanders en handelt daarnaar. Hoewel er in beperkte omstandigheden simulatie van empathie mogelijk is voor AI, zien wij dit in het algemeen geen beschikbare functionaliteit is bij de politie. Het NPAI adresseert deze valkuilen middels deelname aan consortia die dit thema vanuit meerdere disciplines aanvliegen. Voorbeelden hiervan zijn de aansluiting met het ALGOPOL-project[21] en ELSA-lab initiatieven[22]. Deze samenwerkingen bieden denkkaders, richtlijnen en methoden om normen en waarden mee te nemen in de socio-technische aspecten van AI-toepassingen.

Breder binnen de politie lopen ook verschillende andere ontwikkelingen op dit thema, zoals begeleidingsethiek in samenwerking met het Electronic Commerce Platform (ECP) (Borst, 2019 Begeleidingsethiek verkent ethische vragen op basis van de context en de relevante actoren, effecten en waarden (Verbeek & Tijink, 2019). Met behulp van deze methode kunnen gesprekken – genaamd ethiektafels - methodologisch gevoerd worden die leiden tot handelingsopties om ethische risico's te verkleinen en kansen te verzilveren. De handelingsopties worden opgedeeld in maatregelen in de omgeving van de techniek (by environment), maatregelen die we kunnen nemen om de techniek aan te passen (by design), en maatregelen die gebruikers kunnen nemen (by user). De kwaliteit van een ethiektafel is afhankelijk van de deelnemers. In het algemeen is het goed om zoveel mogelijk perspectieven en disciplines aan tafel te krijgen.

Vertrouwen

Achter de bovenstaande risico's schuilt een groter probleem, namelijk wantrouwen vanuit de samenleving jegens AI gebruik door de politie.[23] Hoewel niet specifiek benoemd, zit hier vaak een aspect van autonomie in. Waar angst voor is, is dat we een deel van de menselijke autonomie bij de politie opgeven en dat deze vervangen wordt door autonomie van AI.[24] Ook bestaat het gevaar van zogenaamde *automation bias*,[25] waar zelfs met een *human-in-the-loop*

[21] https://algopol.sites.uu.nl/.
[22] Zie https://www.nwo.nl/nieuws/meer-dan-10-miljoen-euro-voor-mensgerichte-ai-onderzoek-elsa-labs. In het bijzonder het AI-MAPS initiatief.
[23] Nederlanders wantrouwen de algoritmes die het leven in toenemende mate bepalen, Trouw, 6-6-2019, https://www.trouw.nl/nieuws/nederlanders-wantrouwen-de-algoritmes-die-het-leven-in-toenemende-mate-bepalen~bf196475/
[24] De Kool et al. (2020).
[25] Skitka et al. (2012).

politiemedewerkers te veel beïnvloed worden door het systeem, waardoor AI feitelijk zelf beslissingen kan nemen. In hoeverre deze angsten rationeel (or irrationeel) zijn, maakt niet uit. Ze hebben namelijk invloed op het vertrouwen van burgers in het gebruik van AI door de politie.

Dit vertrouwen van burgers in AI wordt op twee manieren beïnvloed: *generiek* in het algemene functioneren van de politie met betrekking tot AI en meer *specifiek*, in bepaalde toepassingen. Een voorbeeld waardoor het generieke vertrouwen aangetast kan worden is het rapport van de Rekenkamer, waarin verschillende algoritmes getoetst worden aan de hand van een uitgebreid toetsingskader.[26] Ook een algoritme van de politie wordt getoetst. De Rekenkamer concludeert dat dit algoritme op alle getoetste onderdelen ondermaats scoort en dat er voor de politie veel werk aan de winkel is om te zorgen dat ze algoritmes op een eerlijke en verstandige manier kunnen gebruiken. Dit rapport bekritiseert de huidige gang van zaken en kan daarmee zorgen voor maatschappelijke onrust en wantrouwen rondom het gebruik van algoritmen door de politie.

Een voorbeeld waarin het vertrouwen in een specifieke technologie ter discussie staat, is het rapport van Amnesty International over het Sensing-Project in Roermond.[27] In dit rapport wordt beschreven hoe discriminatie kan ontstaan door hogere risicoscores voor mogelijk mobiel banditisme toe te kennen aan mensen met een Oost-Europese kentekenplaat dan zonder. Hiermee wordt specifiek AI bekritiseerd, wat kan leiden tot argwaan en weerstand van burgers tegen AI gebruik door de politie. Kortom, een toename in het aantal negatieve publicaties rondom AI gebruik door de politie, zowel generiek als specifiek, kan leiden tot een groeiend wantrouwen van burgers. Als het vertrouwen in AI gebruik door de politie laag is, zal het lastig zijn om autonome systemen te laten landen in de samenleving omdat er weinig draagvlak zal zijn voor deze technologieën.

De politie moet echter niet het kind met het badwater weggooien. Ten eerste omdat enige mate van wantrouwen gezond is in een democratische samenleving. Een voorzichtige houding zorgt ervoor dat burgers politieke en sociale gebeurtenissen in hun gemeenschap nauwlettend in de gaten houden (Lenard, 2008). Als er geen gezond wantrouwen bestaat, is het onmogelijk om misstanden te signaleren en aan de kaak te stellen. Ten tweede omdat er manieren zijn om irrationele angsten en uitingen van wantrouwen weg te nemen en daarmee vertrouwen te waarborgen.

Eén van die manieren om vertrouwen te waarborgen is door het realiseren van transparantie rondom het gebruik van AI door de politie. Maar hoe doe je dat, transparant zijn over AI gebruik van de politie? Bij de beantwoording van deze vraag kan de volgende indeling houvast bieden om na te denken over welke zaken men transparant kan zijn:

- Microniveau: informatie over het algoritme zelf en/of de gebruiker(s)
- Mesoniveau: informatie over de organisationele inbedding van algoritmen
- Macroniveau: informatie over de institutionele inbedding van algoritmen in wet- en regelgeving

Transparantie op *microniveau* – informatie over hoe een bepaalde AI-toepassing werkt en hoe het tot een bepaalde uitkomst gekomen is - werd eerder al besproken als manier om de keuzes

[26] Algemene Rekenkamer (2022).
[27] Amnesty International (2020).

van een AI toepassing inzichtelijk te maken. Men kan hierbij denken aan informatie over een AI-toepassing zoals gegenereerd door de Explabox, maar ook een algemene uitleg die in samenwerking met politiemedewerkers geschreven en vooraf gecontroleerd is. Eerder onderzoek laat zien dat transparantie over een specifiek algoritme van groot belang is.[28] In dit onderzoek is gekeken naar het eerdergenoemde adviessysteem voor aangiften, waar AI gebruikt wordt om sneller vast te stellen of er sprake is van een strafbaar feit in het geval van online handelsfraude. Hier krijgt de aangever direct een advies van het systeem, waardoor burgers eerder duidelijkheid hebben en onterechte aangiftes voorkomen kunnen worden. Als er bij de aanbeveling geen uitleg volgt over hoe dit advies tot stand is gekomen of hoe de slimme keuzehulp werkt, dan zien we grote afname in het vertrouwen van burgers in dit algoritmische advies. Transparantie heeft in dit geval dus een groot positief effect op vertrouwen en onderstreept het belang van informatie verschaffen over specifieke algoritmen.

Op *mesoniveau* betreft transparantie de informatie over de organisationele inbedding van algoritmen, bijvoorbeeld welke experts er betrokken zijn bij de totstandkoming van algoritmen, het organisatiebeleid rondom algoritmen of de monitoring en evaluatie van algoritmen (Grimmelikhuijsen & Meijer, 2020). Het gedetailleerde etnografische onderzoek bij de politie door Waardenburg (2021) maakt transparant hoe AI het politiewerk en de organisatie verandert. Door onderzoekers een langere tijd mee te laten lopen in de politieorganisatie creëert de politie de mogelijkheid om inzicht te krijgen in de organisationele inbedding van hun algoritmen. Een ander voorbeeld van transparantie op *mesoniveau* betreft de ontwikkeling van een algoritmeregister voor de politie. Op dit moment wordt er binnen de politie gewerkt aan het opzetten van een algoritmeregister voor de hele organisatie. Hierin kunnen verschillende onderwerpen uitgelicht worden, zoals wie er betrokken zijn bij de ontwikkeling van een algoritme, binnen welke risicoclassificatie een algoritme valt, wie of welke afdeling verantwoordelijk is voor een algoritme of hoe een algoritme gemonitord en/of geëvalueerd wordt. Dit geeft burgers en andere betrokken (maatschappelijke) partijen inzicht in hoe algoritmen ingebed zijn in de politieorganisatie. Naar het algoritmeregister wordt tegelijkertijd ook wetenschappelijk onderzoek gedaan door ALGOPOL en het NPAI, zodat de funderingen van dit project gebaseerd worden op lessen die we trekken uit recente onderzoeken.

Op *macroniveau*, tot slot, kan informatie worden gegeven over hoe algoritmen zijn ingebed in wet- en regelgeving. Zo vallen algoritmen van de politie in Nederland onder de Algemene verordening gegevensbescherming (AVG) en wanneer het om persoonsgegevens gaat ter voorkoming, opsporing of vervolging van strafbare feiten onder de Wet politiegegevens (Wpg). Een voorbeeld op dit niveau betreft het eerdergenoemde rapport van de Rekenkamer, waarbij onder andere is getoetst of algoritmen van verschillende organisaties, waaronder ook de politie, voldoen aan de eisen van de AVG. Dit rapport heeft maatschappelijke en politieke discussies teweeggebracht, waardoor zowel binnen als buiten de politieorganisatie gereflecteerd en gesproken kan worden over de institutionele inbedding van algoritmen.

Kortom, door transparant te zijn op elk niveau wordt informatie beschikbaar gemaakt die de samenleving tot zich kan nemen en het hier over kan hebben. Transparantie zorgt er namelijk voor dat we een gesprek kunnen voeren over belangrijke onderwerpen zoals welke waarden we belangrijk vinden in de totstandkoming, het functioneren en het gebruik van algoritmen of welke hoeveelheid autonomie binnen AI-toepassingen van de politie we als samenleving aanvaardbaar vinden. Transparantie maakt een breder debat mogelijk. Als we het niet weten, kunnen we het er ook niet over hebben. Het is een belangrijke stap in het wegnemen van

[28] https://www.uu.nl/en/news/research-contributes-to-crime-reporting-tool-national-police

wantrouwen en kan uiteindelijk zorgen voor meer draagvlak en vertrouwen in autonomie in AI door de politie.

Conclusie

In dit artikel hebben we in de context van de Nederlandse politie besproken hoe autonomie in AI kan voorkomen. Daarna hebben we de (technische) valkuilen van autonomie bij AI besproken:

1. Autonome AI kan keuzes op grond van de verkeerde of irrelevante factoren maken.
2. Autonome AI kan leiden tot een verantwoordelijkheidsvacuüm.
3. Autonome AI kan niet met normen en waarden redeneren.

Voor elk van deze valkuilen hebben we een voorbeeld gegeven van hoe de politie hiermee omgaat. Wat betreft keuzes (1) kan transparant gemaakt worden waarom een AI-systeem of algoritme keuzes maakt, zowel in het algemeen als bij individuele beslissingen of keuzes. Een verantwoordelijkheidsvacuüm (2) kan worden verkleind door AI-toepassingen zo te ontwerpen dat ze zich aan de basisprincipes van de politie en de wet houden *by design*. En wat betreft normen en waarden (3) wordt in de ontwikkeling van AI-toepassingen begeleidingsethiek toegepast, alsmede bredere samenwerking gezocht met verschillende disciplines uit de sociale en geesteswetenschappen.

De drie genoemde valkuilen kunnen onder andere leiden tot een verminderd vertrouwen van burgers in de politie. De politie ligt terecht onder een vergrootglas als het gaat om de inzet van AI en andere nieuwe technologieën. Wij betogen dat transparantie een belangrijke manier is om het vertrouwen van de maatschappij in de politie te waarborgen. Deze transparantie moet dan wel op alle niveaus, van (technisch) microniveau tot (institutioneel) macroniveau gerealiseerd worden.

De valkuilen en mogelijke oplossingen genoemd in dit artikel zijn zeker niet uitputtend, en de politie blijft samen met haar partners bezig (het gebruik van) AI voor en door de politie te verbeteren. Alleen door samen te praten over deze onderwerpen kunnen we bepalen wat ongewenste effecten zijn en welke handelingsopties we hebben om deze effecten te voorkomen of te verminderen. In dit artikel is een startpunt gepresenteerd voor deze discussie.

De titel van dit artikel vraagt wat het ertoe doet dat je een mens bent. Het korte antwoord is dat alleen mensen kunnen aanvoelen wat voor een soort keuzes de politie dagelijks moet maken en hoe dit moet gebeuren. Dat bepaalt uiteindelijk welke autonome AI-toepassingen voor de politie gewenst zijn.

Bibliografie

Algemene Rekenkamer (2022) Algoritmes getoetst: De inzet van 9 algoritmes bij de overheid. https://www.rekenkamer.nl/publicaties/rapporten/2022/05/18/algoritmes-getoetst.

Amnesty International (2020) We Sense Trouble: Automated Discrimination and Mass Surveillance in Predictive Policing in the Netherlands. https://www.amnesty.nl/content/uploads/2020/09/Amnesty-International_We-Sense-Trouble_EUR-35_2971_2020.pdf?x64788.

Borst, E. (2019) Begeleidingsethiek bij de politie. *Politie*

Chesterman, S. (2020). Artificial intelligence and the problem of autonomy. *Notre Dame Journal on Emerging Technologies*, 1, 210-250.

Commissie Schneiders. (2022). *Ruimte voor slagvaardig politiewerk.* https://www.rijksoverheid.nl/documenten/kamerstukken/2022/06/30/tk-bijlage-eindadvies-adviescommissie-voor-de-landelijke-eenheid.

Council of the European Union (2020) *Presidency conclusions - The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change*, 11481/20, 2020. https://www.consilium.europa.eu/media/46496/st11481-en20.pdf.

Dechesne, F.; Dignum, V.; Zardiashvili, L.; Bieger, J. (2019) AI & Ethics at the Police: Towards Responsible use of Artificial Intelligence in the Dutch Police. *External research report*. https://hdl.handle.net/1887/85954.

Dignum, V. (2017). Responsible autonomy. *arXiv preprint arXiv:1706.02513*.

Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123-153.

European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Gehem, M., Blink, S. & Verbraak, M. (2021) Waardenkaart voor goed politiewerk. De Argumentenfabriek

Grimmelikhuijsen, S. G., & Meijer, A. J. (2020). Verantwoorde algoritmisering: zorgen waardengevoeligheid en transparantie voor meer vertrouwen in algoritmische besluitvorming?. Bestuurskunde, 2020(4), 7-20.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

de Kool, D., Vermeeren, B., & Steijn, B. (2020). Kunstmatige Intelligentie bij de politie: Praktische en sociale lessen ten aanzien van het aangifteproces. *Risbo Erasmus Universiteit Rotterdam*.

Lenard, P. T. (2008). Trust your compatriots, but count your change: The roles of trust, mistrust and distrust in democracy. Political Studies, 56(2), 312-332.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 175-183.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35.

Ministerie van Justitie en Veiligheid (2021) Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses. https://www.rijksoverheid.nl/documenten/richtlijnen/2021/09/24/richtlijnen-voor-het-toepassen-van-algoritmen-door-overheden-en-publieksvoorlichting-over-data-analyses.

Odekerken, D. & Bex, F (2020) Towards transparent human-in-the-loop classification of fraudulent web shops. Legal Knowledge and Information Systems. JURIX 2020: The Thirty

Third Annual Conference. Frontiers in Artificial Intelligence and Applications 334, pp. 239 - 242.

Odekerken, D., Borg, A. & Bex, F. (2020) Estimating Stability for Efficient Argument-based Inquiry. Computational Models of Argument. Proceedings of COMMA 2020, Frontiers in Artificial Intelligence and Applications 326, pp. 307-318.

Osoba, O. A., & Welser IV, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701-717.

Testerink, B., Odekerken, D.& Bex, F.(2019) A Method for Efficient Argument-based Inquiry, 13th International Conference on Flexible Query Answering Systems (FQAS 2019). Lecture Notes in Artificial Intelligence, volume 11529, p. 114-125, Springer.

Verbeek, P., & Tijink, D. (2019). Aanpak begeleidingsethiek: een dialoog over technologie met handelingsperspectief.

Waardenburg, L. (2021). Behind the scenes of artificial intelligence: Studying how organizations cope with machine learning in practice.

Walsh, K. R., Mahesh, S., & Trumbach, C. C. (2018). Autonomy in AI Systems: Rationalizing the Fears.