Bijlage bij Brief over waarborgen tegen risico's van data-analyses door de overheid

Bijlage 1. Richtlijnen voor het toepassen van algoritmes door overheden

Inleiding
Doel van deze richtlijnen is het geven van handvaten ten behoeve van de ontwikkeling en het gebruik van algoritmes door de overheid en voorlichting over data-analyses door overheden aan het publiek.

Voor de toepassing van algoritmes gaat het om richtlijnen over:
- bewustzijn risico's,
- uitlegbaarheid,
- gegevensherkenning,
- auditeerbaarheid,
- verantwoording,
- validatie,
- toetsbaarheid.

Voor de voorlichting aan het publiek over data-analyses gaat het om informatieverschaffing door een overheidsdienst over onder meer:
- waarom zij data-analyses uitvoert (wat het doel ervan is en wat met de resultaten daarvan wordt gedaan),
- wat de eventuele consequenties van de analyse voor betrokken burgers zijn,
- welke databronnen van welke organisaties daarvoor worden gebruikt, en wat de kwaliteit daarvan is,
- wie verantwoordelijk voor de analyse is,
- welke kwaliteitsborging er plaatsvindt.

Aanleiding voor het opstellen van deze richtlijnen is het kabinetsstandpunt over het rapport "Big Data in een vrije en veilige samenleving" van de Wetenschappelijke Raad voor het Regeringsbeleid (WRR). Daarin heeft het kabinet aangegeven te zullen onderzoeken hoe, rekening houdend met alle relevante belangen, voor toezicht en rechterlijke toetsing, voldoende inzicht kan worden gegeven in gebruikte algoritmen en analysemethoden, met name voor situaties waarin besluitvorming op basis van een Big Data analyse rechtsgevolgen of anderszins een aanmerkelijke impact op burgers heeft. Om voor meer transparantie rond Big Data analyses door overheidsdiensten te zorgen, heeft het kabinet ook aangegeven te zullen stimuleren dat deze diensten op hun websites informatie opnemen over het doel van analyses die zij uitvoeren, en de databestanden die daarvoor worden gebruikt.[1] Met deze richtlijnen geeft het kabinet aan die toezeggingen uitvoering.

Deze richtlijnen zijn niet enkel relevant op de door de WRR geformuleerde Big Data analyses, maar op data-analyses in brede zin, zeker waar data-analyse rechtsgevolgen of anderszins een aanmerkelijke impact op burgers heeft.
Om die reden wordt hierna niet gesproken over Big Data analyses, maar over data-analyses in het algemeen. Overigens hoeft het hier niet alleen om beslissingen van de overheid te gaan. Ook beleidsonderbouwing met behulp van algoritmes kan impact op burgers hebben

[1] Kamerstukken II 2016/17, 26643, nr. 426, p. 9.

De richtlijnen zijn voorbereid door een tweetal interdepartementale werkgroepen, één op het gebied van de ontwikkeling en het gebruik van algoritmes en één op het gebied van publieksvoorlichting. De eerstbedoelde werkgroep bestond uit vertegenwoordigers van organisaties als de Belastingdienst, Politie, CBS, CJIB, NFI, Inspectie SZW, ICOV (Infobox Crimineel en Onverklaarbaar Vermogen) en met participatie vanuit de Universiteit Leiden en de Universiteit van Amsterdam.

De richtlijnen zijn vervolgens besproken in het CIO-beraad Rijk van 29 mei 2019. Vanuit het CIO-beraad is positief gereageerd op de richtlijnen, met dien verstande dat men deze graag nog verder in detail wil bespreken. De richtlijnen zijn op 27 juni 2019 ook besproken in het Overheidsbreed Beleidsoverleg Digitale Overheid (OBDO). Ook het OBDO was in algemene zin positief. Wel had men een voorkeur voor een "handreiking" boven "richtlijnen". De term "handreiking" dekt echter onvoldoende de lading en suggereert een te grote mate van vrijblijvendheid. Aan de andere kant vloeit uit de aard van richtlijnen voort dat zij geen dwingend karakter hebben. Wel is het uiteraard wenselijk deze te volgen, bijvoorbeeld in het kader van de P&C-cyclus. Daarbij zou dan bij voorkeur het principe "*comply or explain*" moeten worden gevolgd.

In vervolg op het OBDO is met de VNG afgesproken een proces in te richten dat ervoor moet zorgen dat de richtlijnen ook in voldoende mate op gemeenten worden afgestemd. In dat proces zal onder meer worden gekeken naar de uitvoerbaarheid van de richtlijnen en zal onder auspiciën van de VNG een impactanalyse voor gemeenten worden uitgevoerd.

In dit verband is tot slot nog van belang dat de richtlijnen in het Transparantielab dat door BZK is ingericht[2], zullen worden getoetst aan de hand van concrete casussen waarin data-analyse plaatsvindt. Op grond van de resultaten van deze toets kunnen zij, zo nodig, worden aangescherpt of anderszins verbeterd. Naar aanleiding van de vraag van het lid Futselaar tijdens het plenaire debat op 10 september 2019 zal in relatie tot de norm "uitlegbaarheid" tevens aandacht worden besteed aan de mate en gevallen waarin de beslisregels van het algoritme voor de burger inzichtelijk worden gemaakt.

Een en ander bevestigt het karakter van deze richtlijnen als "levende" spelregels, die onder invloed van nieuwe ontwikkelingen en ervaringen met enige regelmaat zullen worden aangepast, te beginnen met een eerste evaluatie die kort na de zomer van 2020 zal worden afgerond.

Typen algoritmes
Er bestaan verschillende typen algoritmes. Deze kunnen variëren van een eenvoudige beslisboom met een beperkt aantal variabelen tot complexe en zelflerende algoritmes, zoals machine learning of deep learning algoritmes.

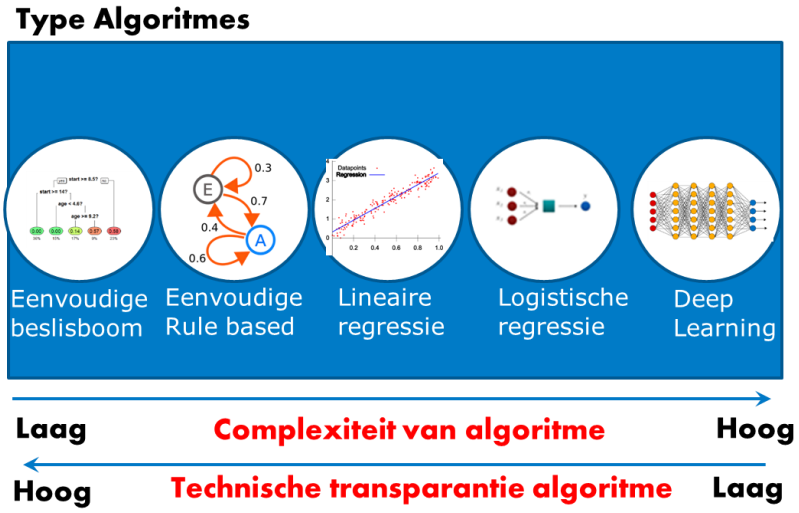
Deze laatste twee varianten vallen onder de categorie kunstmatige intelligentie. Hiermee kunnen complexe verbanden worden vastgesteld die een mens zelf moeilijk of niet kan vinden. Hierin is het redeneerproces vaak ook inherent ondoorzichtig en voor een mens lastig of niet te begrijpen.

Algoritmes zijn gelet op hun technische aard niet altijd transparant. Soms zijn deze door hun complexiteit ondoorzichtig. In andere gevallen wordt transparantie van algoritmes bewust achterwege gelaten, omdat getracht wordt "*gaming the system"* te voorkomen en zo hun werkzaamheid te behouden. Ondoorzichtigheid

[2] Zie: NL Digibeter 2019, p. 78, bijlage bij Kamerstukken II 2018/19, 26643, nr. 621.

van algoritmes kan ook het gevolg zijn van bescherming van eigendomsbelangen van hun makers te beschermen of van de enorme complexiteit van algoritmes die van nature extreem dynamisch en ondoorgrondelijk kunnen zijn.

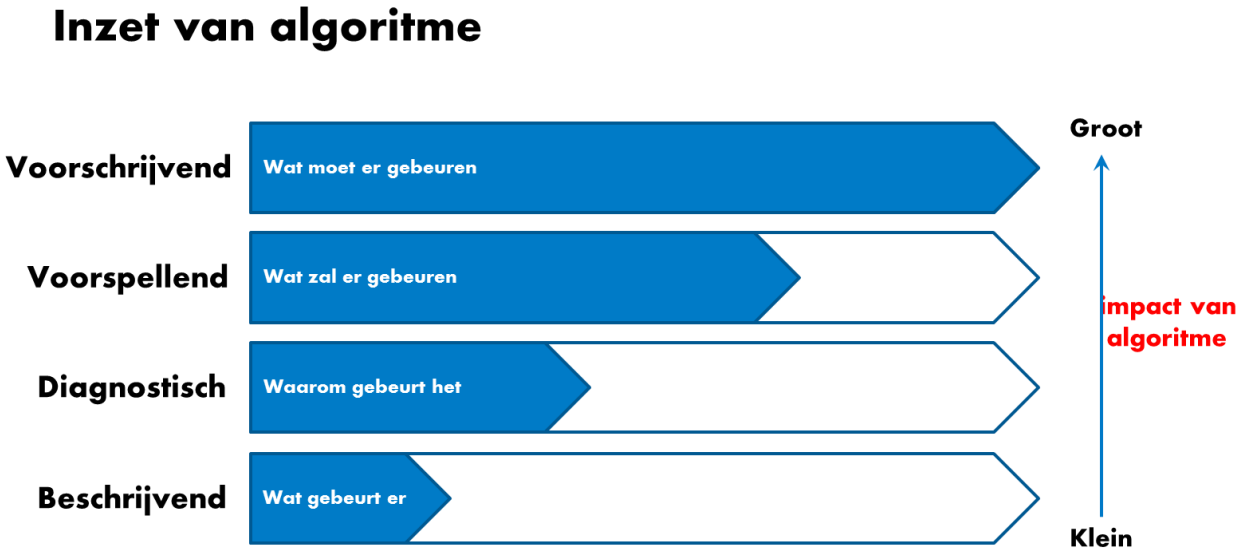
Naarmate het algoritme complexer wordt, levert "technische transparantie" (zie hierna) steeds minder begrip van het algoritme op.



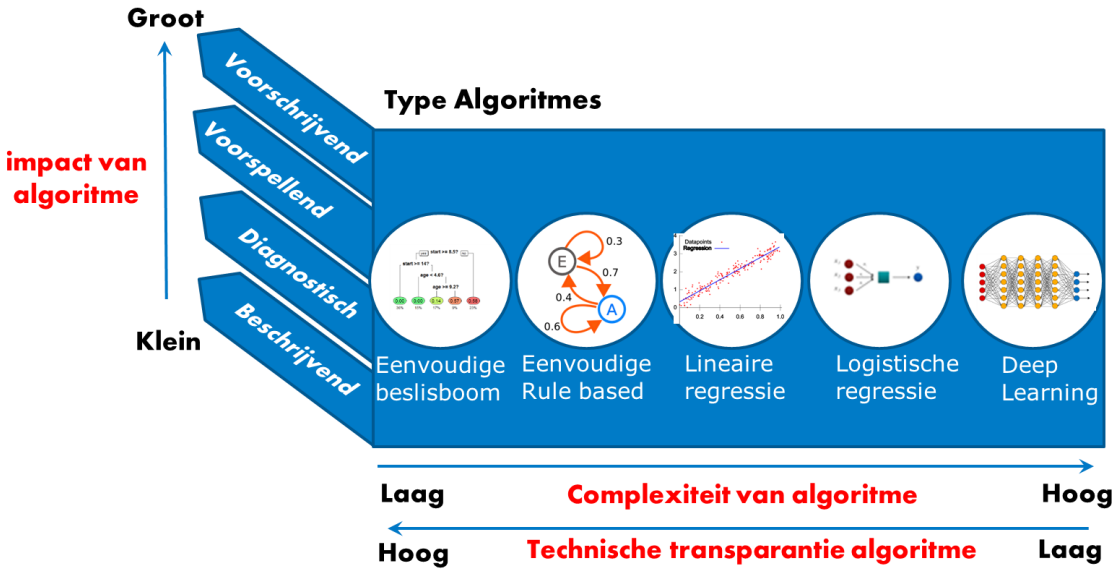
Daar waar "technische transparantie" niet steeds de gewenste helderheid brengt, kan uitlegbaarheid een belangrijke rol spelen. Bij uitlegbaarheid gaat het om uitleggen van de inzet, het doel en de uitkomst van een algoritme in begrijpelijke taal (zie ook hierna).

Inzet van algoritmes
Naast de uitleg is van belang hoe en waarvoor het algoritme wordt ingezet (welk doel). Met betrekking tot het eerste, hoe het algoritme wordt ingezet, kunnen we onderscheid maken in de volgende vier "inzetgebieden":

1. Beschrijvend – Analyse van "Wat gebeurt er?"
2. Diagnostisch – Analyse van "Waarom gebeurt het?"
3. Voorspellend – Analyse van "Wat zal er gebeuren?"
4. Voorschrijvend – Analyse van "Wat moet er gebeuren?"



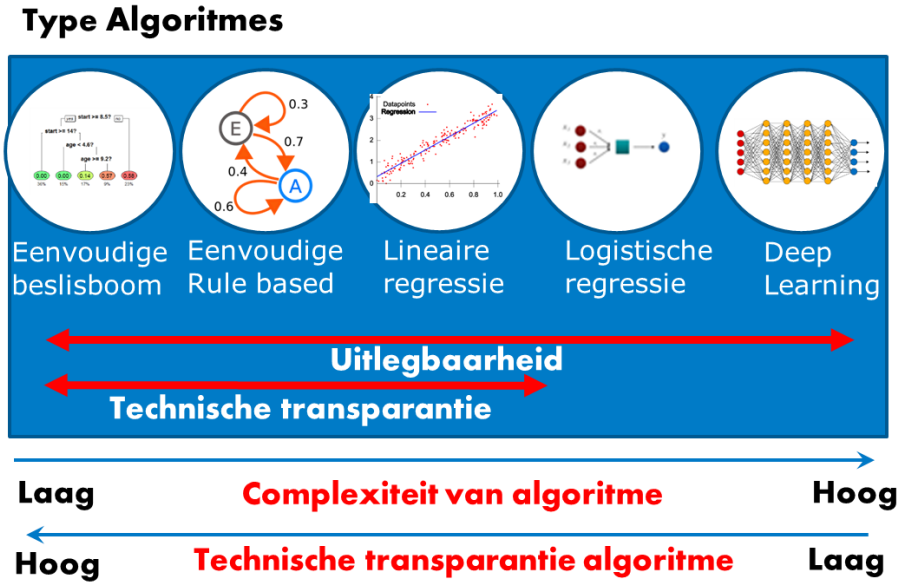
De inzet van het algoritme is relevant voor de impact ervan: een voorschrijvend algoritme heeft doorgaans meer impact dan een beschrijvend algoritme. Naast de impact neemt in het algemeen daarbij ook de complexiteit toe. Wanneer een data-analyse leidt tot een voorschrijvende uitkomst (bijvoorbeeld bij geautomatiseerde besluitvorming) zal in die analyse ook een beschrijvende, diagnostische en eventueel een voorspellende analyse moeten plaatsvinden. Dat is immers nodig om adequaat te analyseren wat de beste maatregel of beslissing is.



Technische transparantie en uitlegbaarheid
Om effectief toezicht te kunnen houden op de totstandkoming van besluitvorming door de overheid op basis van data-analyses is het nodig dat de relevante toezichthouders voldoende inzicht hebben in dataverwerkingsprocessen. Het is van belang hierbij onderscheid te maken tussen "technische transparantie" en "uitlegbaarheid".

Onder technische transparantie verstaan we het begrip van de algoritmische methode en het beschikbaar stellen van de broncode en eventuele invoer variabelen, parameters en drempelwaarden die gebruikt zijn. Bij complexe algoritmes, zoals *deep learning* algoritmes, met een grote hoeveelheid variabelen en neurale lagen kan het ook voor experts lastig zijn om op basis van deze technische transparantie het algoritme en de werking daarvan voldoende te doorgronden.

Bij uitlegbaarheid van het algoritme gaat het om het verklaren van de uitkomst van het algoritme in begrijpelijke taal. Technische transparantie kan hieraan tot op zekere hoogte bijdragen, maar kent bij zeer complexe algoritmes ook haar grenzen. Voor inherent ondoorzichtige modellen zijn wel technieken ontwikkeld of in ontwikkeling om achteraf te achterhalen op welke informatie een algoritme zijn uitkomst baseert. Tegen deze achtergrond wordt bij uitlegbaarheid van algoritmes de focus gelegd op het beschrijven van het doel dat met het algoritme wordt nagestreefd, welke variabelen doorslaggevend zijn geweest voor de uitkomst, het type gegevens dat wordt gebruikt (de kwaliteit ervan, hoe de gegevens worden gecombineerd), en de eventuele beslisregels.[3] Daarmee is uitlegbaarheid in veel gevallen veelzeggender dan technische transparantie.



Naast technische transparantie en uitlegbaarheid van het algoritme, van zowel de procedures die door het algoritme gevolgd worden als de specifieke beslissingen die zijn gemaakt, is het tevens noodzakelijk dat al bij de ontwikkeling van het algoritme rekening wordt gehouden met een aantal richtlijnen. Daarvoor zijn er de volgende richtlijnen geformuleerd:

<u>Bewustzijn risico's</u>
Eigenaren, ontwerpers, bouwers, gebruikers en andere belanghebbenden moeten zich bewust zijn van discriminerende- of stigmatiserende factoren die het ontwerp, de implementatie daarvan en het gebruik kunnen opleveren, de impact of schade die bias kunnen creëren voor individuele burgers en samenleving, en het onderscheid tussen causaliteit en correlatie voor de keuze van het algoritme. Daarbij dient ook acht te worden geslagen op het eventuele risico dat bij

[3] Kamerstukken II 2018/19, 26643, nr. 570, p. 4.

compromittering van het algoritme (door bijvoorbeeld een hack) de integriteit wordt geschaad, en als gevolg daarvan de uitkomsten van het algoritme kunnen worden gemanipuleerd. Het Cybersecurity Beeld Nederland 2019 laat zien dat er significante dreiging uitgaat van cybercriminelen en statelijke actoren.[4] Hierom dienen organisaties in te zetten op het versterken van hun digitale weerbaarheid. Door het nemen van gepaste beheersmaatregelen, waar het kabinet op inzet middels de Nederlandse Cybersecurity Agenda, kan dit risico aanzienlijk worden verkleind.[5]

Uitlegbaarheid
Overheden die gebruik maken van algoritmische besluitvorming met specifieke consequenties voor individuele burgers moeten uitleg kunnen geven over zowel de procedures die door het algoritme gevolgd worden, als de specifieke beslissingen die zijn genomen. Dit brengt als uitgangspunt mee dat overheidsorganisaties in beginsel geen algoritmes mogen hanteren die te complex zijn om te kunnen worden uitgelegd.

Gegevensherkenning
Wanneer gebruik gemaakt wordt van methoden waarbij vooraf parameters moeten worden vastgesteld of trainingsgegevens worden gebruikt, beschrijf dan de wijze waarop de parameterisering en de keuze voor trainingsgegevens tot stand is gekomen, vergezeld van een verkenning van de potentiële discriminerende factoren.

Auditeerbaarheid
Modellen, algoritmes, data en beslissingen met specifieke consequenties voor individuele burgers moeten worden vastgelegd, zodat ze geverifieerd kunnen worden in gevallen waarin schade wordt vermoed. Dit betekent een gedegen R&D proces plus documentatie waarin het gebruik van algoritmen in productie navolgbaar is.

Verantwoording
Overheden zijn verantwoordelijk voor beslissingen die door hun gebruikte algoritmes worden gemaakt, ook als deze door derden zijn gemaakt, zelfs als deze niet in detail uitlegbaar zijn. Zij dienen daarover dan ook verantwoording te kunnen afleggen.

Validatie
Overheden moeten gebruik maken van strikte methoden om hun modellen te valideren en deze methoden en resultaten documenteren. In het bijzonder moeten ze routinematig testen uitvoeren om te beoordelen en bepalen of het model het beoogde doel bereikt en geen bijkomende schade oplevert. Overheden worden aangemoedigd om de resultaten van dergelijke tests openbaar te maken.

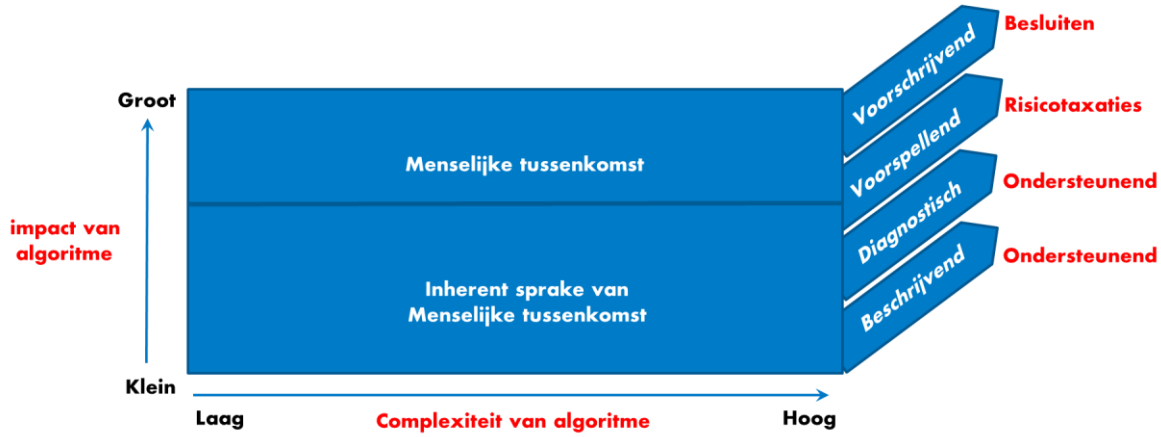
Toetsbaarheid
Data-analyse dient zo te worden ingericht dat de methode van data-analyse, de gehanteerde algoritmes, datasets en de feitelijke verwerkingen (gerechtelijk) kunnen worden getoetst.

Hierbij is van belang dat er voldoende aandacht is voor de kwaliteit van zowel het ontwikkelproces als het valideren van de resultaten.

[4] Cybersecuritybeeld Nederland 2019 (Kamerstuk 26 643, nr. 614).
[5] Nederlandse Cybersecurity Agenda (Kamerstuk 26 643, nr. 536), Voortgangsrapportage NCSA (Kamerstuk 26 643, nr. 614).

Zie voor een nadere toelichting op en uitwerking van bovenstaande zeven richtlijnen bijlage 1.1

De uitvoering van de hierboven weergegeven richtlijnen vergt relatief meer aandacht indien er sprake is van automatische besluitvorming. In onderstaande figuur is getracht duidelijk te maken dat daar waar de inzetgebieden beschrijvend, diagnostisch en voorspellend zijn er inherent sprake is van menselijke tussenkomst, dat wil zeggen dat op de uitkomst van het algoritme altijd een menselijke handeling volgt. En daar waar het inzetgebied voorschrijvend en dus besluitvormend is, er sprake is van een recht op betekenisvolle menselijke tussenkomst (artikel 22 AVG).



Dit betekent niet dat in *alle* gevallen van automatische besluitvorming er sprake is van menselijke tussenkomst. De Wet Mulder, die de administratiefrechtelijke afdoening van snelheidsovertredingen regelt, biedt een voorbeeld van volledige geautomatiseerde besluitvorming zonder menselijke tussenkomst. De keuzes en aannames zijn hier volledig uit bestaande wetten en beleidsregels af te leiden.

Op basis van de benoemde kenmerken en inzetgebieden van algoritmes kunnen we tot een categorie indeling komen.

Categorie 1: Eenvoudige algoritmes waarbij de gemaakte keuzes en aannames volledig uit wetgeving en beleidsregels zijn af te leiden.

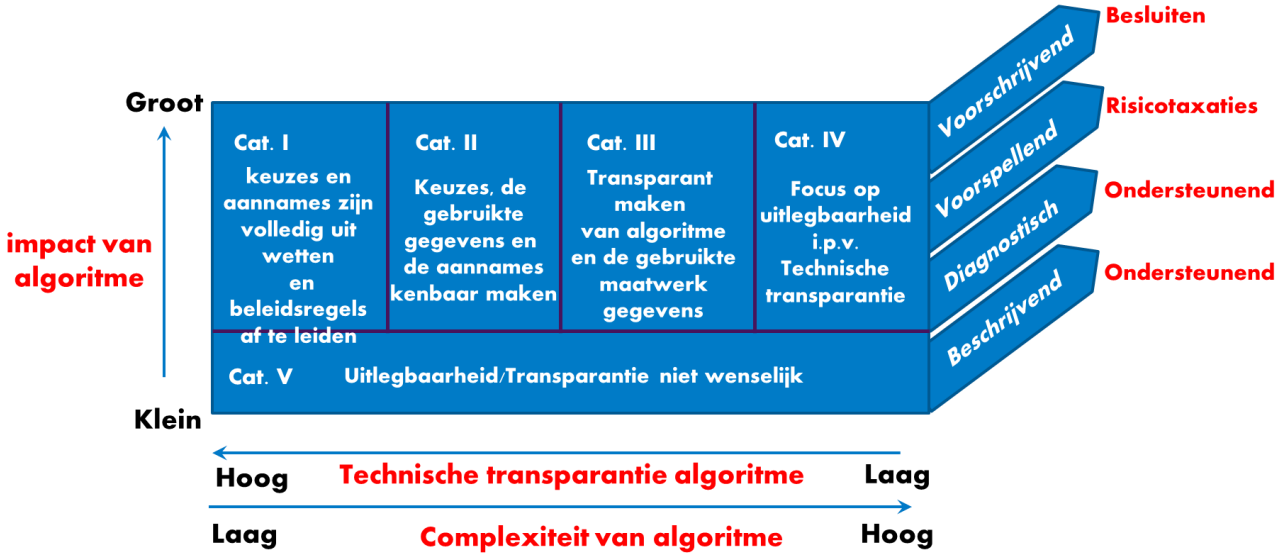
Categorie 2: Algoritmes waarbij de keuzes, de gebruikte gegevens en de aannames kenbaar moeten worden gemaakt, ter voorkoming van een ongelijkwaardige procespositie, zodat deze keuzes, gegevens en aannames kunnen worden beoordeeld en zo nodig gemotiveerd worden betwist, zodat reële rechtsbescherming tegen besluiten die op deze keuzes, gegevens en aannames zijn gebaseerd, mogelijk is.

Categorie 3: Complexe algoritmes, waarbij het technisch transparant maken van algoritmes en gebruikte maatwerkgegevens verduidelijking en inzicht oplevert van het redeneerproces.

Categorie 4: Zeer complexe algoritmes, daar waarvan technische transparantie geen inzicht geeft, maar uitlegbaarheid wordt nagestreefd. Uitgelegd dient te worden welk doel het algoritme nastreeft, welke aannames zijn gemaakt, wat het doet, hoe het wordt ingezet en de beschrijving van welke type gegevens worden gebruikt. Daarnaast dient informatie te worden verschaft over de wijze waarop

deze gegevens worden gecombineerd, de kwaliteit ervan, de keuze voor bepaalde technieken en methoden (algoritmes) en de reproduceerbaarheid en valideerbaarheid
van de resultaten.

Categorie 5: Categorie waar de uitlegbaarheid/transparantie niet wenselijk is om de werkzaamheid te behouden en ontwijkende/calculerend gedrag te voorkomen. Dit geldt bijvoorbeeld bij algoritmes die ten behoeve van opsporing en cybersecurity worden ingezet.



Publieksvoorlichting
Om voor meer transparantie rond data-analyses door overheidsdiensten te zorgen is het wenselijk dat overheidsdiensten daarover op hun websites informatie opnemen. In bijlage 1.2 zijn richtlijnen opgenomen met betrekking tot de informatie die deze diensten daartoe op hun websites kunnen opnemen.

Bijlage 1.1: Nadere toelichting/uitwerking richtlijnen algoritmes

Inleiding
Algoritmes en methoden die bij data-analyses worden gebruikt, dienen deugdelijk te zijn en aan wetenschappelijke criteria voor goed (statistisch) onderzoek te voldoen. Hierbij is van belang dat voldoende aandacht is voor zowel de kwaliteit van het proces als het valideren van de resultaten.

Uitgangspunt dient verder te zijn dat de logica achter de analyse en dus de gebruikte algoritmes transparant zijn. Het gaat daarbij niet alleen om gekozen algoritmes, maar ook om andere aspecten waaronder bijvoorbeeld de kwaliteit van gebruikte databronnen.

Het is vooraf niet exact voor te schrijven waaraan de analysefase moet voldoen, want dit is per geval verschillend. Wel kan in algemene zin gesteld worden dat naarmate de impact van de uitkomst van het algoritme groter is, de eisen aan de kwaliteit van de data en de deugdelijkheid van de gehanteerde methoden zwaarder dienen te zijn. Zeker daar waar algoritmes worden gebruikt met mogelijke specifieke gevolgen voor individuele burgers.

1. Bewustzijn risico's
Eigenaren, ontwerpers, bouwers, gebruikers en andere belanghebbenden van analytische systemen die gebruikt worden voor toepassingen met specifieke consequenties voor individuele burgers, moeten zich bewust zijn van de mogelijke discriminerende- of stigmatiserende factoren die het ontwerp, de implementatie daarvan en het gebruik kunnen opleveren, de impact of schade die biases kunnen creëren voor individuele burgers en samenleving, en het onderscheid tussen causaliteit en correlatie voor de keuze van het algoritme.

Met als specifieke aandachtspunten:
- Test het algoritme op basis van test cases of scenario's en evolueer test cases periodiek en elke keer als de software verandert om te voorkomen dat nieuwe fouten ontstaan dan wel functionaliteit onbedoeld wordt aangepast.
- Houd rekening met het feit dat een toename van de gegevens in de tijd van invloed kan zijn op de uitkomsten van een algoritme en corrigeer daar zo nodig op.
- Er moeten binnen de grenzen van de wetgeving[6] controlemechanismes worden opgebouwd die specifiek toetsen of er geen sprake is van discriminatie of stigmatisering.
- Onderzoek de kwaliteit van databronnen en breng in kaart welke beperkingen het gebruik van een databron kent, bijvoorbeeld doordat er rechten zijn gevestigd op de databron of op het algoritme en analysemethode. Een ander voorbeeld is dat een databron waarvan de kwaliteit van de informatie slechts globaal of veranderlijk is, minder geschikt kan zijn voor een werkwijze die leidt tot het doen van uitspraken over individuen, maar een dergelijke bron kan wel geschikt zijn voor het beschrijven van groepen. Datakwaliteit is essentieel voor het kunnen uitvoeren van een gedegen analyse en brongegevens bevatten eigenlijk altijd biases en fouten. Beschrijf daarom het doel van de analyse en hoe je omgaat met de fouten in de datasets en uitkomsten.
- Maak een bewuste keuze voor data-analyse technieken. Het is bijvoorbeeld niet zondermeer nodig kunstmatige intelligentie methoden in te zetten op

[6] De vigerende wetgeving laat controle met behulp van zgn. bijzondere persoonsgegevens, zoals gegevens over iemands ras, vooralsnog niet toe.

data. Vaak zijn min of meer traditionele analysemethoden geschikt om de kwaliteit van bronnen te onderzoeken of om patronen te vinden. Belangrijk is ook of je wilt werken met vooraf bedachte hypothesen die je wilt toetsen d.m.v. data-analyse, of dat je zonder hypothese te werk wilt gaan. In dat laatste geval is het in het algemeen lastiger om een werkwijze te legitimeren.
- Hanteer een standaard methodiek van werken, bijvoorbeeld CRISP DM (Cross Industry Standard Process for Data Mining). Deze methodiek is toepasbaar bij data mining (techniek om patronen in datasets te ontdekken). Wanneer je een hypothese wil toetsen aan de hand van data-analyse is deze methodiek minder van toepassing.
- Hanteer gelet op de aard van gegevensverwerking juridische en beleidsmatige toetsingskaders en gebruik instrumenten als een gegevensbeschermingseffectbeoordeling om risico's voor de privacyrechten van betrokkenen zoveel mogelijk te beperken.
- Ga na of het algoritme per se van causaliteit moet uitgaan. Dat zal in het algemeen gelden voor algoritmes die gebruikt worden voor besluitvorming. Dat impliceert dat in dat geval het gebruik van deep learning minder voor de hand ligt, omdat die techniek in toenemende mate gebruik gemaakt van correlaties, d.w.z. statistische verbanden.

2. Uitlegbaarheid
Overheden die gebruik maken van algoritmische besluitvorming met specifieke consequenties voor individuele burgers moeten in begrijpelijke taal uitleg kunnen geven over zowel de procedures die door het algoritme gevolgd worden, als de specifieke beslissingen die zijn genomen. Dit brengt als uitgangspunt mee dat overheidsorganisaties in beginsel geen algoritmes mogen hanteren die te complex zijn om te kunnen worden uitgelegd.

Hierbij zijn onder meer de volgende vragen van belang: hoe is het model ontwikkeld, welke data en welke algoritmes zijn gebruikt, hoe zijn deze verkregen, hoe zijn deze intern getoetst en in welke vorm worden de resultaten van data-analyse gepubliceerd.
Met als specifieke maatregelen:
- Organiseer de code in modules welke separaat en gecombineerd kunnen worden geëvalueerd.
- Test deze modules op correcte functionaliteit zowel afzonderlijk als in combinatie.
- Leg de gehanteerde analysemethode uit en meet de nauwkeurigheid.
- Leg de input gegevens (brondata/datasets) vast die gebruikt worden en gebruik daarbij enkel relevante data. Documenteer en leg dit vast.
- Beschrijf en onderzoek de kwaliteit van de gebruikte databron(nen).
- Leg de aannames die gehanteerd zijn vast.

3. Gegevensherkenning
Wanneer gebruik gemaakt wordt van methoden waarbij vooraf parameters moeten worden vastgesteld of trainingsgegevens worden gebruikt, beschrijf dan de wijze waarop de parameterisering en de keuze voor trainingsgegevens tot stand is gekomen, vergezeld van een verkenning van de potentiële discriminerende factoren. Leg de trainingsgegevens of andere gebruikte informatie om te komen tot parametrisering vast, zodat het mogelijk is resultaten te reproduceren. Maak, indien mogelijk, analyses met betrekking tot de gevolgen die een andere keuze van parameterisering of inzet van trainingsdata (meer of minder, volgorde van aanbieden van trainingsdata) heeft op de resultaten.

4. Auditeerbaarheid
Modellen, algoritmes, data en beslissingen met specifieke consequenties voor individuele burgers moeten worden vastgelegd, zodat ze geverifieerd kunnen worden in gevallen waarin schade wordt vermoed. Dit betekent een gedegen R&D proces plus documentatie waarin het gebruik van algoritmen in productie navolgbaar is.

Met als specifieke aandachtspunten:

- Werk niet met confidentiële algoritmes, maar met open algoritmes, welke toegankelijk zijn voor toezichthouders ter controle en bij voorkeur ook voor experts en burgers. Dat impliceert het volgende:
    - het algoritme dient niet-confidentieel te zijn;
    - het algoritme dient gedocumenteerd te zijn;
    - gebruik zoveel mogelijk algoritmes en analysemethoden die wetenschappelijk gevalideerd zijn;
    - gebruik indien mogelijk algoritmes die reeds open source zijn of stel deze als open source beschikbaar.

  Er kunnen redenen zijn om van bovenstaande richtlijnen op basis van een juiste onderbouwing af te wijken.
- Onderbouw keuzes, zoals de keuze voor specifieke algoritmes en gebruikte data.
- Noteer waarnemingen, zoals afwijkingen in de gegevens of onverwachte/onverklaarbare resultaten.
- Gebruik eenvoudige methoden boven complexe methoden daar waar mogelijk. Dit komt ten goede aan de uitlegbaarheid, auditeerbaarheid en beperking van risico's.
- Verifieer zowel voor statistische analyse methoden als complexere methoden de uitkomsten gebaseerd op de specifieke input.
- Lever een gedetailleerde omschrijving van het algoritme en de werking ervan, samen met een machine-controleerbare validatie dat de code overeenkomt met de specificatie.
- Zorg voor reproduceerbaarheid.

5. Verantwoording
Overheden zijn verantwoordelijk voor beslissingen die door hun gebruikte algoritmes worden gemaakt, ook als deze door derden zijn gemaakt, zelfs als deze niet in detail uitlegbaar zijn. Zij dienen daarover dan ook verantwoording af te leggen.

6. Validatie
Overheden moeten gebruik maken van strikte methoden om hun modellen te valideren en deze methoden en resultaten documenteren. In het bijzonder moeten ze routinematig testen uitvoeren om te beoordelen en bepalen of het model het beoogde doel bereikt en geen bijkomende schade oplevert. Overheden worden aangemoedigd om de resultaten van dergelijke tests openbaar te maken.

7. Toetsbaarheid
Data-analyse dient zo te worden ingericht dat de methode van data-analyse, de gehanteerde algoritmen, datasets en de feitelijke verwerkingen (gerechtelijk) kunnen worden getoetst. Wanneer data-analyse wordt toegepast ten behoeve van besluitvorming dient het algoritme, gelet op de transparantieverplichting die onderdeel uitmaakt van de algemene beginselen van behoorlijk bestuur, te beschikken over een toereikend niveau van transparantie, verifieerbaarheid en toetsbaarheid. De bestuursrechter stelt op grond van het bestuursrecht eisen aan

de inzichtelijkheid, controleerbaarheid en toegankelijkheid in geval van geautomatiseerde besluitvorming door bestuursorganen. Voor een nieuw toetsingskader voor de beoordeling van geautomatiseerde besluitvormingsprocessen m.b.v. algoritmes kan de uitspraak van de Afdeling bestuursrecht van de Raad van State van 17 mei 2017als uitgangspunt worden genomen. [7]

Wanneer persoonsgegevens worden verwerkt door middel van een algoritme, zal dit gelet op de beginselen inzake verwerking van persoonsgegevens op een manier moeten plaatsvinden die voor betrokkenen rechtmatig, behoorlijk en transparant is.

[7] ECLI:NL:RVS:2017:1259, i.h.b. rechtsoverwegingen 14.3 en 14.4, https://www.recht.nl/rechtspraak/uitspraak/?ecli=ECLI:NL:RVS:2017:1259.

Bijlage 1.2 : Richtlijnen inzake publieksvoorlichting over data-analyses

Inleiding
Deze richtlijnen hebben vooral betekenis in het geval dat bij het uitvoeren van data-analyses persoonsgegevens worden verwerkt. Met het oog daarop is ook aandacht besteed aan relevante voorschriften uit de Algemene verordening gegevensbescherming (AVG).
De gewenste mate van publieksvoorlichting zal altijd gebonden zijn aan de context waarin de data-analyses plaatsvinden: een overheid als dienstverlener zal doorgaans meer transparantie kunnen betrachten dan een overheid die als toezichthouder of opsporingsinstantie optreedt.

Waarom transparant?
Transparantie rond data-analyses kan bijdragen aan het vertrouwen dat burgers in deze analyses hebben. In zoverre is transparantie vooral een middel, waar vertrouwen het doel moet zijn. Transparantie zal de burger ook beter in staat stellen de werkwijze van de overheid bij een data-analyse te controleren en aldus kunnen bijdragen aan een zo evenwichtig mogelijke verhouding tussen burger en overheid. Transparantie kan ook leiden tot een betere naleving van wet- en regelgeving: het noemen van variabelen of drempelwaarden die de overheid bij data-analyses hanteert, kan enerzijds calculerend gedrag in de hand werken, maar anderzijds juist gedaan worden om bepaalde acties van burger te voorkomen (functie van *nudging*). Als in een data-analyse persoonsgegevens worden verwerkt, heeft transparantie op grond van artikel 5 AVG bovendien het karakter van een zorgplicht.[8] Tegen die achtergrond kunnen deze richtlijnen ook worden gezien als een invulling van deze zorgplicht. Het gaat in deze richtlijnen louter om algemene informatie, bestemd voor het publiek. Verplichtingen uit de verordening om een betrokken burger individueel over verwerking van hem betreffende persoonsgegevens te informeren[9], blijven hier dus buiten beschouwing.

Algemene voorwaarden aan transparantie
De algemene informatie die over data-analyses aan het publiek wordt verstrekt, dient, waar het de verwerking van persoonsgegevens betreft, aan een aantal voorwaarden te voldoen. Zij moet in de eerste plaats beknopt en transparant, begrijpelijk en gemakkelijk toegankelijk zijn. "Beknopt en transparant" wil in dit verband zeggen dat de informatie efficiënt en bondig moet worden gepresenteerd. Als een overheidsdienst de informatie in een *privacystatement* op haar website opneemt, kan dit betekenen dat de informatie daarin "gelaagd" wordt opgenomen, waarbij men snel kan navigeren naar relevante passages, zonder door het gehele privacystatement te hoeven scrollen. "Begrijpelijk" impliceert dat een gemiddelde vertegenwoordiger van het beoogde publiek de informatie moet kunnen begrijpen. Het kan in dit verband nuttig zijn om af en toe bij het actuele publiek te checken of dit het geval is, bijvoorbeeld door middel van een gebruikerspanel. "Gemakkelijk toegankelijk" betekent dat men weinig moeite hoeft te doen om toegang tot de informatie te krijgen. Opneming daarvan in een *privacystatement* op de eigen website met een duidelijke link op de *homepage* kan daaraan bijdragen. De informatie moet ook in duidelijke en eenvoudige taal worden verschaft, vooral als de informatie voor kinderen is bestemd. Zij moet

[8] Transparantie is niet een principe dat expliciet in een artikel van Richtlijn 2016/680 ("Richtlijn gegevensbescherming opsporing en vervolging") wordt genoemd, maar wel in overweging 26 van die richtlijn.
[9] Zie de artikelen 13 en 14 AVG.

geen ruimte voor verschillende interpretaties geven en in ieder geval duidelijkheid verschaffen over het doel en de wettelijke grondslag van de analyses.[10]

Waarover transparant?
Als een overheidsdienst data-analyses verricht, dient deze dienst op haar website het publiek te informeren over:

- dat zij data-analyses uitvoert,
- waarom zij data-analyses uitvoert (wat het doel ervan is en wat met de resultaten daarvan wordt gedaan),
- wat de eventuele consequenties van de analyse voor betrokken burgers zijn,
- eventuele toepassing van *machine learning* en de uitleg daarvan,
- wat de wettelijke grondslag voor het uitvoeren van deze analyses is,
- welke databronnen van welke organisaties daarvoor worden gebruikt, en wat de kwaliteit daarvan is,
- wie verantwoordelijk voor de analyse is,
- wat de rol van eventuele derden bij deze analyses is,
- welke kwaliteitsborging er plaatsvindt (welke risico's worden onderkend, welke maatregelen daartegen worden genomen en op welke wijze toetsing plaatsvindt; hoe data en model "schoon" worden gehouden om *loop* effecten te voorkomen),
- hoe er tussen analyse en een eventueel besluit menselijke tussenkomst plaatsvindt, en
- welke toetsingskaders er zijn en hoe de uitvoering van de analyses wordt geëvalueerd.

Inzichtelijkheid algoritmes?
Een overheidsdienst kan ervoor kiezen de algoritmes die zij voor haar data-analyses gebruikt, inzichtelijk te maken voor het publiek, maar kan daarmee niet volstaan. Een burger zal deze algoritmes immers doorgaans niet of nauwelijks begrijpen. Informatie over de toepassing van algoritmes en over het proces van bijvoorbeeld kwaliteitsbewaking zijn belangrijker. Dit laat onverlet dat een overheidsorganisatie zelf altijd zoveel mogelijk inzicht moet hebben over hoe een algoritme werkt en dit moeten kunnen uitleggen aan derden.

Crystal box
Om de transparantie rond data-analyses te vergroten zou een overheidsdienst een zgn. C*ristal box* kunnen ontwikkelen waarmee je inzicht in de analyse kunt krijgen en daarop controle kunt uitoefenen, zonder dat je van de technische details op de hoogte hoeft te zijn. Zo'n *Cristal box* geeft inzicht in het analyseproces, de gebruikte algoritmes en de gebruikte datasets. Zo kan men beter achterhalen welke beslissingen in het analyseproces zijn genomen en welke variabelen zijn gebruikt.

Hoe groter de gevolgen, des te belangrijker de transparantie
Naarmate de gevolgen van data-analyses voor burgers groter zijn, is transparantie belangrijker. Dat geldt zeker in gevallen waarin dergelijke analyses tot verkeerde conclusies of besluiten (kunnen) leiden.

Getrapte transparantie
De gehanteerde transparantie kan getrapt zijn. Dit betekent dat, als een overheidsdienst over haar data-analyses wel in algemene termen transparantie betracht maar bepaalde, meer gedetailleerde aspecten met het oog op bijvoorbeeld het belang van de opsporing niet openbaar maakt, zij een dergelijke

[10] Zie nader artikel 12, eerste lid, AVG en de *Guidelines on transparency under Regulation 2016/679 (AVG) van de Article 29 Data protection working party*, paragraaf 6-18.

handelswijze tenminste compenseert met voldoende intern en extern toezicht op die aspecten.

Voorkom *gaming the system*
Bij het betrachten van transparantie zal rekening moeten worden gehouden met het risico van *gaming the system*: wanneer (veel) inzicht in methoden en data wordt gegeven kunnen kwaadwillenden hier misbruik van maken. Zo kan het wenselijk zijn dat een overheidsdienst in haar privacystatement wel informatie verschaft over variabelen die zij in haar analyses hanteert, maar niet over drempelwaarden. Dit laat onverlet dat overheidsdiensten bij verwerking van persoonsgegevens in de data-analyse op grond van de AVG een betrokken burger individueel over verwerking van hem betreffende persoonsgegevens moet informeren. Maar ook dat zij dan op grond van sectorspecifieke wetgeving deze verplichting mogelijk categorisch buiten toepassing kunnen laten dan wel op grond van artikel 41 van de Uitvoeringswet AVG in individuele zaken buiten toepassing kunnen laten.[11]

Transparantie over *qualifiers*
Met het oog op de transparantie van data-analyses zijn vooral de *qualifiers* (variabelen en drempelwaarden) binnen een algoritme van belang. Welke *qualifiers* zorgen ervoor dat men tot een risicoprofiel komt, en kunnen deze *qualifiers* inzichtelijk worden gemaakt? Een toetsingscommissie zou aan de hand van *case studies* kunnen toetsen wanneer diensten over deze *qualifiers* wel en niet transparant kunnen zijn. Daarbij zou kunnen worden overwogen om, indien een risicoprofiel invloed heeft op de rechten en plichten van iemand, deze *qualifiers* voor hem of haar transparant te maken. Deze vorm van transparantie zou dan om *gaming the system* te voorkomen bij voorkeur niet vooraf in het proces moeten plaatsvinden maar achteraf.

Transparantie en toetsbaarheid
Transparantie van algoritmes voor het grote publiek en de mogelijkheid tot toetsing van een algoritme dienen niet met elkaar te worden verward. Ook al zou het weinig zin hebben om transparant te zijn over de gehanteerde algoritmes zelf, zij dienen wel toetsbaar te worden gemaakt. Transparantie en toetsbaarheid hebben in zoverre weer wel met elkaar te maken dat, indien een data-analyse door een deskundige "onafhankelijke" partij is getoetst en in orde bevonden, het wenselijk is het publiek daarvan in kennis te stellen, zodat een burger zelf niet alles tot in detail hoeft te begrijpen.

Proeftuinen
Als een overheidsdienst in een zgn. proeftuin met data-analyses wil gaan experimenteren, is het wenselijk in dat kader ook duidelijkheid te verkrijgen over de wijze waarop men transparantie kan en wil betrachten. Bij twijfel hierover is het wenselijk daarover de discussie aan te gaan en naar buiten te treden.

[11] Bij een beroep op artikel 41 zal dan moeten worden aangetoond dat dit noodzakelijk of evenredig is ter waarborging van een aantal in dat artikel genoemde belangen. Daartoe behoren, kort gezegd, onder meer de nationale veiligheid, de landsverdediging, de openbare veiligheid, de voorkoming, het onderzoek, de opsporing en de vervolging van strafbare feiten en toezicht of inspectie.

Bijlage 2. Toelichting op typen data-analyses waarvoor wettelijke waarborgen zullen gaan gelden

Het kabinet heeft voor ogen om wettelijke waarborgen te realiseren voor twee typen data-analyses door de overheid, te weten:

- Profilering, in de betekenis die daaraan wordt gegeven in de AVG en Richtlijn;
- Gebiedsgebonden analyse waarbij ook sprake is van verwerking van persoonsgegevens en van soortgelijke risico's als die welke zich bij profilering voordoen.

In deze bijlage wordt een nadere toelichting gegeven op beide typen data-analyses.

Profilering

Onder 'profilering' verstaat artikel 4, onderdeel 4, AVG: 'elke vorm van geautomatiseerde verwerking van persoonsgegevens waarbij aan de hand van persoonsgegevens bepaalde persoonlijke aspecten van een natuurlijke persoon worden geëvalueerd, met name met de bedoeling zijn beroepsprestaties, economische situatie, gezondheid, persoonlijke voorkeuren, interesses, betrouwbaarheid, gedrag, locatie of verplaatsingen te analyseren of te voorspellen'.[12] Artikel 3, onderdeel 4, Richtlijn bevat eenzelfde definitie.

In de Richtsnoeren van de *European Data Protection Board* (EDPB) is deze definitie als volgt uitgewerkt: Het moet een g*eautomatiseerde* vorm van verwerking zijn, die betrekking heeft op *persoonsgegevens*, met als doel het *evalueren of beoordelen van persoonlijke aspecten* van een natuurlijk persoon.[13] Profilering wordt, aldus de Richtsnoeren, toegepast om bepaalde voorspellingen te doen of conclusies te trekken over een individu en diens gedrag.[14] Aan deze individuele beoordeling gaat een geautomatiseerde verwerking vooraf waarbij persoonsgegevens uit verschillende bronnen worden verzameld, geanalyseerd en/of gecombineerd met als doel een persoon eigenschappen toe te kennen op basis van kenmerken van anderen die tot een bepaalde categorie (profiel) behoren.[15] Het kan hierbij ook gaan om kenmerken van anderen die in statistisch opzicht vergelijkbaar zijn.[16]

Dit element, het toekennen van eigenschappen op basis van kenmerken van anderen, is een wezenlijk kenmerk dat profilering aanzienlijk risicovoller maakt dan andere vormen van geautomatiseerde verwerkingen van persoonsgegevens. Op deze wijze kunnen de negatieve kenmerken van een bepaalde groep namelijk aan een individu worden toegewezen dat deze kenmerken helemaal niet hoeft te

[12] Zie ook Overweging 72 AVG: 'Voor profilering gelden de regels van deze verordening betreffende de verwerking van persoonsgegevens, bijvoorbeeld de rechtsgronden voor verwerking of beginselen van gegevensbescherming'.
[13] EDPB Richtsnoeren inzake geautomatiseerde individuele besluitvorming en profilering voor de toepassing van Verordening (EU) 2016/679, gewijzigd en vastgesteld op 6 februari 2018.
[14] Richtsnoeren, p. 7-8. Wanneer een bedrijf alleen voor statistische of administratieve doeleinden zijn klanten indeelt op basis van bepaalde kenmerken zonder daar conclusies aan te verbinden voor dat individu is er geen sprake van profilering. Hiervan is wel sprake in het bekende voorbeeld van webwinkels die, op basis van analyse en vergelijking van koopgedrag, bezoekers suggesties geven over mogelijke aankopen: *"Andere bezoekers die dit product kochten, waren ook geïnteresseerd in de volgende producten".*
[15] Hieruit volgt dat het opleggen van snelheidsboetes op grond van bewijs van flitscamera's een geautomatiseerd proces is waarbij geen sprake is van profilering. De beoordeling vindt namelijk plaats op basis van vooraf en objectief vastgestelde variabelen die in een directe relatie tot de betrokken persoon staan; uitsluitend de individuele kenmerken van de betrokken persoon zijn beslissend. Zie ook Richtsnoeren p. 9. Hetzelfde geldt voor het doorvoeren van een correctie op een belastingaanslag na automatische vergelijking van de aangifte met bij de belastingdienst beschikbare contra-informatie over de belastingplichtige.
[16] Richtsnoeren p. 7-8. Dit betekent dat er bij profilering sprake kan zijn van het blootleggen en vaststellen van statistische verbanden of gevolgtrekkingen (correlaties) die op voorhand nog niet bekend waren en vervolgens op een persoon worden toegepast om kenmerken van diens huidig of toekomstig gedrag vast te stellen.

hebben (*false positive*) of andersom ten onrechte niet worden toegewezen (*false negative*). De bijzondere risico's die hiermee samenhangen maken, volgens het kabinet, dat specifiek gekeken moet worden naar de benodigde waarborgen bij het toepassen van profilering.

Vergelijkbare risico's doen zich voor bij risicotaxatiemodellen zoals in gebruik bij diverse overheidsorganisaties. In dergelijke modellen worden persoonsgegevens geanalyseerd met als doel om op basis van risico-indicatoren op individueel niveau voorspellingen te doen ten aanzien van mogelijke onregelmatigheden. Ook voor deze modellen doet het risico zich voor dat men ten onrechte een bepaalde eigenschap krijgt toegekend (*false positive*) of ten onrechte niet krijgt toebedeeld (*false negative*). In het vervolgtraject zal dan ook nader onderzocht worden of ook deze instrumenten omgeven moeten worden met de nodige waarborgen.

Gebiedsgebonden analyses
Het is ook mogelijk om analyses uit te voeren die, anders dan profilering, niet op individuen zijn gericht (en dus ook niet onder de AVG-definitie van profilering vallen), maar op geografische gebieden, zoals een wijk of straat.

Van gebiedsgebonden analyses zijn enkele voorbeelden te geven. Te denken valt aan het analyseren van (persoons)gegevens met als doel het doen van een voorspelling over de (verhoogde) kans dat een strafbaar feit wordt gepleegd in een bepaald gebied in een bepaalde tijdsperiode.[17] Ook valt te denken aan het uitvoeren van gegevensanalyses om bepaalde vormen van ondermijnende criminaliteit in verschillende delen van een stad inzichtelijk te maken.[18] Bij deze analyses worden weliswaar persoonsgegevens verwerkt, maar de uitkomsten van de analyses worden niet toegepast op een natuurlijke persoon. Er worden patronen en fenomenen blootgelegd die op geaggregeerd niveau inzicht geven in crimineel gedrag. Hierdoor kan schaarse overheidscapaciteit effectiever worden ingezet. Voorbeelden daarvan zijn surveillance inzet, een thematische aanpak of afspraken met externe partners zoals gemeenten. Sommige risico's die zich voordoen bij profilering zijn ook aanwezig bij deze analyses. Te denken valt aan de aanwezigheid van bias waardoor bepaalde gebieden onevenredig vaak naar voren komen uit data-analyses. Ook voor dit type analyses acht het kabinet het daarom aangewezen om te bezien of daarvoor aanvullende waarborgen dienen te gelden.

[17] Zie hierover ook: Vetzo c.s., p. 26-27, onder *Predictive policing*.
[18] Zie hierover ook: Stcrt. 2017, nr. 48699.