Gemeente Amsterdam

Versie 1.0
27 februari 2019

Algoritmes, accountability en duurzame toegankelijkheid

Een verkenning naar de toepassingen binnen gemeentes en uitdagingen met betrekking tot verantwoording en informatiebeheer

Inhoud

1 Inleiding

De interesse in algoritmes voor toepassingen die helpen om ons werk voor de stad nog beter te doen, neemt snel toe. Veel gemeentes en ook de gemeente Amsterdam experimenteren met deze technieken, in de hoop dat deze onze dienstverlening effectiever en efficiënter kunnen maken. Tegelijkertijd zien we de signalen van organisaties die zich zorgen maken over of we wel grip kunnen houden op deze technieken. Zo heeft de Raad van State onlangs in een kritisch rapport gewaarschuwd voor het uithollen van onze burgerrechten als gevolg van verregaande digitalisering, waaronder automatisering van besluitvorming op basis van technieken als zelflerende algoritmes[1]. Minister voor Rechtsbescherming Sander Dekker beloofde na kamervragen om onderzoek te doen naar de transparantie van zelflerende algoritmes[2].

Het Stadsarchief heeft als één van haar taken om te zorgen dat de besluitvorming binnen de gemeente duurzaam toegankelijke gemaakt wordt, zodat terug herleid kan worden op basis waarvan de gemeente haar besluiten heeft genomen. Gemeentelijke informatie moet duurzaam toegankelijk worden beheerd, zodat besluiten reconstrueerbaar zijn. Zo kan de gemeente blijvend transparant zijn over haar handelen.

Het Stadsarchief stelt de kaders voor het informatiebeheer in de stad en houdt hier toezicht op. Vanwege die rol zijn de experimenten met algoritmes van belang voor het Stadsarchief. Immers, wanneer onze besluiten worden genomen op basis van dergelijke algoritmes, hoe leg je daar verantwoording over af en hoe borg je dat je beslissingen neemt die corresponderen met je beleid? En hoe ga je dan om met het bewaren en vernietigen van informatie en data, zodat je later de juiste informatie kan terugvinden?

Ik heb gepoogd een antwoord te geven op deze vragen, door enkele cases te onderzoeken. De vraag is in de eerste plaats op welke manieren we deze technieken op dit moment gebruiken. Vervolgens is de vraag waar de knelpunten met betrekking tot het beheren en bewaren van deze algoritmes zitten.

Allereerst ga ik iets dieper in op wat we bedoelen met "slimme" of "zelflerende" algoritmes en hoe ze op dit moment worden toegepast binnen gemeenteland. Vervolgens beschrijf ik aan de hand van enkele onderzochte cases de processen waarin deze technieken worden gebruikt. Als laatste volgt een conclusie met enkele aanbevelingen over de techniek en over het inrichten van de processen.

[1] Raad van State (2018). Ongevraagd advies over de effecten van de digitalisering voor de rechtsstatelijke verhoudingen [W04.18.0230/I]. Online via: https://www.raadvanstate.nl/adviezen/zoeken-in-adviezen/tekst-advies.html?id=13065
[2] Sander Dekker (2018, 9 oktober). Transparantie van algoritmes in gebruik bij de overheid [kamerbrief]. Online via: https://www.digitaleoverheid.nl/document/kamerbrief-over-motie-over-transparantie-van-algoritmes-in-gebruik-bij-de-overheid/

2 Wat zijn algoritmes?

We leven in een tijd van razendsnelle technologische ontwikkelingen, die volgens het Rathenau Instituut[3] leidt tot een samenleving waarin de digitale wereld onlosmakelijk verbonden raakt met de fysieke wereld. Hierin leunt de mens op computers voor veel belangrijke beslissingen, en laat deze vaak zelfs geheel over aan computers. Deze tijd van digitalisering en automatisering drijft op Big Data, Internet of Things (IoT) en Kunstmatige Intelligentie (KI)[4]. De onderliggende technologie? "Slimme" of "zelflerende" algoritmes. Waar voorheen machines en computers "domme" menselijk taken overnamen, beloven "slimme", "zelflerende" algoritmes nu ook denk- en besliswerk over te kunnen nemen.

Big Data kenmerkt zich door grote hoeveelheden data uit verschillende bronnen die *real-time* geraadpleegd en aangevuld kunnen worden[5]. Analyse van Big Data is niet langer hypothese gedreven, maar data gedreven. Dit betekent dat niet langer vooraf hypotheses worden opgesteld die getoetst worden, maar dat het vinden van patronen en verbanden in datasets een doel *an sich* is. Dit kan leiden tot verrassende verbanden en inzichten, waarbij de causaliteit, oftewel de oorzaak van het verband, niet per se belangrijker is dan de correlatie zelf.

Een bekend voorbeeld is het Criminaliteits Anticipatie Systeem (CAS), waarbij gebieden op basis van een flinke hoeveelheid gegevens zoals aanwezigheid van horeca, woonplaats van bekende criminelen, uitvalswegen, demografische en sociaal-economische gegevens een risicoprofiel krijgen, op basis waarvan over politie-inzet kan worden besloten.

IoT omvat het geheel van met het internet verbonden fysieke, alledaagse objecten, oftewel 'dingen'[6]. Deze 'dingen' bevatten sensoren om data te verzamelen, soms mogelijkheden om deze data te verwerken en een actie te koppelen aan de metingen, en een verbinding met het internet, zodat de metingen ter opslag en analyse verstuurd kunnen worden naar een server die al dan niet in de cloud ligt.

[3] Kool, L. e.a. (2017). Opwaarderen – borgen van publieke waarden in de digitale samenleving. Online via: https://www.rathenau.nl/nl/digitale-samenleving/opwaarderen
[4] Gerards, J.; Nehmelman, R.; Vetzo, M. (2018). Algoritmes en Grondrechten. Universiteit Utrecht. Online via: https://www.rijksoverheid.nl/documenten/rapporten/2018/03/01/algoritmes-en-grondrechten
[5] WRR (2016). Big Data in een vrije en veilige samenleving. Online via: https://www.wrr.nl/onderwerpen/big-data-privacy-en-veiligheid/documenten/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving.
[6] Gerards, J.; Nehmelman, R.; Vetzo, M. (2018). Algoritmes en Grondrechten. Universiteit Utrecht. Online via: https://www.rijksoverheid.nl/documenten/rapporten/2018/03/01/algoritmes-en-grondrechten

IoT-toepassingen kennen we bijvoorbeeld in het asset management van onze bruggen, kademuren, lantarenpalen, vuilnisbakken en andere objecten in de publieke ruimte die eigendom zijn van de gemeente. Gemeente Amsterdam bijvoorbeeld koppelt sensoren aan vuilnisbakken, waarmee gemeten kan worden welke vuilnisbak vol zit, op basis waarvan de route van de vuilnisophaaldienst kan worden geoptimaliseerd (met behulp van algoritmes). Een ander voorbeeld zijn straatlantaarns die automatisch aan of uit gaan op basis van meetgegevens over de hoeveelheid daglicht, de weersomstandigheden en de verkeersdrukte (denk aan geluid als proxy voor de hoeveelheid verkeer, bijvoorbeeld). Deze voorbeelden laten zien dat je vaak ook opslagcapaciteit, analysesoftware en toegang in de vorm van een applicatie nodig hebt om gebruik te maken van de meetgegevens. Ook in het gezondheidsdomein zien we toepassingen in de vorm van apparaten die hartslag, lichaamstemperatuur, ademhaling meten en het aantal stappen tellen. Deze informatie kan geanalyseerd worden en helpen bij diagnostiek.

KI, tot slot, gaat grofweg om het nabootsen van menselijk denken en handelen. Kenmerkend is beslissingen nemen, problemen oplossen en leren. Hierin zit een hoge mate van autonomie. Met KI kunnen complexe taken verricht worden zonder menselijk ingrijpen. *Machine Learning* (ML) is een vorm van KI die veel toegepast wordt op data. Machine Learning omvat algoritmes die in staat zijn om op basis van 'trainingsdata' zelf conclusies te trekken over een volgende, nog onbekende casus.

Een toepassing waar ML uitermate geschikt voor is, is beeldherkenning. Op basis van duizenden foto's van bijvoorbeeld stoplichten, kan het 'zelflerende' algoritme zelfstandig herkennen of een nieuwe foto een stoplicht bevat. Een bekend voorbeeld is de scanauto die onder andere de gemeentes Amsterdam, Rotterdam en Utrecht inzetten om foutparkeren tegen te gaan. Er wordt een foto gemaakt van het kenteken, met behulp van ML wordt in de foto het kenteken herkend en wordt vervolgens op de achtergrond (met behulp van andere algoritmes) gecontroleerd of ze wel staan aangemeld voor betaald parkeren. Overigens moet opgemerkt worden dat niet automatisch een boete wordt uitgeschreven. Een handhaver rijdt achter de scanauto aan om de boete daadwerkelijk uit te schrijven. Naast juridische overwegingen is een voordeel hierbij dat de handhaver de context kan meenemen in zijn besluit om een boete uit te schrijven, bijvoorbeeld wanneer iemand aan het laden en lossen is.

Bij Machine Learning komt het ondoorzichtige 'black-box' aspect naar boven waar men vaak aan refereert als men het heeft over algoritmes[7]. Het algoritme bouwt als het ware een netwerk van op elkaar volgende afwegingspunten, waarbij aan iedere afweging een waarde en een gewicht wordt gehangen[8]. Het algoritme "leert" wanneer het feedback krijgt op de output die het levert en op basis van die feedback de waardes en gewichten verandert. Hierdoor wordt het algoritme steeds preciezer. Maar voor een individueel geval is het pad langs al die (sets van) variabelen voor ons mensen praktisch betekenisloos en in elk geval losgezongen van een redelijke, begrijpelijke uitleg (als je het pad überhaupt kunt achterhalen). Dan krijg je dus op de vraag "op basis waarvan heb je

[7] Zie bijvoorbeeld: Wouter van Bergen in de Telegraaf (25 oktober 2018). "Algoritmes vaak een black box". Online via: https://www.telegraaf.nl/financieel/2727422/algoritmes-vaak-een-black-box.
[8] Dit noemt men ook wel een "neural network". Zie voor een kraakheldere, uitgebreide introductie: https://youtu.be/aircAruvnKk

deze beslissing genomen?" een antwoord als: "Tja, het algoritme: een schier eindeloze reeks afwegingen langs honderden variabelen met veranderlijke waardes en gewichten"[9].

2.1 conclusie

In gemeenteland zijn er veel toepassingen waarvoor bovengenoemde technologieën met behulp van al dan niet "zelflerende" algoritmes gebruikt worden. In de verschillende voorbeelden is soms het Big Data aspect wat groter aanwezig, soms wordt de data verzameld via IoT-objecten, en de analysemethodes variëren van klassieke hypothese-gestuurde statistiek, tot *Machine Learning* technieken. Wanneer we het dus hebben over algoritmes, is het belangrijk om heel scherp te houden welke aspecten van de techniek per toepassing belangrijk zijn voor het informatiebeheer. Is het de verzameling en opslag van grote hoeveelheden data? Is het het *real-time* aspect van de dataverzameling? Is het de verzameling uit verschillende bronnen? In hoeverre is sprake van een ondoorzichtige techniek?

In hoeverre het van belang is om op deze technische aspecten beheermaatregelen toe te passen, is onder meer afhankelijk van de mate van impact die met een algoritmisch besluit uitgeoefend wordt op de persoonlijke levenssfeer van individuele burgers. In het voorbeeld van de scanauto's wordt een boete uitgeschreven, hetgeen meer impact heeft op een individu. Dit verzwaart het belang van goede verantwoording en informatiebeheer. Tegelijk zien we in dit voorbeeld dat de belangrijkste beslissing, namelijk om een boete uit te schrijven, nog bij de ambtenaar ligt, en niet bij de machine.

Kortom, je kunt in je proces dingen doen om de noodzaak van verantwoording over de techniek te verminderen. Verder kun je naast technische informatie, ook allerlei procesinformatie bewaren om zodoende het informatiebeheer en de verantwoording te borgen. Om deze uitspraak wat nader toe te lichten worden hieronder twee casussen beschreven als voorbeeld. Hierbij kijk ik naar de techniek en naar de maatregelen die in het proces worden genomen om tot zorgvuldige besluitvorming en goed informatiebeheer te komen. Hierbij kijk ik in elk geval naar: (1) de (ondoorzichtigheid van de) techniek van het algoritme; (2) de dataverzameling en –verwerking; (3) welke beslissingsmogelijkheden aan het algoritmisch model worden gegeven en in hoeverre mensen daar (nog) bij worden betrokken en (4) welke impact deze beslissingen hebben op de persoonlijke levenssfeer van individuen.

[9] Vergelijk dit met menselijke hersenen: wanneer ik een beslissing neem, kan ik die weliswaar rationeel uitleggen, maar intussen knetteren er in mijn hersenen een heleboel neuronen waarvan niet te herleiden is waarom specifiek déze knetteren en hoe dat dan bijdraagt aan mijn beslissing.

3 Toepassingen

Ik laat een vergaande toepassing en een vrij simpele toepassing zien. De eerste betref t een werkend risicomodel dat moet inschatten hoe groot het risico is dat bepaalde burgers uitkeringsfraude plegen. Dit model is daadwerkelijk in gebruik genomen en is onderdeel van het gemeentelijke proces geworden, wat het interessant maakt om nader te bekijken welke maatregelen men heeft genomen ten behoeve van verantwoording en reconstrueerbaarheid van beslissingen. De tweede toepassing is juist erg simpel en omhelst een pilot om e-mails te laten categoriseren door een algoritmisch model. Het bekijken van deze toepassing is nuttig omdat hierin de relatie tussen mens en machine mooi naar voren komt.

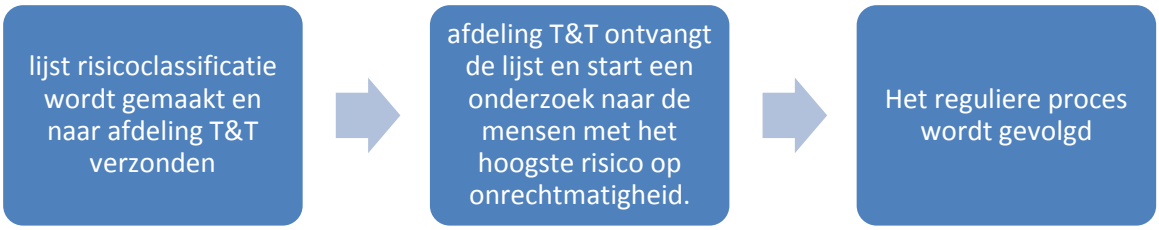
3.1 Bijstandsfraude voorspellen

In verschillende gemeentes, zoals de gemeente Rotterdam en de uitvoeringsinstantie Orionis Walcheren van de gemeentes Middelburg, Veere en Vlissingen[10], wordt geëxperimenteerd met het profileren en categoriseren van mensen die een uitkering ontvangen, met als doel uitkeringsfraude tegen te gaan. Voor onderstaande beschrijving heb ik gebruik gemaakt van projectdocumentatie.

3.1.1. het proces

Als eerste stap in het proces wordt door het algoritmisch model een lijst van personen geproduceerd met een inschatting van het risico dat deze persoon zijn uitkering op onrechtmatige gronden ontvangt. Op basis van deze categoriseringslijst wordt door de handhavers van afdeling Toetsing en Toezicht (T&T), dat wil zeggen, de mensen die controleren op bijstandsfraude door middel van contact zoals huisbezoeken of uitnodigingen tot gesprek, een onderzoek gestart naar de mensen met het hoogste risico. De eerstvolgende stap in dit onderzoek is om de persoon uit te nodigen om op gesprek te komen. Vervolgens wordt het reguliere proces gevolgd. Wanneer de lijst "op" is, wordt een nieuwe lijst geproduceerd. Hierbij is vooralsnog geen automatische feedbackloop ingebouwd. Dat wil zeggen dat de resultaten van de vorige lijst niet worden gebruikt om het algoritmisch model verder te trainen. Het model dat ten grondslag licht aan de tweede lijst blijft dus onveranderd na de initiële trainingsfase.

[10] Zie: M. Heijink (2018). Algoritme voorspelt wie fraude pleegt bij uitkeringsfraude. Online via: https://www.nrc.nl/nieuws/2018/04/08/algoritme-voorspelt-wie-fraude-pleegt-bij-bijstandsuitkering-a1598669



3.1.2. Dataverzameling

Voor dit project is gebruik gemaakt van eerder gegenereerde data die op basis van de Participatiewet (2015) verzameld wordt voor de processen van Werk & Inkomen (W&I). Deze data wordt vastgelegd in de systemen Socrates en RAAK/RMW, de informatie- en registratiesystemen van Wigo4it. Ook werd gebruik gemaakt van lijsten met resultaten van eerdere rechtmatigheidsonderzoeken, waarin ook gegevens stonden van de onderzochte personen en hun partners en kinderen. Oftewel, er is voor dit doel geen nieuwe data gegenereerd. Er is dus buiten de reeds bekende databronnen niet gekeken naar het koppelen met andere databronnen die niet gebruikt worden in het reguliere proces. De data wordt opgeslagen in het centrale datawarehouse. Deze data wordt niet per individueel geval in het zaaksysteem opgeslagen. Wel wordt genoteerd dat een persoon geselecteerd is om op gesprek te komen.

Er is dus sprake van koppeling van data uit twee verschillende bronnen, namelijk uit de systemen van Socrates en RAAK. Het model wordt niet *real-time* van nieuwe data voorzien. Met betrekking tot het beheer hiervan geldt dus het reguliere beleid met betrekking tot datawarehouses.

3.1.3. (On)doorzichtigheid van de techniek

Voor de analyse is gebruik gemaakt van een model dat een 'black box' te noemen valt. Het sorteert de cases op de kans dat een persoon onrechtmatig een uitkering ontvangt. Het model is getraind op gevallen waarvan al bekend was of ze hun uitkering onrechtmatig hadden ontvangen. Nieuwe gevallen worden vergeleken met deze voorbeelden. Er wordt gezocht naar patronen in de nieuwe gevallen die overeenkomen met patronen in de trainingsgevallen. Om patronen te herkennen, zijn zoveel mogelijk kenmerken of *features* nodig per geval. Per case zijn er zo'n 400 features. Het model zoekt naar patronen met betrekking tot deze features en kiest hierin zelf welke features in welke mate gebruikt worden voor de risicovoorspelling. We kunnen zien welke features over het algemeen belangrijk waren voor de output van het model, maar per individuele case zegt dit niets. Op de vraag: "waarom heb je me uitgenodigd om op gesprek te komen?" kan geen specifieker antwoord worden gegeven dan: "dat is op basis van ons algoritme". Wel zou je kunnen beargumenteren dat verwacht mag worden dat de gemeente überhaupt zo nu en dan contact onderhoudt met de mensen die een uitkering krijgen, ook als er geen directe aanleiding is om te denken dat er iets aan de hand is.

3.1.4. Menselijke tussenkomst

Zoals beschreven doet het algoritme autonoom een suggestie voor verder onderzoek. Hoewel de suggestie door technologie tot stand komt en bovendien rechtstreeks overgenomen wordt door de handhavers, is het aan de handhavers om verder onderzoek te doen en te beoordelen of er

inderdaad sprake is van een onrechtmatig verkregen uitkering. Uiteindelijk wordt op basis van dit onderzoek een beslissing genomen over de uitkering. Zodoende is de mens *part of the loop* en worden er geen directe beslissingen genomen door het model.

3.1.5. Impact op individuele burgers

De data bevatten persoonsgegevens, maar werden al verzameld in het kader van de uitvoering van de taken van W&I. De analyse van deze gegevens leidt tot een risicoclassificatie op basis waarvan een onderzoek gestart kan worden. De handhavers nemen uiteindelijk besluiten die impact hebben op individuen. Indirect kan de inzet van het model dus gevolgen hebben voor burgers. Aan de andere kant kun je je afvragen hoeveel méér gevolgen de inzet van het model veroorzaakt. De afdeling Toetsing en Toezicht startte ook onderzoeken vóór de ingebruikname van dit model. Toen gebeurde dit op steekproefsgewijze basis. Bij het ontvangen van een uitkering hoort nou eenmaal een zekere mate van contact met de gemeente, vindt men. Toename van het aantal onderzoeken en contactmomenten met individuele burgers, onafhankelijk van het gebruik van een algoritmisch risicomodel, heeft waarschijnlijk meer impact op het individu dan het gebruik van het risicomodel.

Wel kun je vragen stellen bij de proportionaliteit van deze methode. Vinden we bijstandsfraude werkelijk zó ernstig dat we willen proberen met behulp van risicomodellen fraude op te sporen? Moeten we onze middelen richten op de zwakkere onderkant van onze samenleving en kunnen we onze pijlen niet beter richten op bijvoorbeeld ontduiking van vermogensbelasting via belastingparadijzen? In dit kader kun je je ook afvragen of je niet beter een andere vraag kunt stellen aan je data, zoals bijvoorbeeld een vraag geframed vanuit preventie of hulp, in plaats van toezicht, toetsing en opsporen van fraude.

3.1.6. Conclusie

Het gebruik van een risicomodel om bijstandsfraude op te sporen is momenteel in gemeenteland een vrij vergaande toepassing van algoritmes. Er wordt geprobeerd aan de hand van allerlei kenmerken van individuen geprobeerd te voorspellen of iemand zijn uitkering onrechtmatig ontvangt. Er wordt een ondoorzichtige techniek gebruikt, er is potentieel veel impact op individuen.

We zien aan de andere kant dat er geen gebruik gemaakt wordt van een zichzelf *real-time* verversend model, hetgeen het beheer vergemakkelijkt. Er is één versie van het model, dat getraind is op historische data. Dit lijkt me niet moeilijk om te bewaren. Per individueel geval dat geselecteerd is kan dan in het persoonlijk dossier verwezen worden naar het model en trainingsdata. Een andere belangrijke mitigerende maatregel is dat de mens uiteindelijk de belangrijke beslissingen neemt. De output is een lijst met een prioritering. De hoogstgenoteerden op deze lijst worden uitgenodigd op gesprek. Op basis daarvan wordt nader onderzoek gestart. De invloed van het model zou nog meer verminderd kunnen worden wanneer de mens nog een afweging kan maken binnen die lijst op basis van eigen ervaring of inzichten. De vraag is natuurlijk of dit de besluitvorming transparanter en makkelijker reconstrueerbaar maakt.

3.2. E-mails selecteren om te archiveren

Omdat e-mail inboxen van werknemers privé worden geacht, wordt nu gerekend op de waakzaamheid van de werknemers om de juiste e-mails op te slaan in het betreffende zaaksysteem of DMS. E-mails kunnen belangrijke te bewaren gegevens bevatten en het is al sinds de komst van e-mail in de jaren '90 een dilemma voor archiefinstellingen om te zorgen dat de juiste e-mails bewaard blijven. Het Nationaal Archief wilde weten of algoritmische modellen dit werk uit handen van de werknemers te nemen. In een eerste pilot werden algoritmische modellen ontwikkeld die aan de tekst in de e-mail konden herkennen of deze e-mail werkgerelateerd was, of 'ruis'.

3.2.1. het proces

Deelnemers aan de pilot voerden het model met trainingsdata. Handmatig werd voor heel veel mailtjes aangevinkt of deze werkgerelateerd was of 'ruis'. Vervolgens werd het model losgelaten op nieuwe mailtjes. Het model gaf aan in welke categorie het viel en de medewerker kon dit controleren. Ieder gecontroleerd mailtje werd terug opgenomen in het model als trainingsdata, waardoor het model zich bleef aanpassen.

3.3.3. (Ondoorzichtigheid van) de techniek

Voor de analyse is gebruik gemaakt van een model dat een 'black box' te noemen valt. Het model zoekt patronen in woordgebruik en bepaalde woordgroepen om tot een conclusie te komen met betrekking tot de categorie waarin een mailtje moest vallen. Het model werd doorlopend getraind met nieuwe output. Medewerkers konden de categorisering van het model valideren, waardoor het model doorlopend veranderde om een steeds betere voorspelling te kunnen doen.

3.3.4. Menselijke tussenkomst

Aan de medewerkers werd de output getoond van het model. Medewerkers konden bij ieder mailtje de details van de afweging bekijken, bijvoorbeeld het zekerheidspercentage. Na verloop van tijd werd het model zó accuraat dat sommige mensen het begonnen te vertrouwen en de risico-afweging maakten om niet meer te controleren. De redenering was: "ik maak meer fouten dan het model nu en ik neem het enkele foutje dat het model nog maakt op de koop toe". Hieruit concluderen we dat het vertrouwen toenam

3.3.5. Impact op individuele burgers

De data bevatte geen persoonsgegevens. Er werden geen beslissingen genomen die directe gevolgen hadden voor individuele burgers.

3.3.6. Conclusie

Belangrijke maatregelen die zijn genomen ten behoeve van transparantie en verantwoording zijn dus ten eerste dat medewerkers onderdeel werden gemaakt van het proces, doordat zij zelf het model konden trainen. Daarnaast hadden zij het laatste woord en waren zij dus onderdeel van het besluitvormingsproces. Het model staat op die manier ten dienste van het proces en helpt medewerkers hun verantwoordelijkheden uit te voeren.

4 Conclusies en aanbevelingen

Er zijn talloze experimenten en pilots met zelflerende algoritmes. Er zijn zeker uitdagingen op het gebied van verantwoording, reconstrueerbaarheid en het beheer van algoritmische modellen en de daaronder liggende data. Tegelijk zijn de huidige toepassingen nog beperkt in complexiteit en lijkt voldoende grip op verantwoording en reconstrueerbaarheid goed mogelijk.

Wanneer het model ondoorzichtig is of wanneer de gebruikte data uit een veelheid van bronnen afkomstig is en *real-time* ververst wordt, maakt dat de herleidbaarheid van de totstandkoming van de output moeilijk. Wanneer daarbij bovendien de impact van die output hoog is, kun je verschillende dingen doen om te borgen dat tot voldoende verantwoording afgelegd kan worden.

Je kunt allereerst zorgen voor technische transparantie. Hieronder valt bijvoorbeeld publicatie van de broncode, analysemethodes, de variabelen en welke doorslaggevend zijn geweest voor de bouw van het model, de ingevoerde trainingsdata, validatietests, zekerheidsmarge, drempelwaarden, etc. Je kunt deze technische gegevens van woordelijke uitleg voorzien om transparantie richting "leken" te vergroten, al dan niet met behulp van (versimpelde of fictieve) voorbeelden van wat het algoritmisch model doet, welk doel het heeft, het type gegevens dat wordt gebruikt, hoe de output wordt gebruikt et cetera.

Je kunt ook zorgen voor betere verantwoording door het model te versimpelen, waardoor je in beginsel niet met bovengenoemd verantwoordingsprobleem komt te zitten. We zien bijvoorbeeld bij het model dat bijstandsfraude moet voorspellen dat geen nieuwe data *real-time* teruggegeven wordt aan het model. Het model is dus statisch. Op deze manier is voor alle gevallen helder op basis van welk model de voorspelling tot stand kwam. Vervolgens kan veel tijd worden gestoken in de beschrijving testen en validatie van het model. Ook dit komt de mogelijkheden voor verantwoording en reconstrueerbaarheid ten goede.

Ook in het proces waarin de toepassing gebruikt wordt kun je maatregelen treffen om zo te waarborgen dat je je besluiten kunt verantwoorden. In het geval van het Rotterdamse model dat onrechtmatigheid bij uitkeringen moet voorspellen, is er bijvoorbeeld voor gekozen om de output te zien als een prioriteitenlijst, waarbij een ambtenaar verder onderzoek op dient te starten om te valideren of de voorspelling klopt. Dit beperkt de mogelijkheid voor "automatische" besluitvorming door ondoorzichtige algoritmes en maakt de verantwoordelijkheid voor dat besluit duidelijker.

Kortom, de focus moet dus niet alleen liggen op de gebruikte algoritmische modellen, de gebruikte data (soort, hoeveelheid, oorspronkelijke databronnen, veranderlijkheid en opslag), en de output van het model. Ook moet aandacht gegeven worden aan de plek van het algoritmisch model in het proces. Hiermee bedoel ik wat er gebeurt naar aanleiding van de output, in hoeverre

mensen hier nog betrokken bij zijn, in hoeverre besluiten "automatisch" worden genomen door het algoritmisch model.

Hierbij is de aanname dat een belangrijkere en meer autonome plek voor het algoritme in het proces leidt tot grondiger bewaren en beheren van de techniek. Dit alles moet leiden tot een afweging over de maatregelen die je neemt om verantwoording af te kunnen blijven leggen.

In het algemeen kunnen we dus zeggen dat de menselijke betrokkenheid bij het proces toe dient te nemen naarmate de gebruikte techniek ondoorzichtiger, de data complexer en de impact hoger wordt. Sterker verwoord: automatische, algoritmische besluitvorming kan alleen wanneer ofwel de impact laag is, ofwel de techniek navolgbaar is. Voor het beheer van deze toepassingen geldt: een complexe toepassing betekent dat je waarschijnlijk strengere eisen stelt aan verantwoording van keuzes in het proces waar de toepassing deel van uit maakt.