CBPL

Commissie voor de bescherming
van de persoonlijke levenssfeer

Big Data Rapport

Verslaggever: Frank De Smet
bijgestaan door Cliff Beeckman
en Dieter Verhaeghe

Inhoud

Deel 1
Inleiding

Volgens IBM (2016) creëren we elke dag zo'n slordige 2,5 quintiljoen bytes (of 2,3 triljoen gigabytes) aan data – zoveel dat 90% van alle data in de wereld vandaag alleen al in de laatste twee jaar werd gecreëerd.[1] Deze data komen van zowat overal: e- mails, informatie die we achterlaten op sociale media, mobiele apps, digitale foto's en video's, zoekopdrachten in Google of andere zoekmachines, sensoren, online en offline aankooptransacties, GPS signalen, wearables, enz.

Het gegeven waarbij innovatieve technologieën nieuwe mogelijkheden bieden om waarde te putten uit deze tsunami aan beschikbare data kan men vaagweg aanduiden met de term "Big Data".

Dankzij de vooruitgang in de ICT is men vandaag in staat om enorme hoeveelheden aan gegevens te verzamelen, te bewaren en te analyseren. De geautomatiseerde analyse van dergelijke grote gegevensbestanden levert niet alleen veel tijdswinst op, maar kan – indien zorgvuldig uitgevoerd – ook leiden tot betere informatie en kennis, op basis waarvan men vervolgens en potentieel meer accurate beslissingen kan nemen en meer betrouwbare voorspellingen kan maken. Big Data-toepassingen beloven dus veel potentiële voordelen: gerichtere controles in fraudebestrijding, het nauwkeurig in kaart brengen van mensenstromen, gezondheidszorg op maat ("personalised medicine"), enz.

Maar Big Data houdt ook risico's in, in het bijzonder met betrekking tot de privacy van de burger. Zo bestaat er bij Big Data-analyses bijvoorbeeld het risico dat ze gevoelige informatie blootleggen uit oorspronkelijk niet-gevoelige gegevens (zie Punt M hierna). Ook zijn er bijvoorbeeld risico's met betrekking tot de impact op de rechten en vrijheden van de betrokkenen (zie Punten E, N en O hierna) en het gebrek aan transparantie (zie Punt L hierna).

Dit rapport is tweeledig. In het eerste deel gaan we in op de vraag wat Big Data precies inhoudt. Een goed begrip van "Big Data" is immers belangrijk om de gevolgen in te schatten op het vlak van privacybescherming. Desalniettemin moet het onderscheid tussen de eerder klassieke data-analyse en analyse bij Big Data gerelativeerd worden in het kader van privacy-discussies. De regelgeving moet voldoende algemeen, robuust en technologieneutraal zijn (en is dat voor een groot deel ook) zodat bij iedere verwerking de betrokkenen voldoende worden beschermd en een goede balans gevonden wordt om zowel de rechten en vrijheden van de betrokkenen als legitieme opportuniteiten niet teveel te fnuiken. Afdoende waarborgen dienen geboden te worden waaronder legitiem, maatschappelijk verantwoord[2] gebruik mogelijk is. Dat laatste vormt het onderwerp van het tweede deel waarin een analyse wordt gemaakt van de huidige wet- en regelgeving toegepast op Big Data.

[1] IBM: https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html. Geraadpleegd op 18 mei 2016.

[2] Hieronder verstaan we het ethisch en sociaal bewust omgaan met persoonsgegevens. Toepassingen die discriminatoire gevolgen voor bepaalde bevolkingsgroepen hebben (zie "selection bias" hierna) en die niet afdoende rekening houden met de rechten en vrijheden van de betrokkenen vallen hier niet onder. Zie Punten B, F en I hierna. Zie ook: Wetenschappelijke Raad voor het Regeringsbeleid, Synopsis van WRR-rapport,
http://www.wrr.nl/fileadmin/nl/publicaties/PDF-samenvattingen/Synopsis_R95_Big_Data_in_vrije_en_veilige_samenleving.pdf

A. Wat is Big Data?

Zoals wel vaker het geval is bij een nieuw fenomeen bestaat er geen consensus over een definitie. Dat geldt ook voor "Big Data". Definities lopen uiteen en leggen respectievelijk de nadruk op de hoeveelheid data, de verscheidenheid en complexiteit van de data, de nieuwe methoden om met de data om te gaan en op de maatschappelijke, economische en beleidsmatige mogelijkheden die ontstaan door het gebruik van Big Data.[3]

Een eenduidige definitie van Big Data is met andere woorden moeilijk te geven. Een betere manier om het ambigue fenomeen van Big Data te omschrijven, is door het aanreiken van een aantal belangrijke kenmerken. In de eerste plaats dienen we te kijken naar eigenschappen van de gebruikte data in de context van Big Data. Dit wordt veelal omschreven aan de hand van de "3 V's": *Volume*, *Variety* en *Velocity*. Bij Big Data gaat het doorgaans om grote hoeveelheden data die moeten verwerkt worden (*Volume*). Deze data zijn bovendien meestal afkomstig van verschillende bronnen en worden vaak in verschillende formaten gestructureerd, semigestructureerd of ongestructureerd (waaronder tekst, beeld en geluid) opgeslagen (*Variety*). Daarnaast gaat het bij Big Data ook om de snelheid waarmee data ingezameld en verwerkt worden: er moeten dikwijls continu, soms realtime, gegevensstromen verwerkt worden (*Velocity*). Naarmate het domein van Big Data matuurder werd, werden deze 3 V's verder aangevuld met vier andere V's waaronder *Veracity*[4], *Value*[5], *Visualisation*[6] *en Variability*[7].

Het zijn evenwel niet alleen de eigenschappen van de data die bepalen wat Big Data is, maar ook de methoden of technieken om met die data om te gaan.[8] Ietwat eenvoudig uitgedrukt kan men stellen dat de klassieke wetenschappelijke methode vertrekt van hypothesen, veelal over causale mechanismen, die vervolgens empirisch getest worden aan de hand van hiervoor zorgvuldig vergaarde data. Big Data-onderzoeksmethoden daarentegen vertrekken doorgaans van minder selectief verzamelde data en zoeken daarin naar patronen of correlaties (wat niet noodzakelijk een causaal verband impliceert)[9] zonder vooraf opgestelde hypothesen. Waar men de klassieke wetenschappelijke methode dus hypothese-gedreven zou kunnen noemen, spreekt men bij Big Data over een data-gedreven aanpak.

Tot slot kunnen we ook vaststellen dat het gebruik van Big Data nieuwe mogelijkheden biedt. Het gebruik van grotere hoeveelheden data kan immers leiden tot betere, meer gedetailleerde inzichten die op hun beurt leiden tot verbeterde besluitvorming of voorspellingen.
Samenvattend – en in navolging van de Nederlandse WRR[10] – kunnen we stellen dat we Big Data dienen te beschouwen als een samenspel of convergentie van (technologische) factoren, waaronder bovenvermelde, en niet zozeer als een vastomlijnd en definieerbaar gegeven.

---

[3] WRR-rapport 95 (2016). *Big Data in een vrije en veilige samenleving*, pg. 33.

[4] De mate van accuraatheid van de data, omdat big data analyses maar zo sterk zijn als de data waarmee ze werken. Data kunnen ook onvolledig, foutief of waardeloos zijn.

[5] De waarde ligt in de inzichten en de kennis die big data analyses met zich meebrengen.

[6] De presentatie van data in een leesbare en verstaanbare vorm, omdat de weergave van de resultaten van big data analyses een enorme invloed kan uitoefenen op de beslissingen van degenen die de resultaten aanschouwen.

[7] Data waarvan de betekenis constant wijzigt zoals bijvoorbeeld in talen waarbij woorden niet altijd eenzelfde statische betekenis behouden.

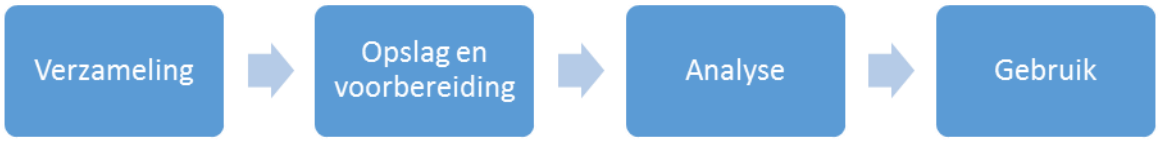
[8] Een gevolg hiervan is onder meer dat de term Big Data ook betrekking kan hebben op kleine datasets. Omgekeerd betekent dit dat niet alle grote datasets typisch Big Data zijn.

[9] Het verschil tussen correlatie en causaliteit is van cruciaal belang. Een correlatie tussen X en Y betekent dat er tussen de variatie van de grootheden X en Y een samenhang bestaat (wiskundig gekwantificeerd door een correlatiecoëfficiënt). Dat betekent echter nog niet dat wijzigingen in X ook een oorzaak zijn van wijzigingen in Y (of omgekeerd). Een correlatie impliceert dus niet automatisch causaliteit. X en Y kunnen immers ook correleren omdat ze bijvoorbeeld beiden afhangen van een derde grootheid Z. Zo zou er bijvoorbeeld een verband kunnen zijn tussen het eten van ijsjes en het aantal verdrinkingsdoden, maar dat betekent nog niet dat mensen door het eten van ijsjes verdrinken. Meer waarschijnlijk is dat het eten van ijsjes gecorreleerd is met verdrinking door een ander gegeven zoals goed, zomers weer.

[10] WRR-rapport 95 (2016). *Big Data in een vrije en veilige samenleving*, pg. 35.

B. De Big Data waardeketen ("value chain")

Op zichzelf hebben data niet zoveel waarde. Waarde ontstaat slechts doorheen een proces van verzameling, voorbereiding en opslag, analyse en gebruik. We noemen dit de Big Data waardeketen of "value chain". Hieronder gaan we verder in op de verschillende fasen van deze waardeketen.



Belangrijk echter om op te merken is dat de fasen elkaar niet steeds sequentieel opvolgen, maar dat het proces vaak in lussen verloopt. Een analyse kan er bijvoorbeeld toe leiden dat er andere databronnen worden gekozen of de voorbereiding herzien wordt. Het gaat dus veeleer om een iteratief dan een sequentieel proces. Automatisering van de verschillende fasen aan de hand van algoritmen[11] speelt hierbij in vergelijking met vroeger in toenemende mate een belangrijke rol, hoewel deskundig toezicht en begeleiding door mensen doorheen hele proces noodzakelijk blijven.

B.1. Verzameling

Door de enorme vooruitgang in de ICT worden er vandaag veel meer gegevens verzameld dan vroeger. De digitalisering heeft er immers toe geleid dat het verzamelen en opslaan van informatie zowel eenvoudiger als goedkoper is geworden. Waar men vroeger gericht onderzoek moest uitvoeren en de data manueel moest verzamelen, worden gegevens vandaag meestal automatisch digitaal geproduceerd – vaak gewoon als bijproduct van dagelijkse handelingen.

Zo wordt bijvoorbeeld elke transactie met uw bankkaart geregistreerd. Ook uw surfgedrag op het internet wordt dikwijls nauwgezet geregistreerd met behulp van cookies, super-cookies of "browser fingerprinting". Daarnaast genereren we enorme hoeveelheden data door het gebruik van zoekmachines, e-mail, sociale media, digitale televisie, mobiele telefoons en apps. Bovendien gebruiken we steeds vaker apparaten die met het internet verbonden zijn en in realtime data produceren – het zogenaamde "Internet of Things". Niet alleen online worden er echter veel data verzameld, ook offline worden er via bijvoorbeeld enquêtes of getrouwheidskaarten nog heel wat gegevens ingezameld (die later gedigitaliseerd worden). Belangrijk om hierbij op te merken is dat de data die zowel online als offline verzameld worden meestal persoonsgegevens betreffen die door hun aard en de schaal van de inzameling vaak een indringend beeld geven van de betrokkenen, zeker als verschillende bronnen worden samengevoegd.

Al deze data worden ofwel zelf verzameld door bedrijven of overheidsadministraties, ofwel verkregen (al dan niet tegen een bepaalde kost) van andere bedrijven of administraties, of van "data brokers"[12]. Data worden daarenboven vaak gedeeld, en soms zelfs vrij beschikbaar gesteld voor het publiek als "open data", hetgeen niet altijd gebeurt conform de wetgeving op de bescherming van de persoonlijke levenssfeer (zie Punt C hierna).

---

[11] Zie de definitie van algoritme in de woordenlijst

[12] Voor een analyse van de activiteiten van data brokers zie bijvoorbeeld het rapport van de FTC: *Data Brokers. A Call for Transparency and Accountability*. (Mei 2014)

Het resultaat van deze diversiteit aan databronnen is een tsunami aan data. Een compromis die we in de meeste gevallen echter moeten sluiten voor het verzamelen van meer gegevens is dat we onze hoge eisen met betrekking tot de exactheid van de gegevens ietwat moeten afzwakken. Veel van de gegevens die worden verzameld zijn immers onbewerkte "ruwe" data die bovendien veel ruis kunnen bevatten waardoor het niet triviaal is om hieruit informatie (gegevens met betekenis) of kennis te halen.

Kenmerkend voor Big Data tot slot is dat zoveel mogelijk data worden verzameld via allerlei bronnen, zonder dat men vooraf precies weet wat men hiermee gaat doen (zie Punt H hierna). Het uitgangspunt bij Big Data lijkt vaak: hoe meer data hoe beter. Pas achteraf bekijkt men hoe men hier het best informatie kan uithalen. Gegevensverzamelingen kunnen dus veelal meerdere doeleinden dienen, die bovendien pas bepaald of duidelijk worden op het moment van het gebruik van de gegevensverzamelingen. In de gegevens zit namelijk dikwijls informatie "verborgen" die op het eerste zicht niets te maken heeft met het initieel doel van de verzameling of zelfs met de betekenis van de variabelen die verzameld worden (bijv. het afleiden van de kans op overlijden uit kredietkaartgeschiedenis, seksuele geaardheid uit connecties op Facebook, de kans op echtscheiding uit activiteit op sociale netwerken, de kans op misdadig gedrag aan de hand van tweets enz.). Dat is een wezenlijk verschil met vroeger waarin gegevens meer verzameld werden in functie van één duidelijk afgelijnd, vooraf bepaald doel, alhoewel secundair gebruik (latere verwerking) uiteraard ook al mogelijk was. Dit houdt rechtstreeks verband met het feit dat het verzamelen van gegevens vroeger een complexe en dure aangelegenheid was.

B.2. Opslag en voorbereiding

De tweede stap in de waardeketen is de opslag van de data. Parallel met de mogelijkheden om data te verzamelen, is ook de opslagcapaciteit toegenomen terwijl de kosten afnamen. Toch blijft de opslag van grote hoeveelheden data een uitdaging daar de productie van data sneller toeneemt dan de opslagcapaciteit.[13] De oplossing om met dit probleem om te gaan is gedistribueerde opslag: data die inhoudelijk of formeel bij elkaar horen op verschillende plaatsen bewaren. Vaak komt dit neer op een vorm van opslag van data in de cloud. Hierbij vormt beveiliging van de data een extra aandachtspunt.[14]

Het feit dat gegevens gedistribueerd worden opgeslagen, evenals de variëteit van de gegevens vereisen tevens een andere manier om met de gegevens om te gaan. Traditionele, zogeheten 'relationele' databases zijn ontworpen voor een wereld waarin gegevens schaars, gestructureerd, vooraf gedefinieerd en exact zijn. Deze benadering staat echter steeds meer haaks op de werkelijkheid met steeds grotere hoeveelheden gegevens van wisselend type en wisselende kwaliteit, verspreid over verschillende harde schijven en computers. Die nieuwe realiteit heeft geleid tot nieuwe databaseontwerpen waarin gegevens gedistribueerd, meestal (semi)ongestructureerd worden opgeslagen en geraadpleegd worden met behulp van nieuwe, speciaal daarvoor aangepaste mechanismen (NoSQL ⇔ SQL15).

Een populair softwareframework in dit verband is Apache Hadoop, een open-source-implementatievan het MapReduce framework dat grootschalige parallelle en gedistribueerde verwerking van data mogelijk maakt.

---

[13] International Data Corporation,The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, april 2014, gepubliceerd op https://www.emc.com/leadership/digital-universe/2014iview/index.htm.

[14] Zie in dit verband ook het advies nr. 10/2016 van de Commissie aangaande de gebruikmaking van cloud computing door de verantwoordelijke voor de verwerking.

[15] SQL staat voor Structured Query Language. Het is een gestandaardiseerde taal voor het bevragen en aanpassen van relationele databases, waarin gegevens worden opgeslagen in tabelstructuur. NoSQL (Not Only SQL) databases daarentegen bieden een mechanisme om data te benaderen die niet noodzakelijk op een traditionele, relationele manier werd opgeslagen.

Nadat de data werden opgeslagen, zijn deze evenwel nog niet meteen klaar voor analyse. Hiervoor moeten de data eerst nog worden voorbereid. Dat begint bij het samenvoegen van verschillende databronnen tot een zogenaamd "data warehouse". Van zodra zo'n "data warehouse" gevormd is, moeten de data erin opgeschoond en gefilterd worden en eventueel een aantal andere "pre-processing" stappen doorlopen (zoals bijvoorbeeld normalisatie of transformatie). Zo worden afhankelijk van de specifieke toepassing bijvoorbeeld niet-essentiële karakters verwijderd, waarden zoals datums en telefoonnummers gestandaardiseerd, inconsistenties in de data waar mogelijk hersteld, en duplicaten en onbetrouwbare data verwijderd.

In deze voorbereidingsfase worden ook privacybevorderende technieken toegepast om de data waar nodig en in de mate van het mogelijke zo min mogelijk herleidbaar te maken tot identificeerbare individuen (zie ook Punt C hierna). Het gaat om technieken zoals pseudonymisation[16], noise addition, substitution, aggregation of K-anonymity, L-diversity, differential privacy en hashing/tokenization.[17] Toch moet het nodige voorbehoud worden geformuleerd bij de effectiviteit van deze technieken. Niet al deze technieken leiden tot gegevens die geen persoonsgegevens meer zijn, d.w.z. anonieme gegevens. Onderzoek toonde immers aan dat het soms relatief makkelijk is om uit schijnbaar anonieme data individuen alsnog uniek te identificeren mits gebruik te maken van kennis betreffende individuen, die diegene die de heridentificatie wenst te doen soms reeds bezit.[18] Het is belangrijk om de effectiviteit van de gebruikte technieken na te gaan om vast te stellen of er echt sprake kan zijn van anonieme gegevens die buiten het toepassingsgebied van de Verordening vallen.

B.3. Analyse

De derde stap in de waardeketen is de analysefase. Men spreekt hier soms ook over "Knowledge Discovery in Databases" (KDD) of data mining[19]. Naast KDD en data mining zijn er nog heel wat andere termen die geassocieerd worden met Big Data-analyse zoals profiling, clustering, text mining, machine learning, (social) network analysis, predictive analysis, natural language processing en visualization. Algoritmes automatiseren beslissingen in (semi-)geautomatiseerde verwerkingen, zoals bij het stellen van prioriteiten (zoekmachines, risicobeoordeling), het klasseren van gegevens (en personen), het maken van associaties (correlatie of causaliteit), filtering (weglaten van informatie)[20]. Toch zou het – zeker in gevallen waar er belangrijke gevolgen kunnen zijn voor individuen - technisch en juridisch gezien de opzet moeten zijn dat een menselijke factor een grote rol van betekenis blijft spelen (zie hierna onder Punt N). Mensen blijven technisch en maatschappelijk gesproken noodzakelijk om het hele proces te ontwerpen, begeleiden en te sturen en om de resultaten van de analyse te interpreteren (wat in de meeste gevallen zelfs redelijk arbeidsintensief is). Data mining blijft een ambacht. Tover- of kant-en-klare druk-op-de-knop oplossingen zijn er hier meestal niet, in tegenstelling tot wat sommige softwareleveranciers van dataminingproducten ons willen laten geloven.

---

[16] Zie de woordenlijst

[17] Voor een overzicht, evenals de sterktes en zwaktes van de technieken: zie Opinion 05/2014 on Anonymisation Techniques van de WP29 (WP216).

[18] Zie bijvoorbeeld de Montjoye et al (2013). Unique in the Crowd: The privacy bounds of human mobility*. Scientific Reports, 3:1376.* doi:10.1038/srep01376; de Montjoye Y. A., Unique in the shopping mall: On the re-identifiability of credit card metadata, Science 347, 30 January 2015; http://science.sciencemag.org/content/347/6221/536, 537.

[19] Zie de woordenlijst

[20] DIAKOPOULOS, N., Accountability in Algorithmic Decision Making, Februari 2016, Communications ACM, http://www.nickdiakopoulos.com/wp-content/uploads/2016/03/Accountability-in-algorithmic-decision-making-Final.pdf

Big Data-analyses hebben als doel om patronen of kenmerken in datasets te ontdekken of voorspellen. Deze patronen drukken evenwel louter correlaties uit, niet noodzakelijk causale verbanden (zie Punt D hierna). Slechts in tweede instantie kunnen deze correlaties verder onderzocht worden op hun validiteit/reproduceerbaarheid en/of mogelijke onderliggende causale mechanismen. Bovendien zeggen correlaties ook slechts iets over statistische verbanden die (mogelijks) opgaan op populatieniveau maar het is perfect mogelijk dat voor individuele gevallen deze algemene verbanden niet tot uiting komen.

Bij analyses kan men een onderscheid maken tussen "supervised learning" en "unsupervised learning".

Bij "supervised learning" vertrekken de analyses van een model dat getraind dient te worden (zie ook Figuur 1 in Punt G). Het model wordt getraind aan de hand van een "trainingset" van data. Deze "trainingset" bestaat uit de waarden van de verzamelde variabelen over een bepaald object (bijv. medische parameters van een patiënt) en het geanticipeerd resultaat of de status, weergegeven door middel van o.a. klasse labels, die men wenst te voorspellen aan de hand van deze variabelen (bijv. de aan- of afwezigheid van een medische diagnose). Om een model te bouwen moet men met andere woorden eerst beschikken over een groep objecten waarbij het resultaat dat men wenst te voorspellen reeds gekend is. Belangrijk is ook dat de performantie van het model (cfr. vals positieven en vals negatieven) wordt nagegaan met behulp van onafhankelijke en representatieve testdata. Van zodra het model getraind en getest is, kan men dit model gebruiken om het resultaat te voorspellen van andere objecten, waarbij het resultaat nog niet gekend is.

Indien men niet beschikt over een "trainingset" (dus als er geen groep objecten beschikbaar is met gekende klasse labels) dan kan men gebruik maken van "unsupervised learning". Hierbij gaat men op zoek naar ongekende structuur in de data. Een voorbeeld van een dergelijke techniek is clustering waarbij algoritmen op zoek gaan naar ongekende groepen, clusters of klassen in de data.

In deze context is het daarnaast ook nuttig om een onderscheid te maken tussen black-box en white-box modellen. Er bestaan verschillende interpretaties van deze termen. Over het algemeen stelt men dat de voorspellingen gemaakt door black-box modellen moeilijk intuïtief kunnen worden uitgelegd en/of bestaan uit een wiskundige relatie tussen input en output die geen verband heeft met de werkelijke wetmatigheden die de relatie tussen in- en output bepalen. Bij white-box modellen is de basis waarop voorspellingen gebeuren wel interpreteerbaar en/of gebaseerd op de wetmatigheden die het proces determineren dat wordt beschreven door het model.

B.4. Gebruik

De vierde en laatste stap in de waardeketen heeft betrekking op het gebruik van de resultaten van de analyse.
Op basis van de modellen, patronen of correlaties die uit de data gedistilleerd werden, kan men hypothesen vooropstellen en/of gedragingen, kenmerken en/of gebeurtenissen met een al dan niet grote nauwkeurigheid voorspellen. Door het gebruik van grotere datasets neemt ook de statistische kracht ("power") toe en zal men vlugger (zelfs reeds bij subtiele of minimale correlaties of verbanden) statistisch significante resultaten bekomen (doch statistisch significant zal in deze context dus niet steeds relevant zijn omdat ook minimale en betekenisloze verbanden worden weerhouden).

De resultaten van dataminingtechnieken op Big Data zijn dikwijls maar hypothese-genererend

en moeten bijna altijd verder geverifieerd of gevalideerd worden (zowel op algemeen of modelniveau als op individueel niveau). Ze zijn dus feitelijk maar een tussenstap. Automatische beslissingen moeten dus met uiterste omzichtigheid gebeuren, in het bijzonder bij belangrijke beslissingen die een impact hebben op de betrokkene zoals met betrekking tot gezondheid, kredietverlening of werkgelegenheid (zie Punt N hierna). Bij "supervised learning" bijvoorbeeld moet men ervan uitgaan dat er altijd een deel van de voorspellingen fout is. Er komen met andere woorden misclassificaties voor (ook vals positieven en vals negatieven genoemd), dit zijn objecten die aan de verkeerde groep worden toegekend door het wiskundig model. Een ander voorbeeld is dat het bij Big Data mogelijk is om zeer veel mogelijke correlaties of verbanden tegelijk te onderzoeken. In de statistiek noemt men dit "multiple testing" en dit heeft als gevolg dat men soms verbanden ziet die er niet zijn, gewoon per toeval. "Overfitting" van het model is een ander voorbeeld waar men verbanden modelleert die er niet zijn waardoor de performantie van het model op onafhankelijke testdata verslechtert.

Niettegenstaande bovenstaande kritische bemerkingen, kunnen meer accurate beslissingen en voorspellingen door middel van datamining op hun beurt potentieel leiden tot onder meer het verminderen van de menselijke foutenlast, grotere operationele efficiëntie, verminderde kosten en risico's, nieuwe producten, maatschappelijke meerwaarde (bv. in de gezondheidszorg) en geoptimaliseerde aanbiedingen. Dit verklaart de grote interesse in de toepassingen en mogelijkheden van Big Data bij zowel bedrijven als overheden.

Voorbeelden:

- Google Flu Trends: Google voorspelt griepepidemieën op basis van miljarden zoekopdrachten.
- Amazon geeft suggesties voor boeken op basis van Big Data-algoritmen gebruik makend van het eigen aankoopgedrag en dat van anderen.
- Overheden kunnen Big Data gebruiken om diverse vormen van fraude op te sporen[21].
- Antiterrorisme (cfr. PRISM-programma van de NSA) en predictive policing[22] (het i-Politie project van de Federale Politie[23])
- Gerichte advertenties ("Online Behavioural Advertising"): analyse van surf-, klik en kijkgedrag[24] levert veel geld op omdat adverteerders bereid zijn om meer te betalen voor tonen van gerichte advertenties.
- Medische wetenschappen: voorspellen van therapeutisch effect en bijwerkingen van medicijnen en andere therapeutische interventies, ondersteuning van artsen bij keuze behandeling (clinical decision support), gerichtere en meer nauwkeurige diagnose, personalised medicine ...
- Bio-informatica: nieuwe technieken in de moleculaire biologie/genetica laten toe om zeer grote hoeveelheiden data te genereren. Het is nu bijvoorbeeld mogelijk om het volledig genoom (next generation sequencing) van een individu voor een betaalbare prijs te bepalen. Het is bijvoorbeeld ook al geruime tijd mogelijk om het transcriptoom (concentratie van alle RNA-moleculen in een cel) van bijv. tumorcellen te bepalen. Deze

[21] Zie Regeerakkoord 9 oktober 2014, http://www.premier.be/sites/default/files/articles/Accord_de_Gouvernement_-_Regeerakkoord.pdf; TOMMELEIN, B., Beleidsverklaring van de staatssecretaris voor bestrijding van de sociale fraude, privacy en Noordzee van 13 november 2014, Kamer, DOC 54, 0020/004, gepubliceerd op http://www.dekamer.be/FLWB/PDF/54/0020/54K0020004.pdf

[22] Zie bijvoorbeeld de verwijzingen naar de toepassingen bij de Los Angeles Police Department op pagina 20 van Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Mei 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

[23] MEEUS, R., Longread: Hoe iPolice de natie veiliger maakt, 24 juni 2016, http://datanews.knack.be/ict/nieuws/longread-hoe-ipolice-de-natie-veiliger-maakt/article-longread-720899.html; Ponciau, L, Les détails du projet iPolice dévoilés, Le Soir, 17 september 2016.

[24] Proximus TV gaat reclame aanpassen aan uw kijkgedrag, Express Business, 1 augustus 2016; Proximus ouvre la voie à la publicité personnalisée, le Soir, 2 augustus 2016, Telenet begint met tv-reclame op maat, De Tijd, 13 september 2016.

vorm van big data gecombineerd met dataminingtechnieken laat gepersonaliseerde therapeutische schema's toe waarbij de behandeling veel nauwer aansluit bij het genetisch profiel van de patiënt (of tumor).

Deel 2

Dataprotectiebeginselen en aanbevelingen met betrekking tot de Big Data analyses

A. Algemene opmerkingen – methodologie

De wetgeving op de bescherming van persoonsgegevens is van toepassing van zodra er sprake is van een verwerking van persoonsgegevens[25]. Het volstaat hierbij dat personen kunnen worden afgezonderd in datasets (zie verwijzing naar "single out" in Punt C hierna) waarbij er een redelijke kans bestaat dat een aspect van hun identiteit zou kunnen worden achterhaald (dit kan ook een digitale identiteit zijn, zoals bv. een mailadres, waardoor de betrokkene benaderbaar wordt) om te kunnen spreken van een verwerking van persoonsgegevens[26], zonder dat het nodig is om noodzakelijk hun naam of burgerlijke identiteit te kennen.

In het kader van de plicht om de doelstelling van de verwerking uitdrukkelijk te omschrijven (artikel 5.1.b) AVG), en de risico-gebaseerde aanpak ("**risk based approach**")[27] van de AVG (waarbij de impact op de rechten en vrijheden van de betrokkenen moet worden nagegaan), is het van belang steeds na te gaan welk effect het gebruik van big data analyses in de praktijk kan hebben ten aanzien van de betrokkenen.

Er is een onderscheid tussen de aard van de data-analyses en het gebruik ervan. Qua kan gebruik kan men een onderscheid maken tussen het voorspellend[28], beschrijvend[29] of voorschrijvend[30] gebruik van big data analyses[31]. Dergelijk gebruik kan een invloed hebben op de wijze van behandeling en rechten en vrijheden van de betrokkenen (zie hierna onder punt E.2.).

De uitgangspunten van veel big data projecten staan in een gespannen verhouding tot de privacy- en dataprotectiebeginselen[32], en de schaalbaarheid[33] en risico-gebaseerde aanpak onder de AVG (zie hierna).

Er bestaat een duidelijke behoefte aan meer **(rechts)zekerheid** over de vraag of en hoe de

---

[25] Zie de definities in artikel 1 § 1 en § 2 Wet 8 december 1992 en artikel 4 1) en 2) AVG.

[26] Zie overweging 26 AVG waarvan de Engelse versie verwijs naar "singling out", en de Opinie 08/2012 WP 199 van de Groep 29, gepubliceerd op http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp199_en.pdf. Volgens het Hof van Justitie in de zaak C-582/14 van 19 oktober 2016 Patrick Breyer / Bondsrepubliek Duitsland vormt het dynamische internetprotocoladres van een bezoeker voor de exploitant van de site een persoonsgegeven wanneer hij beschikt over wettige middelen waarmee hij de betrokken bezoeker kan identificeren aan de hand van extra informatie die bij diens internetprovider berust.

[27] Zie WP 218, Verklaring van de groep 29 van 30 mei 2014 over de rol van een risico-gebaseerde aanpak in juridische kaders voor gegevensbescherming, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf

[28] Zoals hierna blijkt vertonen vooral de voorspellende en voorschrijvende analyses bepaalde privacygevoelige kenmerken een risico op oneerlijke verwerking door discriminatie

[29] Hierna wordt onder beschrijvende modellen verstaan: het in kaart brengen van correlaties of patronen in digitale sporen. Bij voorspellende modellen tracht men – door gebruik te maken van deze correlaties - outputvariabele(n) (bv. de kans op fraude) te voorspellen aan de hand van inputvariabelen.

[30] Gebruik van mathematische modellen om het (aankoop of normconform) gedrag te sturen (bv. ontraden van offline aankopen door een goedkoper digitaal alternatief aan te bieden dat meer digitale sporen nalaat, zoals onder klantengetrouwheidsprogramma's van grootwarenhuizen, of beleidsmaatregelen die het anoniem aankopen van prepaidkaarten onmogelijk maken).

[31] Zie pagina 21 van MOEREL, Lokke en PRINS, Corien, Privacy for the homo digitalis. Proposal for a new regulatory framework for data protection in the light of Big Data and the Internet of Things, 25 mei 2016, beschikbaar op http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784123, en de definities hierna en VAN DER SLOOT, B. en VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p 40, http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf

[32] International Working Group on Data Protection in Telecommunications, Working paper on Big Data and Privacy. Privacy principles under pressure in the age of Big Data analytics, 5-6 mei 2014, gepubliceerd op http://www.datenschutz-berlin.de/attachments/1052/WP_Big_Data_final_clean_675.48.12.pdf

[33] Deze "schaalbaarheid" betekent dat men voor verwerkingen met een hoog risico meer maatregelen zal dienen te nemen om de vereisten van gegevensbescherming na te leven dan ten opzichte van verwerkingen met een laag risico. Het idee illustreert een overgang van de nadruk op het verzamelen van persoonsgegevens naar het gebruik van persoonsgegevens, zoals in discussies rond big data. Zie WP 218, Verklaring van de groep 29 van 30 mei 2014 over de rol van een risico-gebaseerde aanpak in juridische kaders voor gegevensbescherming, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf

huidige en toekomstige[34] privacy-en dataprotectieregels dienen te worden toegepast.

De Commissie beschouwt het als een uitdaging om de nieuwe aanpak onder de AVG en de klassieke beginselen voor de bescherming van privacy en persoonsgegevens een nuttig effect te laten hebben in de context van big data analyses. Hierbij kan het niet de bedoeling zijn om het gebruik van deze nieuwe informatietechnologieën onnodig af te remmen, gelet op hun dikwijls aantoonbare nut voor de maatschappij.

Zoals hiervoor vermeld in deel I moet de **regelgeving algemeen, robuust en technologieneutraal** genoeg zijn (en dat is ze voor een groot deel ook) zodat bij iedere verwerking de privacy van de betrokkenen voldoende gegarandeerd wordt en een **goede balans** gevonden wordt om legitieme opportuniteiten niet teveel te fnuiken, doordat de voorwaarden worden geboden waaronder legitiem gebruik mogelijk is.

Om deze uitdaging aan te gaan formuleert de Commissie hierna **concrete aanbevelingen**, die kunnen worden gebruikt om de mate van naleving van de privacy- en dataprotectieregels bij big data analyses te toetsen en te bevorderen. Om tegemoet te komen aan de bekommernis van een aantal stakeholders om een "**level playing field**" na te streven t.o.v. andere Europese landen werd. Er werd een **publieke consultatie gehouden in april 2017** aangaande deze aanbevelingen. Er werd ook naar gestreefd om deze te schrijven **vanuit de toepassing van de AVG**. Deze aanbevelingen gelden onverminderd de standpunten van het Europees Comité voor de gegevensbescherming.

Bij de toepassing van de regels tot bescherming van de privacy en persoonsgegevens is het van belang om een inzicht te hebben in de verschillende fasen[35] die bij big data projecten kunnen worden onderscheiden. Het maakt immers een wezenlijk verschil dat de privacy- en dataprotectiebeginselen worden toegepast op alle hiervoor vermelde fasen in plaats van op een enkel onderdeel van een big data project

B. AVG

De Commissie vestigt er de aandacht op dat er recent nieuwe Europese regelgeving inzake de bescherming persoonsgegevens werd uitgevaardigd: de algemene Verordening betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en de Richtlijn voor Politie en Justitie[36]. Deze teksten verschenen in het Europese Publicatieblad van 4 mei 2016[37].

De verordening, meestal AVG (Algemene Verordening Gegevensbescherming) genaamd, is van kracht geworden twintig dagen na publicatie, nl. op 24 mei 2016 en wordt, twee jaar later, automatisch van toepassing: 25 mei 2018. De richtlijn voor politie en justitie moet via nationale wetgeving omgezet worden tegen uiterlijk 6 mei 2018.

---

[34] Zie verwijzing naar de AVG in Punt B hierna

[35] Zie de verwijzing in deel I hiervoor naar "verzameling", "opslag en voorbereiding", "analyse" en "gebruik".

[36] Richtlijn 2016/680 van het Europees Parlement en de Raad van 27 april 2016 *betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens door bevoegde autoriteiten met het oog op de voorkoming, het onderzoek, de opsporing en de vervolging van strafbare feiten of de tenuitvoerlegging van straffen, en betreffende het vrije verkeer van die gegevens en tot intrekking van Kaderbesluit 2008/977/JBZ van de Raad*, PB L 119, 4 mei 2016, p. 89–131

[37] Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (algemene verordening gegevensbescherming)

Richtlijn (EU) 2016/680 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens door bevoegde autoriteiten met het oog op de voorkoming, het onderzoek, de opsporing en de vervolging van strafbare feiten of de tenuitvoerlegging van straffen, en betreffende het vrije verkeer van die gegevens en tot intrekking van Kaderbesluit 2008/977/JBZ van de Raad

http://eur-lex.europa.eu/legal-content/NL/TXT/?uri=OJ:L:2016:119:TOC

http://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ%3AL%3A2016%3A119%3ATOC

Voor de AVG betekent dit dat vanaf 24 mei 2016, en gedurende de termijn van twee jaar voor de tenuitvoerlegging, op de lidstaten enerzijds een positieve verplichting rust om alle nodige uitvoeringsbepalingen te nemen en anderzijds ook een negatieve verplichting, de zogenaamde "onthoudingsplicht". Laatstgenoemde plicht houdt in dat er geen nationale wetgeving mag worden uitgevaardigd die het door de Verordening beoogde resultaat ernstig in gevaar zou brengen. Ook voor de Richtlijn gelden gelijkaardige principes.

Het verdient dan ook aanbeveling om desgevallend nu reeds op deze teksten te anticiperen. De Commissie heeft in onderhavig rapport, in de mate van het mogelijke en onder voorbehoud van mogelijke bijkomende toekomstige standpunten, alvast gewaakt over de hoger geschetste negatieve verplichting.

Hoewel de AVG geen definitie bevat van "big data" als dusdanig, voert de AVG tal van nieuwe elementen in die relevant lijken in de context van big data analyses. Zonder volledig te willen zijn, wijst de Commissie op de definitie van profilering in de AVG[38], die in navolging van een eerdere aanbeveling van het Comité van Ministers van de Raad van Europa[39], de nadruk legt op het analytisch en voorspellend karakter van deze techniek. De AVG hanteert een "risk based approach" (gekoppeld aan de gegevensbeschermingseffectbeoordeling[40] (zie hierna onder Punten E en S). De AVG bevat ook nieuwe plichten en beginselen zoals de verantwoordingsplicht ("accountability"[41]) (zie hierna onder Punt S), de meldingsplichten[42] en de raadplegingsplicht voor residuele risico's[43], het aanleggen van interne documentatie[44], de beginselen van gegevensbescherming door ontwerp ("privacy by design")[45] en gegevensbescherming door standaardinstellingen ("privacy by default")[46], en de rol van gedragscodes en certificering[47] (gedragscodes en certificering kunnen zeker een belangrijke rol spelen bij het beschermen van persoonsgegevens in de big data context - deze concepten zullen nog verder worden toegelicht in bijkomende opinies van de toezichthouders op Europees of nationaal vlak).

Kunnen daarnaast nog worden vermeld : transparantie (als algemeen beginsel)[48], de voorwaarden voor toestemming[49], de regeling van de gevoelige persoonsgegevens[50], en het recht op informatie[51] (bv. bestaan van geautomatiseerde besluitvorming)

Een van de nieuwe elementen is ook dat de AVG de evaluatie van persoonlijke aspecten waaraan rechtsgevolgen zijn gekoppeld voor de betrokkene of die betrokkene op vergelijkbare

---

[38] Artikel 4, 4) AVG bevat de volgende definitie van profilering : *"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".*

[39] *Zie ook de definitie in* artikel 1. *e. van de* Aanbeveling CM/Rec(2010)13 van 23 november 2010 van de Raad van Ministers over de bescherming van personen ten opzichte van de geautomatiseerde verwerking van persoonsgegevens in de context van profilering, gepubliceerd op
https://wcd.coe.int/wcd/ViewDoc.jsp?Ref=CM/Rec(2010)13&Language=lanEnglish&Ver=original&Site=CM&BackColorInternet=DBDCF2&BackColorIntranet=FDC864&BackColorLogged=FDC864

[40] Artikel 35 AVG. Het verrichten van een gegevensbeschermingseffectbeoordeling zal (onder meer) nodig zijn als een big data project systematisch en uitgebreid persoonlijke aspecten van natuurlijke personen beoordeelt, en wordt gebruikt voor nemen van beslissingen ten aanzien van natuurlijke personen, waaraan rechtsgevolgen zijn verbonden of die natuurlijke personen op vergelijkbare wijze wezenlijk treffen.

[41] Artikel 5.2 AVG

[42] Artikelen 33 en 34 AVG

[43] Artikel 36.1 AVG

[44] Artikel 30 AVG (register van de verwerkingsactiviteiten), artikel 33.5 AVG (voor alle inbreuken in verband met persoonsgegevens)

[45] Artikel 25 AVG

[46] Artikel 25 AVG

[47] Artikel 40 AVG

[48] Overwegingen 38,58 en 100, artikel 5 1. a), en 88 2. AVG

[49] Zie onder meer artikelen 4 11), 7 en 8 AVG, en de overwegingen 32, 33, 38, 40, 42, 43 en 50 en

[50] Zie de artikelen 9 en 10 AVG, en de overwegingen nrs. 10, 34, 35, 54-54, 71, 80, 91 en 97.

[51] Zie de artikelen 12 tot en met 15, 19, 21.4, AVG en de overwegingen nrs. 58, 60-63, 71, 156.

wijze aanmerkelijk treffen[52], beschouwt als een **risico** voor de rechten en vrijheden van de betrokken natuurlijke personen. Dit impliceert dat de verwerkingsverantwoordelijke op objectieve wijze [53] een **continue risicobewaking** [54] dient te verrichten, waarvan de risiciobeoordelingsmethode **periodiek dient te worden geherevalueerd**. Er zal (onder meer) [55] sprake zijn van een **hoog risico** indien er sprake is van een systematische en uitgebreide beoordeling van persoonlijke aspecten[56].

Indien bijvoorbeeld een ipolicing model de reacties van burgers gaat analyseren op sociale media om opdrachten van bestuurlijke of gerechtelijke politie uit te voeren, zal de effectbeoordeling de kans en impact op de vrijheid van meningsuiting op sociale media van de betrokkenen en de gelijke behandeling (verbod op discriminatie) door de politie moeten worden ingecalculeerd, rekening houdende met de reeds bestaande controlemaatregelen op de politiediensten. Naargelang de toepassingen zullen ook de impact voor het recht op eigendom[57], de persoonlijke vrijheid (van toepassing bij predictive policing[58]), moeten worden afgewogen. Anderzijds is het ook zo dat er voor de toepassing van de risico-gebaseerde aanpak onder de AVG[59] een grote behoefte bestaat om een aanvaardbare methode uit te werken voor het toepassen van risicoafweging op big data analyses.

De toegepaste risicobeoordeling zal een methode dienen te hanteren die een aantal basiselementen zal bevatten die de Commissie verder zal toelichten ter gelegenheid van een aparte aanbeveling Het begrip "risico" kan op verschillende wijzen worden geïnterpreteerd. In de literatuur omschrijft men het begrip risico doorgaans als de kans dat een bepaalde bedreiging zich voordoet, met een welbepaalde impact ("ernst") tot gevolg.[60]

Bij continue risicoafweging zal er ook regelmatig een evaluatiemoment moeten worden voorzien, waarbij het risicoafwegingsmodel wordt geherevalueerd. Bij het beoordelen van de risico's verbonden aan big data analyses (zie hiervoor) **onderscheidt en beschouwt men best alle verschillende fases (zie hiervoor)** [61]. Het enkel toelichten van een enkele bewerking (zoals het verzamelen, ordenen,… van gegevens[62]) en/of risico gekoppeld aan een bewerking (bv. codering van gegevens) kan immers een onvolledig beeld geven op de risico's bij de toepassing op natuurlijke personen.

De Commissie stelt vast dat er in het professionele ("B2B") verkeer een groeiende promotie en

---

[52] Overweging 71 AVG.

[53] Overweging 76 AVG bevestigt het objectieve karakter van deze beoordeling van het risico ten opzichte van de verwerking en de gevolgen hiervan voor de rechten en vrijheden van personen: "*De waarschijnlijkheid en de ernst van het risico voor de rechten en vrijheden van de betrokkene moeten worden bepaald onder verwijzing naar de aard, het toepassingsgebied, de context en de doeleinden van de verwerking. Het risico moet worden bepaald op basis van een objectieve beoordeling en vastgesteld moet worden of de verwerking gepaard gaat met een risico of een hoog risico.*"

[54] Zie de risicoklasse vermeld in overweging 75 van de AVG en ondermeer de Artikelen 32.1, 33, 34 en 35 AVG

[55] Voor een lijst van risico's die als hoog moeten worden bestempeld : zie Groep 29, WP 248, - Guidelines van 4 April 2017 on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, gepubliceerd op http://ec.europa.eu/newsroom/document.cfm?doc_id=44137; en bijlage 2 bij de CBPL (ontwerp) aanbeveling mbt de DPIA, gepubliceerd op https://www.privacycommission.be/sites/privacycommission/files/documents/CO-AR-2016-004_NL.pdf

[56] Overweging 91 AVG

[57] Zie het werkdocument n° 112 van de Nationale Bank van 2011 dat een verband aantoonde tussen betalingsachterstanden inzake mobiele telefonie en betalingsachterstanden bij krediet, https://www.nbb.be/en/articles/working-paper-ndeg-212. Indien de centrale voor kredieten aan particulieren wordt herzien om te werken op basis van dergelijke correlaties, kan dit een impact hebben op de mogelijkheid om eigendom te verwerven.

[58] Zie hiervoor en het lexicon. Een voorbeeld vormt de toepassing bij de Los Angeles Police Department beschreven op pagina 20 van de Executive Office of the President: Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Mei 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[59] Zie de aankondiging van ENISA van januari 2016 mbt (onder meer) het aspect van risicobeoordeling van de AVG op dit vlak https://www.enisa.europa.eu/publications/enisa-position-papers-and-opinions/enisa2019s-position-on-the-general-data-protection-regulation-AVG/

[60] Zie bijv. I. Naumann (ed.), "Privacy and Security Risks when Authenticating on the Internet with European eID Cards", ENISA, 26 November 2009.

[61] Zie deel I hiervoor : verzameling, opslag en voorbereiding, analyse en gebruik

[62] In de zin van artikel 1 § 2 WVP.

verkoop is van allerlei diensten en oplossingen in verband met big data. Vaak gaat deze promotie gepaard met het beloven van mogelijks onrealistische verwachtingen[63] naar de verantwoordelijken toe, bijvoorbeeld inzake het besparen van uitgaven bij de aanpak van fraudebestrijding. Men vergeet hierbij dat het potentieel van big data pas kan worden gerealiseerd als op zijn minst voldaan is aan een aantal randvoorwaarden[64], waaronder de beschikbaarheid en kwaliteit van data die de gewenste informatie bevat (zie hierna), het bestaan en correct toepassen van expertise in de organisatie van de verantwoordelijke, en het voorhanden zijn van een goed regelgevend kader dat tegelijk efficiënte big data analyses mogelijk maakt en de rechten en vrijheden van personen vrijwaart.

➢ Aanbeveling 1

Een verantwoordelijke die wenst in te gaan op een aanbod van derden om big data technologie te testen of toe te passen, kan steeds vragen naar de informatie of documentatie die de impact beschrijft van de technologie op de persoonlijke levenssfeer van natuurlijke personen ("gegevensbeschermingseffectbeoordeling" van het product of de dienst)[65] en/of een relevant gegevensbeschermingszegel[66]. Deze informatie of documentatie mag geen statische beoordeling zijn, noch een attest die de AVG compliance bevestigt van het gebruik van de technologie.

Gelet op de beginselen van gegevensbescherming door ontwerp ("privacy by design") en gegevensbescherming door standaardinstellingen ("privacy by default") en de verantwoordingsplicht ("accountability") bevat de AVG een plicht voor verwerkingsverantwoordelijken om te kunnen aantonen dat de big data technologie die zij gebruiken conform de AVG is.

➢ Aanbeveling 2

Elke promotie voor het inzetten van big data technologie voor specifieke toepassingen[67] zou de potentiële professionele ("B2B") klanten (organisaties die deze technologie willen gebruiken) ook duidelijk (bv. via een relevant gegevensbeschermingszegel of in de productfiche)[68] moeten informeren of en hoe de specieke inzet van deze technologie de Europese reglementering inzake privacy en gegevensbescherming respecteert.

Het beoordelen en aantonen[69] van het **al dan niet handelen in overeenstemming met de AVG ("AVG compliance") en/of van de verantwoordelijkheid en aansprakelijkheid bij big data** lijkt op het eerste zicht niet evident. De voorliggende aanbeveling probeert alvast

---

[63] Zie de verwijzing in deel I naar big data als een ambacht

[64] Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, p 84 punt 4.4.

[65] Hoewel er om pro forma ("check the box") oefeningen te vermijden geen vaste vorm noch inhoud van deze beoordelingen kan worden opgelegd, kan wel worden verwezen naar een aantal minimale kenmerken waaraan deze effectbeoordeling dient te beantwoorden. Zie bijlage 1 bij het ontwerp van aanbeveling uit eigen beweging met betrekking tot de gegevensbeschermingseffectbeoordeling, 20 december 2016, https://www.privacycommission.be/sites/privacycommission/files/documents/CO-AR-2016-004_NL.pdf. Zie ook de voorbeelden en criteria in de bijlage van Groep 29, WP 248, - Guideliness van 4 April 2017 on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, gepubliceerd op http://ec.europa.eu/newsroom/document.cfm?doc_id=44137.

[66] Overweging 100 en Artikelen 28.5 en 42 AVG

[67] Sommige big data technologie is generisch en kan voor verschillende toepassingen worden ingezet waardoor het moeilijk is om iets te zeggen over de compliance met de AVG.

[68] Een klassieke privacyverklaring in de algemene productvoorwaarden zal hierbij niet volstaan.

[69] Overweging 90 en artikel 24.1 AVG.

een aanzet te geven om hierbij te helpen. Ook de AVG bevat uiteraard ook elementen die hierbij relevant kunnen zijn zoals het al dan niet rekening houden met het risico van bepaalde "big data" toepassingen voor de rechten en vrijheden van natuurlijke personen[70], het verschaffen van ("B2C") transparantie (aan de betrokkene)[71], het al dan niet gebruik van relevante gegevensbeschermingszegels of afsluiten van gedragscodes[72].

➢ **Aanbeveling 3**

Via de gegevensbeschermingseffectbeoordeling kunnen de verantwoordelijken aangeven in welke mate rekening werd gehouden met de AVG en eventueel de voorliggende aanbevelingen van de Commissie.

➢ **Aanbeveling 4**

Bij gebruik van big data analyses die een hoog risico[73] vormen voor de rechten en vrijheden van de betrokkenen, kan worden aanbevolen om de betrokkenen en diverse stakeholders[74] uit het maatschappelijk veld te betrekken via bijvoorbeeld feedbackmechanismen zoals enquêtes, en desgevallend een gedragscode af te sluiten om verantwoord gebruik van deze verwerkingen aan te moedigen.

C. Mogelijke toepasselijkheid van dataprotectiewetgeving bij toepassing van pseudonimiserings- of anonimiseringstechnieken, al dan niet in combinatie met de aggregatie van gegevens

Een vaak voorkomend misverstand is dat het werken met gegevens in een testomgeving of pilootproject geen verwerking van persoonsgegevens zou zijn die zou vallen onder de dataprotectiewetgeving.

Verantwoordelijken stellen soms dat gegevens worden gecodeerd (pseudonimisering[75]) of geanonimiseerd, waarbij men soms al te vlug redeneert dat het gegevensbeschermingsrecht niet van toepassing zou zijn. Zo vergeet men bovendien vaak dat het omzetten van persoonsgegevens in (zogenaamd) anonieme gegevens ook een verwerking van persoonsgegevens inhoudt. Bij zogezegde "anonimisering" en/of aggregatie van data die in een bepaalde fase van het project wordt toegepast zijn er ook niet altijd afdoende garanties voorhanden die de mogelijkheid van heridentificatie volledig uitsluiten. Er is met andere woorden geen zekerheid dat er geen gegevens worden verwerkt betreffende identificeerbare natuurlijke personen. Soms zijn er dus onvoldoende argumenten voorhanden om de toepasselijkheid van de dataprotectiewet compleet en definitief uit te sluiten.

Men vindt soms de redenering terug dat het feit dat de gegevens worden gecodeerd of geanonimiseerd moet worden beschouwd als een sluitende waarborg voor de bescherming van de privacy. Dit uitgangspunt is niet altijd correct.

---

[70] Overwegingen 74, 84 en 150 AVG.

[71] Overwegingen 39, 58 en 78 en artikel 5.1 a) AVG.

[72] Overwegingen 98, 148 en artikelen 24.3 en 28.5 AVG.

[73] Zie voetnoot 56 hiervoor

[74] De industrie, consumenten, belangenverenigingen en het middenveld, bevoegde overheden en toezichthouders.

[75] Zie de definitie in artikel 4 5) AVG.

Onderzoekers hebben de afgelopen jaren herhaaldelijk aangetoond dat zelfs zogezegd geanonimiseerde of gecodeerde [76] data zonder veel moeite tot specifieke individuen te herleiden zijn (zgn. afzonderen van personen in datasets of "**to single out**" - zie ook de woordenlijst)[77] met het daarbij horende risico op heridentificatie. Bepaalde projecten - zoals bij de analyse van financiële persoonsgegevens bij aankopen [78] of bij het tracken van telecommunicatie [79] - of online activiteit met behulp van cookies - zijn rijk aan data die heridentificatie potentieel mogelijk maken, hoewel ze geen namen, rekeningnummers of evidente identificatienummers bevatten. Projecten waarbij gebruik wordt gemaakt van datasets die rijk zijn aan variabelen (zoals plaats, prijs(categorie), tijdstip en locatie van aankoop bij transactiedata)

waarbij de kans groot is dat een combinatie hiervan een individu ondubbelzinnig bepaalt, hebben een groot risico op "singling out" en daardoor dus op heridentificatie (uiteraard hangt het feitelijk risico op heridentificatie ook af van de gegevens over de betrokkenen waartoe een partij - die tot heridentificatie zou kunnen/willen overgaan - kan verondersteld worden reeds toegang te (kunnen) hebben). Dit is bijvoorbeeld het geval bij data gegenereerd door sommige mobiele apps, financiële metadatasets, surfgeschiedenis, of slimme metergegevens. In dergelijke gevallen is de uniciteit (zie woordenlijst) van de data gewoonweg te groot. Het is dan ook geen toeval dat een aantal van deze datasets reeds (bijkomend) zijn beschermd binnen het communicatiegeheim[80].

In al deze gevallen is er dus sprake is van een verwerking van persoonsgegevens en is de dataprotectiewetgeving zondermeer van toepassing.

Om de kans op singling out of heridentificatie te verkleinen of zelfs uit te sluiten, gaat men soms over tot aggregatie van data[81]. Men kan bij aggregatie dikwijls correct argumenteren – vermits heridentificatie niet meer mogelijk is – dat de dataprotectiewetgeving niet (meer) toepasselijk is (op een deel van het project), bijvoorbeeld bij de mededeling van statistieken naar derden toe. Echter, wanneer het resultaat van de aggregatie vervolgens wordt gebruikt ten aanzien van personen[82] (bv. bij gebruik van correlaties – correlaties kunnen beschouwd worden als een vorm van samenvattende of geaggregeerde informatie – om potentieel gedrag op individuele basis te voorspellen), is er (opnieuw) een (andere) verwerking van persoonsgegevens. De dataprotectiewetgeving zal in dat geval opnieuw van toepassing worden, ook al is er (gedeeltelijk) sprake van het gebruik van anonieme of geaggregeerde gegevens.

Naast aggregatie bestaan er nog andere technieken om data te anonimiseren (zoals randomisatie (ruistoevoeging, permutatie, differentiële privacy) en generalisatie (waarvan technieken voor aggregatie of k-anonimiteit toe behoren, maar verder ook technieken voor l-diversiteit en t-gelijkenis)) [83]. We gaan hier in dit rapport niet verder op in.

---

[76] Zonder dat men in dit geval toegang heeft tot de vertaling van de code naar een betekenisvolle identificator

[77] Zie het eerder aangehaalde onderzoek van De Montjoye Y.-A., Hidalgo C. A., Verleysen M. and Blondel V. D.,"Unique in the Crowd: The privacy bounds of human mobility", Nature Scientific Report 3, 25 March 2013, http://www.nature.com/articles/srep01376, aangehaald op p 67 van Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, gepubliceerd op http://www.wrr.nl/fileadmin/nl/publicaties/PDF-Rapporten/rapport_95_Big_Data_in_een_vrije_en_veilige_samenleving.pdf

[78] De Montjoye Y. A., Unique in the shopping mall: On the re-identifiability of credit card metadata, Science 347, 30 January 2015; http://science.sciencemag.org/content/347/6221/536, 537.

[79] Zie De Montjoye Y.-A., Hidalgo C. A., Verleysen M. and Blondel V. D.,"Unique in the Crowd: The privacy bounds of human mobility", Nature Scientific Report 3, 25 March 2013, http://www.nature.com/articles/srep01376

[80] Zie artikel 5 van de Richtlijn 2002/58/EG van 12 juli 2002 die de vertrouwelijkheid van telecommunicatie beschermt, en artikel 124 van de Wet van 13 juni 2005 betreffende de elektronische communicatie, B.S., 20 juni 2005.

[81] Zie de woordenlijst

[82] Het bepalen van een eigenschap (bv. berekenen van het risico op fraude met behulp van een samenvattend resultaat dat volgt uit big data analyses) en vervolgens automatisch toepassen op een persoon vormt al vrij vlug een geautomatiseerde verwerking van persoonsgegevens.

[83] Zie het advies 05/2014 WP 216 van 10 april 2014 over anonimiseringstechnieken, gepubliceerd http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_nl.pdf

Big data analyse wordt reeds toegepast in België met aggregatie van data (dus ontdaan van elk klantgegeven en de technische identifier), zoals bijvoorbeeld de analyse van geolokatiedata door de Belgische telecomoperatoren Proximus voor toerisme [84] en Orange voor grote evenementen [85]. Het werken met gegroepeerde locatiegegevens van minstens 30 mobiele eindgebruikers achtte de Commissie bij dergelijke projecten aanvaardbaar[86].

Het feit of in een big data project al dan niet sprake is van aggregatie kan tegelijk zowel een toetsingscriterium zijn om de toepasselijkheid van de dataprotectiewetgeving te beoordelen, als om de proportionaliteit van verwerkingen te beoordelen. Of aggregatie kan beschouwd worden als een werkelijke waarborg om de vereiste van een proportionele inmenging in de persoonlijke levenssfeer te garanderen, zal geval per geval moeten worden bekeken.

➢ Aanbeveling 5

Aggregatie, pseudonimisering of (zogezegde) anonimisering van data is niet altijd voldoende om de persoonsgegevens afdoende te beschermen of om heridentificatie onmogelijk te maken. Dit dient steeds geval per geval te worden beoordeeld. Om te kunnen aantonen dat bijvoorbeeld aggregatie van gegevens in een fase van een big data project ook een voldoende waarborg is wordt deze aggregatie daarom best gekoppeld aan **een kwantitatieve beoordeling die de kans op singling out of heridentificatie berekent** op de (geaggregeerde) datasets (bijvoorbeeld **"Small Cells risicoanalyse" of "SCRA"** voor geaggregeerde data)**.** Gezien het soms dynamische karakter van het risico op heridentificatie wordt deze beoordeling in vele gevallen ook best herhaald in de tijd (**periodieke waarborg**).

De Small Cells Risico Analyse (SCRA) is een methode waarbij de kans op indirecte identificatie in geaggregeerde data (in dit rapport is dit dus data waarbij enkel samenvattende statistieken worden gerapporteerd voor groepen van individuen) ingeschat wordt op basis van de combinatie van quasi ID variabelen[87].

De SCRA is een theoretische analyse (die gebeurt zonder dat men de dataset die geaggregeerd dient te worden reeds ter beschikking moet hebben) van het aantal unieke combinaties van de gerapporteerde waarden van deze quasi ID variabelen ten opzichte van het aantal punten in de data op persoonsniveau.

Indien het risico op indirecte identificatie (als gevolg van het potentieel aanwezig zijn van small cells, dit zijn groepen in de geaggregeerde data met een te klein aantal punten op persoonsniveau) te hoog is, kan er worden geadviseerd om een aantal restricties toe te passen (bv. schrappen van een of meerdere quasi ID variabelen, aggregeren van een quasi ID variabele zoals leeftijd tot leeftijdscategorie, …). Bij de risico-evaluatie wordt er ook gekeken of de samenvattende statistieken - die gerapporteerd worden voor de mogelijke small cells - extra en/of sensitieve informatie bevatten over de individuen die in de small cells zitten.

---

[84] http://www.proximus.be/en/id_b_cl_tourism/large-companies-and-public-sector/discover/blog/customer-stories/the-tourism.html

[85] https://business.orange.be/en/node/24126

86 Privacycommissie reageert: 'Iedere koppeling met enig klantgegeven is voor ons onder géén beding aanvaardbaar, 25 november 2016, Datanews, http://datanews.knack.be/ict/nieuws/privacycommissie-reageert-iedere-koppeling-met-enig-klantgegeven-is-voor-ons-onder-geen-beding-aanvaardbaar/article-opinion-781325.html.

[87] Quasi ID variabelen zijn kenmerken die gebruikt worden om de groepen in geaggregeerde data af te lijnen maar die in combinatie ook zouden kunnen worden gebruikt ter identificatie van een individu. Voorbeelden zijn het land van de woonplaats, postcode, geslacht, leeftijd of de geboortedatum. Zie glossary

Het risico van identificatie hangt af van het aantal en de aard van dit type variabelen in de gegevens en van de a priori kennis van diegene die de identificatie probeert uit te voeren.

Dergelijke risico-analyse gebeurt veelvuldig voor dataleveringen in de sector van de sociale zekerheid (bijvoorbeeld door het Intermutualistisch Agentschap (IMA[88]) en betreffende het Thales project[89]).

Indien de aanwezigheid op small cells dient geëvalueerd te worden op het moment dat de dataset die geaggregeerd dient te worden reeds beschikbaar is, is het uiteraard niet meer nodig om een risico analyse te doen maar dient men eenvoudig na te gaan of er groepen (small cells) in de data aanwezig zijn waarvan het aantal punten op persoonsniveau beneden een bepaalde drempel ligt. Indien dit het geval is kan men – en opnieuw alleen indien de samenvattende statistieken die gerapporteerd worden extra en/of sensitieve informatie bevatten over de individuen in de small cells – de informatie over deze groepen maskeren om de anonimiteit maximaal te waarborgen.

Deze small cells risicoanalyse moet dus de uniciteit/"granulariteit" van de datasets of de mogelijkheid "to single out" of heridentificatie beoordelen. Pas indien dergelijke analyse uitwijst dat de mogelijkheid "to single out" afdoende blijft uitgesloten, zijn de data afdoende geaggregeerd of geanonimiseerd. Deze small cells risicoanalyse heeft verder ook alleen zin als er kan vanuit gegaan worden dat diegene die tot heridentificatie wil/kan overgaan, kennis heeft over wie er juist behoort tot de groep individuen die is gebruikt om de dataset op te stellen.

➢ Aanbeveling 6

Aggregatie of anonimisering van gegevens in een fase van een big data project biedt pas sluitende waarborgen voor de bescherming van de persoonsgegevens indien deze maatregel wordt gekoppeld aan de **waarborg**[90] dat er **geen pogingen worden ondernomen tot heridentificatie**, bijvoorbeeld door koppeling met andere datasets.

➢ Aanbeveling 7

De verwerkingsverantwoordelijke dient aggregatie of anonimisering van gegevens te voorzien rekening houdend met de aard en context van de verwerking. Indien anonimisering van gegevens vereist is, hangt het aanvaardbaar risico op singling out of indirecte identificatie immers af van de aard van de partij die de gegevens zal gebruiken en/of het doel waarvoor de data zal worden gebruikt. Voor commercieel hergebruik of verwerking van gezondheidsgegevens bijvoorbeeld, zal derhalve dikwijls een hoog aggregatieniveau aangewezen zijn, terwijl voor andere toepassingen (bv. bestrijden van energiefraude of wetenschappelijk onderzoek met een afdoende reglementair kader) geval per geval kan worden afgewogen of een lager aggregatieniveau kan worden verdedigd.

---

[88] Zie bijvoorbeeld de Small cell analyse van het volledig gegevensbestand van de Gezondheidsenquête 2013 Analyse uitgevoerd door het Intermutualistisch Agentschap, Juli 2015. Zie de Aanbeveling nr 11/03 van 19 juli 2011 van het Sectoraal Comité van de Sociale Zekerheid en van de Gezondheid, Afdeling Gezondheid met betrekking tot een nota van het federaal kenniscentrum voor de gezondheidszorg betreffende de small cell analyse van gecodeerde persoonsgegevens afkomstig van het intermutualistisch agentschap, https://www.ehealth.fgov.be/sites/default/files/assets/nl/pdf/sector_committee/sector_committee_11-03-089_nl.pdf

[89] Zie de Beraadslaging nr. 14/059 van het Sectoraal Comité van de Sociale Zekerheid en van de Gezondheid, Afdeling Gezondheid van 15 juli 2014 met betrekking tot de mededeling van gecodeerde persoonsgegevens die de gezondheid betreffen in het kader van het Thales project, https://www.ehealth.fgov.be/sites/default/files/assets/nl/pdf/sector_committee/2014/14-059-n114-thales-project.pdf

[90] Contractuele verbintenissen en (de mogelijkheid van) controles via audits kunnen naargelang de situatie passende waarborgen vormen.

D. "Juiste Persoonsgegevens"

Volgens de dataprotectiewetgeving [91] heeft de betrokkene het recht om van de verwerkingsverantwoordelijke onverwijld rectificatie van hem betreffende onjuiste persoonsgegevens te verkrijgen.

In de context van big data rijst de vraag wat de correctie van "onjuiste persoonsgegevens" precies betekent. Moet men hier zowel de technische (wiskundige – Punt D.1 hierna) als de maatschappelijke (Punt E hierna) dimensie bekijken. Ook is het nuttig om te wijzen op het verschil tussen associatie door correlatie en associatie door causaliteit voor het leveren van een bewijs (Punt D.2). De Commissie formuleert hierna enkele observaties en aanbevelingen aangaande deze elementen.

D.1. Technisch optimale (zo juist mogelijke) voorspellingen over natuurlijke personen

Technisch optimale voorspellingen over natuurlijke personen bekomen impliceert bijvoorbeeld dat de verantwoordelijke ervoor moet zorgen dat de gebruikte wiskundige modellen optimaal getraind zijn volgens de regels van de kunst, dat het model dus zo goed mogelijk generaliseert op onafhankelijke en representatieve testdata (zie ook Figuur 1 in Punt G) zodanig dat de nauwkeurigheid van de voorspellingen optimaal is[92], dat de variabelen (features) en datasets (voorkomen van data bias) optimaal worden gekozen en dat de resultaten (correlaties) correct worden geïnterpreteerd en de nauwkeurigheid van de voorspellingen juist worden ingeschat. Concreet wil dit onder andere het volgende zeggen:

- ✓ Gebruik van een trainingset die representatief is voor de populatie waarop men het model wil gebruiken om voorspellingen te maken (dit is meestal het geval indien de trainingset een random steekproef is van deze populatie – dit wil dus zeggen dat de trainingset dezelfde statistische distributie moet volgen als deze populatie; zie ook verder bij data bias).
- ✓ Gebruik van technieken die overfitting[93] (dit is het modelleren van random variatie in de trainingset die niets te maken heeft met de onderliggende relatie die men wil capteren, bijvoorbeeld bij het gebruik van een te groot aantal variabelen t.o.v. het aantal punten of observaties) van het model beperken bij het trainen[94]. De nodige waarborgen dienen dus genomen te worden opdat het model zo goed mogelijk generaliseert op onafhankelijke testdata zodanig dat de nauwkeurigheid van de voorspellingen optimaal is.
- ✓ Gebruik van een testset die eveneens representatief is voor de populatie waarop men het model wil gebruiken om voorspellingen te maken. Dit is belangrijk om een nauwkeurige schatting te maken van de nauwkeurigheid van het model (zie ook hieronder bij data bias).
- ✓ Gebruik van een onafhankelijke testset, d.w.z. dat deze op geen enkele manier mag gebruikt zijn tijdens het trainen van het model (indien dit wel gebeurt zal dit over het algemeen aanleiding geven tot het overschatten van de performantie van het model).

[91] Zie artikel 12 WVP en 16 AVG.

[92] Zie randnummer 35 van het advies nr. 25/2016 van 8 juni 2016 betreffende het ontwerp van decreet " tot wijziging van het Energiedecreet van 8 mei 2009, wat betreft het voorkomen, opsporen, vaststellen en bestraffen van energiefraude"

[93] Zie woordenlijst

[94] Voor een uitleg rond supervised learning en model training : zie onder meer deel I en figuur 1.

- ✓ Gezien de populatie waarvoor men voorspellingen wil maken in veel gevallen dynamisch is (d.w.z. dat de kenmerken ervan in de loop van de tijd veranderen), is het belangrijk om de modellen op geregelde tijdstippen te her-trainen met een nieuwe en voor dat moment representatieve trainingset. Het evalueren van het model moet vervolgens ook gebeuren met een nieuwe en voor dat moment representatieve testset.

- **Keuze van de datasets en voorkomen van data bias**

Volgens de literatuur[95] is "big data" dikwijls synoniem voor het beschikken over "meer data". Het risico is hierbij dat bij het maken van voorspellingen er gebruik wordt gemaakt van verouderde data, gebiased data of niet relevante of niet representatieve (voor de populatie waarvoor men voorspellingen wil doen) data[96]. Dit verschilt met data met veel ruis, waaruit door de grootte van de datasets wel dikwijls nauwkeurige voorspellingen kunnen worden gedaan[97].

Ernstige risico's voor de rechten en vrijheden van de betrokkenen zijn te vinden bij beslissingen die gekoppeld zijn aan voorspellende big data modellen met ingebouwde vooroordelen (training gebeurt bijvoorbeeld aan de hand van data die niet representatief is voor de populatie waarvoor men voorspellingen wil doen, zgn. "data bias")[98]. Een voorspellend big data model dat getraind werd op biased data impliceert ook dat het risico toeneemt dat personen verkeerd worden geclassificeerd en bovendien dat bepaalde groepen personen systematisch verkeerdelijk aan een bepaalde klasse (bijvoorbeeld van "mogelijke fraudeur") worden toegekend.

Bij het voorkomen van data bias bij testdata bestaat het risico dat de nauwkeurigheid van het model verkeerd wordt ingeschat.

➢ Aanbeveling 8

Gezien het voormelde risico van "data bias" zich kunnen manifesteren bij toepassing van big data, zullen de verantwoordelijken (bv. in hun gegevensbeschermingseffectboordelingen) steeds moeten **nagaan of er data bias verscholen zit in de gekozen datasets en methodes om bijvoorbeeld wiskundige modellen te trainen of te testen**. Vragen zullen moeten worden toegelicht zoals waarom bepaalde bevolkingsgroepen worden uitgesloten, welke datapunten minder zichtbaar zijn bij training of testen, … .

- **Misverstanden en randvoorwaarden voor correcte big data analyses**

[95] V. Mayer-Schönberger & K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work and think, Boston:Houghton Mifflin Harcourt, 2013.

[96] Zie de verwijzingen naar een slechte selectie van data, onvolledige, incorrecte of verouderde data, vooroordeel bij de selectie van de populatie, en onbedoelde bevestiging en promotie van historische vooroordelen op pagina's 6 en 7 van Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[97] VAN DER SLOOT, B. en VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p 32, http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf

[98] Verborgen vooroordelen die verscholen zitten in gegevens (data is niet representatief voor de populatie onder beschouwing). zie voor een voorbeeld CRAWFORD, Kate, The hidden biases in big data., 1 april 2013, Harvard Business review, gepubliceerd op https://hbr.org/2013/04/the-hidden-biases-in-big-data, o.m. aangehaald in FTC, Big Data. A Tool for inclusion or exclusion ? Understanding the Issues, January 2016, https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

Er bestaat een misverstand dat big data algoritmes of analyses steeds perfecte resultaten zouden opleveren (bv. foutloze classificaties). Het is een misverstand dat fouten zouden kunnen worden weggefilterd doordat men werkt met grote groepen van datapunten. Zelfs indien wiskundige modellen volgens alle regels van de kunst worden getraind en getest, zullen in de meeste projecten nog steeds personen verkeerd geclassificeerd worden (performantie van wiskundige modellen is nooit perfect).

Algoritmes zijn bovendien en eigenlijk maar zo goed als de data waarmee ze werken[99]. Niet alle data bevatten bijvoorbeeld voldoende informatie over hetgeen men wil modelleren (waardoor men modellen zal bekomen met een performantie die laag of onbestaande is). Zo staat de nood aan een afdoende en correcte data-input in een gespannen verhouding met het feit dat het niet altijd evident of zelfs toegelaten is om de relevante data te bemachtigen of te (her)gebruiken. Zo wordt het meer gangbaar voor de fiscale overheid om "fishing expeditions" op te zetten aan de hand van zo groot mogelijke maar niet altijd betrouwbare databronnen, zoals bij de doorlichting van sociale webpagina's, discussiefora en markplaatsen, of bij het massaal doorlichten van de totale groep van alle gebruikers van (vreemde) betaalkaarten[100]. Het probleem is dat de informatie over die populaties niet steeds betrouwbare en representatieve informatie inhoudt om de ganse populatie van belastingbetalers en –ontduikers te onderzoeken als "potentieel fraudeur".

Bij voorspellende modellen is er ook een verschuiving in het beoordelen van personen van een model gebaseerd op causaliteit naar een model dat is gebaseerd op correlatie[101]. De correlaties die gevonden worden met behulp van big data worden ook niet altijd correct geïnterpreteerd. Het kan gebeuren dat deze resultaten worden veralgemeend (terwijl ze bijvoorbeeld alleen geldig zijn voor de populatie waaruit de trainingset komt), of dat in de verdere berichtgeving of popularisering van de onderzoeksresultaten geen rekening meer wordt gehouden met het verschil tussen causaliteit en correlatie.

Het is dus cruciaal om een minimaal technisch inzicht te hebben in en om maatschappelijk verantwoord om te kunnen gaan met "correlaties", "data bias" en de performantie of de kracht (en in sommige gevallen het gebrek hieraan) van modellen. Zonder dit inzicht kan men door het gebruik van profielen soms ongewenste sociale en ethische effecten veroorzaken.

Uit voorgaande blijkt dus dat het (technisch) correct omgaan met het gebruik van big data analyses, geen evidentie is en gekoppeld aan tal van randvoorwaarden. Best wordt derhalve dienaangaande een brede zorgvuldigheidsverplichting opgenomen.

➢ Aanbeveling 9

Een **zorg(vuldigheids)plicht** ten aanzien van de verwerkingsverantwoordelijken dient te worden opgenomen m.b.t. de kwaliteit van de data en de deugdelijkheid van de analysemethodes.

---

[99] BAROCAS, S. en SELBST, A.D., Big Data's Disparate Impact, California Law Review 671, 11 augustus 2016, gepubliceerd op http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

[100] Fiscus screent massaal bankkaarten, De Tijd, 10 september 2016; Fiscus wil via bankkaarten jacht maken op verborgen buitenlandse rekeningen, http://deredactie.be/cm/vrtnieuws/binnenland/1.2763572.,

[101] FTC, Big Data. A Tool for inclusion or exclusion ? Understanding the Issues, January 2016, https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

➢ **Aanbeveling 10**

Om oordeelkundig om te gaan met big data analyses is het vereist dat toezichthouders, verantwoordelijken en beleidsvoerders hun big data expertise versterken en afdoende en op regelmatige basis **opleiding en training** voorzien om oordeelkundig en verantwoord om te gaan met (het resultaat van) deze technieken.

D.2. Associatie door een algoritme vs juridisch bewijs

Het gebruik van bestandsvergelijking, een statistiek of algoritme die een associatie (verband) aantoont tussen eigenschappen of fenomenen (bv. verband tussen afwijkend energiegebruik en kans op fraude, verband tussen een betalingsachterstand inzake mobiele telefonie en betalingsachterstand bij krediet[102],…) is niet meteen hetzelfde als het leveren van een juridisch bewijs van een eigenschap bij een specifieke persoon of een oorzakelijk verband tussen twee fenomenen. Nochtans bestaat vaak het misverstand dat correlationele verbanden oorzakelijk zijn (zie hierboven), en dat correlationele verbanden of voorspellingen van een wiskundig model (die in individuele gevallen dus verkeerd kunnen zijn) een dringend maatschappelijke noodzaak impliceren om ongewenste fenomenen of individuele "hits" bij voorspellingen door big data analyses direct aan te pakken zoals bij sociale fraude en betalingsachterstand. Hoogstens kan er bij het gebruik van big data analyses en correlaties sprake zijn van een "knipperlicht" of aanwijzing[103] waaraan de wetgever niet meteen een bewijswaarde (tot bewijs van het tegendeel door de betrokkene) aan kan toekennen, maar enkel een indicator die verder dient te worden onderzocht in een individueel en gepersonaliseerd onderzoek mede d.m.v. een tussenkomst door een fysiek persoon. In het kader van bijvoorbeeld het opsporen van fraude dienen dergelijke analyses enkel om de beperkte middelen zo efficiënt mogelijk (met een zo groot mogelijk resultaat) in te zetten doordat men het onderzoek kan richten op een groep personen of organisaties waar een hogere kans bestaat om fraude terug te vinden (i.p.v. dat men bijvoorbeeld ongericht en steekproefsgewijs zou controleren in de algemene populatie), zonder dat daar op het niveau van de data-analyse al noodzakelijk sluitend bewijs zou voor bestaan. De bedoeling is m.a.w. om het zoekvenster te richten en te verkleinen. Men dient wel rekening te houden met de foutenmarge (mogelijkheid van misclassificaties van individuen door modellen), de sociale impact van dergelijke maatregelen, en het voorzien van waarborgen om de betrokkenen afdoende te beschermen.

Ervan (verkeerdelijk) uitgaan dat correlaties met fraude, betalingsachterstand,… altijd een objectieve waarheid inhouden op basis waarvan altijd maatregelen moeten worden genomen noemt men ook wel "**datadeterminisme**"[104].

➢ **Aanbeveling 11**

[102] Zie het werkdocument n° 112 van de Nationale Bank van 2011 dat een verband aantoonde tussen betalingsachterstanden inzake mobiele telefonie en betalingsachterstanden bij krediet, https://www.nbb.be/en/articles/working-paper-ndeg-212

[103] Zie de Wet van 13 mei 2016 tot wijziging van de programmawet (I) van 29 maart 2012 betreffende de controle op het misbruik van fictieve adressen door de gerechtigden van sociale prestaties, met het oog op de invoering van het systematisch doorzenden naar de KSZ van bepaalde verbruiksgegevens van nutsbedrijven en distributienetbeheerders tot verbetering van de datamining en de datamatching in de strijd tegen de sociale fraude, B.S., 27 mei 2016.

[104] Data determinism – can data really speak for itself?, 5 februari 2014, https://datasciencebusiness.wordpress.com/2014/02/05/data-determinism/

Het is belangrijk dat er wettelijke regelingen worden voorzien die bepalen hoe en wanneer het resultaat van data mining en statistische analyses (correlaties) door de overheid kan gebruikt worden als juridisch bewijsmateriaal.

Indien de wetgeving het gebruik van big data analyses oplegt of mogelijk maakt (bijvoorbeeld via voorspellingen door wiskundige modellen of vergelijking van gegevens uit verschillende databases) om een indicatie te geven van een welbepaald risico (bv. op fraude, terrorisme of criminaliteit) en/of om beslissingen te nemen ten aanzien van een individuele betrokkene, dan moet de wetgever wel rekening houden met de bijna steeds aanwezige **foutenmarge** van deze analyses, en de bewijswaarde enkel toekennen aan een bijkomende menselijke beoordeling van de individuele gevallen (waarbij de resultaten uit de big data analyse bijvoorbeeld worden gecombineerd met fysieke vaststellingen en verificaties en met een menselijke beoordeling van ondermeer de betrouwbaarheid en het recent en niet-discriminatoir karakter van de big data indicaties) die volgt op de analyses, alvorens beslissingen worden genomen in individuele dossiers met mogelijke nadelige gevolgen voor de betrokkenen.

➢ **Aanbeveling 12**

Zowel voor toepassingen in de publieke en private sector (bv. bij kredietverlening, aanwervingen,…) dienen dus afdoende transparantie en **procedurele waarborgen** te worden voorzien die een correcte wijze van beslissing ten aanzien van individuen waarborgt die verder gaat dan het handelen op basis van alleen een correlatie of statistisch verband.

Een voorbeeld van een dergelijke waarborg bij de aanpak van fraude is het voorzien van een wettelijke plicht om een proces-verbaal van vaststellingen[105] op te maken waarin de fysieke vaststellingen of de gepersonaliseerde analyse van een onderzoeker worden opgenomen, en een verplichting van betekening van het proces-verbaal met reactietijd voor de betrokkene. Bovendien mag er in dit geval niemand formeel worden beschuldigd of gesanctioneerd op basis van het gebruik van statistische verbanden, het resultaat van wiskundige modellen of bestandsvergelijking. Wel kan dit een rechtvaardiging zijn voor verder onderzoek.

E. Individuele of collectieve impact voor de rechten en vrijheden van betrokkenen

Big data projecten kunnen een impact hebben op de rechten en vrijheden van natuurlijke personen (zeker als er technisch suboptimale keuzes zijn gemaakt maar zelfs ook indien men al de regels van de kunst volgt). Ook kan het resultaat sociaal niet aanvaardbaar zijn, bijvoorbeeld doordat discriminatiebepalingen worden geschonden (Punt E.2). In beide gevallen kan het beginsel van een eerlijke verwerking van persoonsgegevens worden overtreden (zie ook hierna onder Punt F).

E.1. Impact van big data op individuele natuurlijke personen

De AVG[106] vereist, zoals hierboven reeds aangehaald, dat de verantwoordelijke een continue

[105] Zie verwijzing in randnummer 39 ev van het advies 24/2015 van 17 juni 2015 mbt sociale fraude (verwijzing naar waarborgen) in fiscale rechtspraak mbt fiscaal visitatierecht

risicobeoordeling en gegevensbeschermingseffectbeoordeling verricht in functie van de impact voor de rechten en vrijheden van de betrokken personen.

Doorgedreven toepassing van voorspellingen ten aanzien van individuele personen draagt een risico op ernstige inmenging in de rechten en vrijheden van dit individu. Dit kan het geval zijn bij controles door politiediensten gebaseerd op onderzoek van facebookprofielen (predictive policing), gebruik van correlaties tussen hoog of laag energiegebruik en het voorkomen van fraude (verliezen van een uitkering), of het gebruik van gegevens via smartphone applicaties door verzekeraars voor de premiebepaling van een ziekteverzekering of de verzekering burgerlijke aansprakelijkheid voor motorvoertuigen[107].

E.2. Maatschappelijke impact van big data op bepaalde sociale groepen

Er werd reeds elders [108] gewezen op het gebruik van bepaalde correlaties, datasets of algoritmes die sociale stratificatie veroorzaken (zgn plaatsen van een populatie in categorieën of "social sorting"[109]), met mogelijk onwettige en ongelijke behandeling van sociale groepen tot gevolg. Het is weliswaar mogelijk dat deze ongelijke behandeling niet intentioneel is maar verscholen zit in de data of de algoritmes.

Niet elke ongelijke behandeling van personen is echter ook onwettig[110]. Dit is wel het geval indien er bijvoorbeeld sprake is van discriminatie[111].
Op technisch vlak kunnen in dit kader de volgende mechanismen[112] voorkomen:

- ✓ **Definitie van de klasse labels die leden van een sociale groep preferentieel in een categorie plaatst die benadeeld wordt** (bijvoorbeeld wanneer men in het kader van een beslissing tot aanwerven een model gebruikt dat een onderscheid maakt tussen "goede" en "slechte" werknemers waarbij het onderscheid gemaakt wordt door de lengte van de periode gedurende dewelke iemand in dienst blijft te vergelijken met een arbitraire waarde). Wanneer bepaalde sociale groepen om een of andere reden meer frequent van werk veranderen zal dit hun kans – ten gevolge van het gebruik van een model getraind op data met dergelijke klasse labels – op aanwerving verminderen zelfs indien ze een even productief zouden zijn als anderen.
- ✓ **Trainingdata die op een of andere manier gebiased is**. Bijvoorbeeld doordat ze gebaseerd is op trainingdata waar **het toekennen van een klasse label reeds gebaseerd is op een vooroordeel**. In dit geval zal het resulterende model het vooroordeel enkel bestendigen. Of bijvoorbeeld doordat ze gebaseerd is op

[107] VAN DER SLOOT, B. en VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p 100, http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf

[108] Zie het system van kredietbeoordelingen in de VS, aangehaald op pagina's 11 tot en met 13 van Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Mei 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf; Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, punt 4.5.1., pagina 89.

[109] D. Lyon, Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination, London, Routledge, 2003, beschikbaar op https://infodocks.files.wordpress.com/2015/01/david_lyon_surveillance_as_social_sorting.pdf en
BROEDERS, D. en SCHRIJVERS, E., Big Data in een vrije en veilige samenleving, 20 mei 2016, http://njb.nl/njv-jaarvergaderingen/jaarvergadering-2016/artikelen/big-data-in-een-vrije-en-veilige-samenleving.19799.lynkx

[110] Dit belet evenwel niet dat wettelijke maar ongelijke behandeling risico's kan inhouden voor de rechten en vrijheden van de betrokken natuurlijke personen, waarmee de verwerkingsverantwoordelijke zal rekening dienen te houden.

[111] De antidiscriminatie- en antiracismewetten hebben zowel een burgerlijk als een strafrechtelijk luik. Voor het strafrechtelijk luik is een bijzonder opzet vereist maar voor het burgerlijk luik heeft de intentie geen enkel belang.

[112] Solon Barocas and Andrew D. Selbst, (2016) Big Data's Disparate Impact, 104 Cal. L. Rev. 671, gepubliceerd op

trainingdata waar in een bepaalde bevolkingsgroep er preferentieel personen zijn gekozen die in de "slechte" of meest nadelige categorie zitten.

- ✓ Wanneer het lidmaatschap van een bevolkingsgroep (of een proxy hiervoor; zie verder ook i.v.m. beschermde criteria) die op een bepaald gebied minder goed scoort (bv. hogere misdaadcijfers in een bepaalde etnische groep), een criterium is op basis waarvan een model een beslissing neemt in dit kader, dan kan **het gebruik van een te beperkte set** – die niet toelaten om de vanzelfsprekende variabiliteit die er in deze bevolkingsgroepen zit te modelleren – **van variabelen of features** in het model leiden tot een model dat nadelige beslissingen neemt ten aanzien van al de leden van een bepaalde groep (bv. lidmaatschap van een etnische groepering geeft op zich een veel hogere kans tot voorspelling van crimineel gedrag, zonder rekening te houden met andere kenmerken van een individu).
- ✓ Het **intentioneel gebruiken** van hierboven beschreven mechanismen (die meestal on-intentioneel tot stand komen) met het doel om bepaalde groepen te discrimineren.

Het **ongelijk behandelen** van natuurlijke personen die geen **verboden discriminatie** vormt in de zin van de toepasselijke wetgeving[113], kan niet a priori als onwettig worden beschouwd onder de antidiscrimpinatiewetgeving.

Bij het aanpakken van maatschappelijke problemen met behulp van big data, kunnen door schijncorrelaties[114], "datafundamentalisme"[115] en "data bias" nadelige effecten ontstaan voor een specifiek deel van de populatie. Onderzoeken van bepaalde algoritmes in verband met vervoer[116] of verkeer[117] toonden het risico aan dat een op een beslissingsalgoritme gebaseerde verdeling van voordelen, een ongelijke verdeling van (overheids)middelen of deze voordelen kan opleveren (herstellen van wegen, het al dan niet verlenen van een dienst binnen een bepaalde reactietijd, het geven van kortingen op diensten). Dit kan impliceren dat middelen niet altijd efficiënt worden toegepast ten aanzien van bevolkingsgroepen omwille van incorrecte veronderstellingen, of doordat een verschillende behandeling ten aanzien van bepaalde bevolkingsgroepen wordt bestendigd, verscherpt of gemaskeerd[118].

Andere voorbeelden betreffen prijsdiscriminatie en/of big data projecten die gelijke toegang tot kansen op de arbeidsmarkt of diversiteit op de werkvloer[119] ondermijnen. Bijvoorbeeld een HR bureau dat met behulp van rekruteringssoftware blindelings een "hiring for culture fit"[120]

---

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

[113] Zie onder meer de Wet van 10 mei 2007 ter bestrijding van bepaalde vormen van discriminatie, B.S., 30 mei 2007.

[114] Fenomeen dat men een verband vermoedt tussen een sensitief criterium (bv. etniciteit) en hetgeen men wil voorspellen, terwijl andere (minder sensitieve) criteria op zich betere voorspellingen toelaten zonder dat men dit sensitief criterium nodig heeft..

[115] Er verkeerdelijk van uitgaan dat correlatie altijd duidt op een causaal verband en dat massale datasets en voorspellende analyses altijd de objectieve waarheid weergeven. Zie CRAWFORD, Kate, The hidden biases in big data., 1 april 2013, Harvard Business review, gepubliceerd op https://hbr.org/2013/04/the-hidden-biases-in-big-data en pagina 10 van Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[116] Zie bijvoorbeeld het onderzoek naar het "surge pricing" algoritme van Uber, waarbij werd aangetoond dat dit algoritme hogere prijzen en betere dienst opleverde voor bepaalde wijken, en lagere prijzen en een slechtere dienst voor andere buurten. DIAKOPOLOUS, N., How Uber surge pricing really works, 17 April 2015, https://www.washingtonpost.com/news/wonk/wp/2015/04/17/how-uber-surge-pricing-really-works/

[117] Zie de "Streetbump smartphone app" die de stad Boston wou gebruiken om middelen voor wegenwerken in te zetten. Deze werkwijze werd echter geplaagd werd door het verborgen vooroordeel dat elke persoon gelijke toegang had tot deze app. Zie CRAWFORD, Kate, The hidden biases in big data., 1 april 2013, Harvard Business review, gepubliceerd op https://hbr.org/2013/04/the-hidden-biases-in-big-data

[118] Zie p 12 van Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[119] DE ANCA, C., Why hiring for cultural fit can thwart our diversity efforts, Harvard Business review, 25 April 2016, https://hbr.org/2016/04/why-hiring-for-cultural-fit-can-thwart-your-diversity-efforts

[120] Dit vanuit het uitgangspunt dat de gewenste kandidaat best een correlatie vertoont met de dominante etnische afkomst, leeftijdsgroep of ras voor een bedrijf. Zie HUHMAN, H., Not Using Big Data for Hiring? You May Be Missing Out on the Best Candidates, 10 April 2014, https://www.entrepreneur.com/article/232780

strategie zou toepassen met onbewuste[121] discriminatie op basis van beschermde criteria zoals godsdienst of politieke of levensbeschouwelijke overtuiging, etnische afkomst, leeftijd, geslacht of ras, zwangerschap en moederschap, en handicap[122]. In dit voorbeeld zou men er over moeten waken dat een algoritme historisch benadeelde of kwetsbare groepen deelname aan de maatschappij niet ontzegt.

➢ Aanbeveling 13

De **verwerkingsverantwoordelijken** nemen de nodige voorzorgen om zoveel mogelijk uit te sluiten dat big data analyses zouden aanleiding geven tot verboden vormen van discriminatie[123] in bepaalde bevolkingsgroepen of tot ongelijke behandelingen die een hoog risico kunnen uitmaken voor de rechten en vrijheden van de betrokkenen. Het in kaart brengen van de impact op bepaalde sociale of culturele groepen kan gebeuren in de gegevensbeschermingseffectbeoordelingen.

De **wetgever** zou ook het beginsel van "gelijke kansen door ontwerp ("equal opportunity by design")[124] kunnen hanteren, om te vermijden dat big data technieken zouden worden gebruikt die in strijd zijn met antidiscriminatiebepalingen of het eerlijkheidsbeginsel (zie ook hierna onder Punt F). Hierbij zou de aandacht kunnen gaan naar beschermde criteria of variabelen, die niet of alleen met de grootste omzichtigheid mogen worden gebruikt in modellen die beslissingen nemen die personen in belangrijke mate treffen

Sociale correcties in individuele dossiers kunnen voor een deel een oplossing bieden als potentieel modellen zijn gebruikt met data bias of naar verhouding veel misclassificaties in bepaalde groepen. Deze correcties kunnen de vorm aannemen van een mogelijkheid tot verhaal, klacht of beroep, een vraag tot herzien van dataclassificaties in dossiers, en het voorzien van feedbackmechanismen die verkeerde classificaties in gelijkaardige gevallen in de toekomst pogen te reduceren.

Op collectief vlak kan wetgeving de betrokkenen beschermen tegen te strenge reductie van kansen inzake basisdiensten ten gevolge van doorgedreven profilering op basis van beschermde criteria of variabelen (leeftijd of gezondheidsgegevens bijvoorbeeld). Dit was bijvoorbeeld het geval in de wet Partyka[125] aangaande het aangaan van schuldsaldoverzekeringen door personen met een verhoogd gezondheidsrisico, die een systeem van goedgekeurde vragenlijsten introduceerde.

---

[121] De antidiscriminatie- en antiracismewetten hebben zowel een burgerlijk als een strafrechtelijk luik. Voor het strafrechtelijk luik is een bijzonder opzet vereist maar voor het burgerlijk luik heeft de intentie geen enkel belang.

[122] Dit criterium komt expliciet voor in de respectievelijke wetgevingen maar ook in Richtlijn 2000/78/EU en het belang ervan neemt toe gelet op het VN Verdrag inzake rechten van personen met een handicap. Sinds Europa dit Verdrag heeft ondertekend moet de richtlijn conform dit Verdrag geïnterpreteerd worden.

[123] Zie onder meer de Wet van 10 mei 2007 ter bestrijding van bepaalde vormen van discriminatie, B.S., 30 mei 2007.

[124] Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

[125] De "wet Partyka" betreft de artikels 212 tot 224 van de wet van 4 april 2014 betreffende de verzekeringen (Belgisch Staatsblad van 30 april 2014, gepubliceerd in het Staatsblad van 3 februari 2010. De uitvoeringsbesluiten verschenen met het Koninklijk besluit van 10 april 2014 tot regeling van sommige verzekeringsovereenkomsten tot waarborg van de terugbetaling van het kapitaal van een hypothecair krediet (Staatsblad van 10 juni 2014).

✓ **Aanbeveling 14**

Sociale correcties[126] op individueel of collectief vlak moeten voorhanden zijn teneinde in een zo vroeg mogelijke fase vooroordelen te vermijden, antidiscriminatiebepalingen (die van openbare orde zijn) te respecteren, en om tijdig cumulatieve nadelen voor bepaalde bevolkingsgroepen te vermijden (risico op discriminatie of oneerlijke verwerking door ongelijke behandeling).

➢ **Aanbeveling 15**

Het is vanuit de invalshoek van de verantwoordingsplicht [127] en/of de gegevensbeschermingseffectbeoordeling [128] onder de AVG van belang dat de verantwoordelijken verantwoordingsmechanismen voorzien die de basis waarop het algoritme beslissingen neemt transparant maakt (zie ook verder onder Punt L) en voorwerp maakt van individuele en/of maatschappelijke **feedback**, dit laatste vooral als de beslissingen van het algoritme de betrokkenen in belangrijke mate kunnen treffen.

Feedback zal noodzakelijk blijken als er sprake is van de combinatie van een aantal hoge risico's voor de rechten en vrijheden van de betrokkenen, fouten of beveiligingsincidenten[129] in verband met te verregaande vormen van big data analyses (dit wil zeggen met een te hoog residueel risico).

Deze feedback [130] kan naargelang de context verschillende vormen aannemen, zoals het inbouwen van 'interactieve modellering" [131] (waarbij de betrokken personen (klanten) het algoritme mee kunnen sturen), een filtermechanisme[132] (bv. maatregelen die de verspreiding van onjuiste berichten via sociale media zou kunnen inperken), het voorzien van afdoende maatschappelijke transparantie en een formeel beoordelingsmoment en/of bijzondere rapporteringsplicht naar de toezichthouders en populatie toe over de impact van het model.

Men dient er in dit kader wel op te letten dat passende waarborgen worden geboden om een evenwicht te vinden tussen het geven van feedback en de bescherming van de commerciële of algemene belangen van de verwerkingsverantwoordelijken of de beveiliging van verwerkingen[133].

Voor concrete toepassingen kan dit evenwicht ook worden voorzien door de werking van

---

[126] Via bijvoorbeeld bepalingen in de wetgeving die aandacht schenkt aan het risico op discriminatie of ongelijke behandeling van data mining.

[127] Artikel 5.2 AVG.

[128] Artikel 35.9 AVG

[129] STROUD, M., The minority report: Chicago's new police computer predicts crimes, but is it racist? Chicago police say its computers can tell who will be a violent criminal, but critics say it's nothing more than racial profiling, The Verge, 19 februari 2014, http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist.

[130] VEALE, M., Shadow of the smart machine: Computer scientists and social scientists must work together on algorithms, 26 January 2016, Shadow of the smart machine series, http://www.nesta.org.uk/blog/shadow-smart-machine-computer-scientists-and-social-scientists-must-work-together-algorithms

[131] Zie een toepassing in de context van journalistiek : de conclusie van het artikel DIAKOPOULOS, N., Accountability in Algorithmic Decision Making, Februari 2016, Communications ACM, http://www.nickdiakopoulos.com/wp-content/uploads/2016/03/Accountability-in-algorithmic-decision-making-Final.pdf

[132] SOLON, O., Facebook's failure: did fake news and polarized politics get Trump elected?, 10 November 2016, https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories.

[133] Overweging 63 en artikel 35.9 AVG.

algoritmes, bestandsvergelijkingen,… op kritische, onafhankelijke en op wetenschappelijke wijze te laten onderzoeken vanuit een bredere maatschappelijke context.

✓ **Aanbeveling 16**

Modellen zijn nooit perfect en misclassificaties kunnen geconcentreerd zitten in bepaalde bevolkingsgroepen. De output van algoritmes dient, indien zich in deze context mogelijks problemen stellen, kritisch en waar mogelijk onafhankelijk te worden onderzocht op fouten en het model dient desgevallend te worden aangepast of bijgestuurd.

F. Eerlijkheidsbeginsel

Het eerlijkheidsbeginsel dat actueel vervat zit in artikel 4 § 1, 1° WVP en artikel 5.1 (a) AVG[134] staat in direct verband met de vereiste van een transparante verwerking en de wijze waarop gegevens werden verkregen.

De Commissie onderscheidt enkele hypotheses waarbij het eerlijkheidsbeginsel kan worden overtreden.
Enerzijds kan er sprake zijn van een gebrek aan transparante verwerking[135] (zie ook verder onder Punt L). Anderzijds kan de toepassing van een model ervoor zorgen dat diverse verbodsbepalingen op regionaal, federaal of Europees vlak inzake discriminatie[136] worden geschonden (zie ook hierboven onder F.).

G. Proportionaliteit: noodzakelijkheid, wijze van opslag en beginsel van minimale gegevensverwerking

Een inmenging in de persoonlijke levenssfeer is pas legitiem als er kan worden aangetoond dat deze inmenging **noodzakelijk is in een democratische samenleving**[137]. Bovendien moeten de gebruikte gegevens niet overmatig en toereikend zijn ten opzichte van het doel waarvoor zij verwerkt worden.

Gelet op het risico van tal van valkuilen (data bias, vals positieven en vals negatieven, datadeterminisme,…) zijn big data analyses nooit a priori een effectieve methode om individueel gedrag accuraat te voorspellen. De bruikbaarheid van deze methode hangt af van de vraag of er over het te onderzoeken fenomeen voldoende relevante en informatieve gegevens beschikbaar zijn, en/of uit de data betekenisvolle patronen kunnen worden afgeleid aan de hand van de juiste werkwijze, en of burgers hun gedrag aanpassen in reactie op grootschalige observatieparktijken. Zo werd betoogd[138] dat data mining een ineffectieve methode zou kunnen zijn voor het zoeken naar terroristen, georganiseerde misdaad of mensensmokkel gelet op het feit dat deze fenomenen vaak een onvoldoende regelmatig karakter hebben, dat er te weinig betrouwbare of voldoende data is, en dat er te weinig overeenkomsten zijn om een goed profiel te maken.

---

[134] Het Engelse woord "fairness" werd onjuist vertaald in het Nederlands als "behoorlijkheid", en correct in het Frans als "loyauté"

[135] Ondernemingen die zonder afdoende informatie aan de betrokkenen data vergaren of correlaties maken en distribueren (bv. door identificatie aan de hand van koppeling met open data), respecteren het eerlijkheidsbeginsel niet.

[136] Zie voor een overzicht van de relevante wetgeving pagina 7 van het discriminatielexicon van UNIA, gepubliceerd op http://www.unia.be/nl/wetgeving-aanbevelingen/wetgeving/discriminatielexicon.

[137] Deze vereiste volgt niet uit de AVG, maar uit artikel 8 EVRM en 22 Grondwet. Zie ook de artikelen 7 en 52 van het Europees Handvest.

[138] Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, p 83 punt 4.3.2.

Ondanks de beloften tot het bereiken van positieve resultaten bij het inzetten van big data voor de bestrijding van fraude en bepaalde vormen van criminaliteit, ontbreekt het in veel gevallen aan een betrouwbare en objectieve onderbouwing en evaluatie van de effectiviteit van de gebruikte analysemethodes.

✓ Aanbeveling 17

Het gebruik van big data in de publieke sector, in het bijzonder voor de promotie van nationale veiligheid, criminaliteitsbestrijding en ordehandhaving, dient steeds te worden onderworpen aan een onderzoek door de toezichthouders qua wenselijkheid en efficiëntie[139].

Een ander risico in het licht van het proportionaliteitsbeginsel is dat men de mogelijkheden en voorwaarden onder de klassieke surveillemethodes van (politionele) observatie van de publieke ruimte al te vlug gaat gelijkstellen met online observatie. Het digitale "patrouilleren" door de politie (bv. op websites, fora, sociale media) of vormen van "predictive policing" zoals in het recent aangekondigde i-Police project[140], gebeurt op een andere schaal en met andere methodes (opslag en analyse van data) dan voorhanden voor de "offline" proactieve observaties op de voor het publiek toegankelijke plaatsen voor doeleinden van bestuurlijke of gerechtelijke politie.

Een geplande wijziging van de wet op het politieambt[141] probeerde eerder het digitale patrouilleren gelijk te stellen met het betreden van publiek toegankelijke plaatsen door de politie. De Commissie associeerde[142] echter het digitale patrouilleren met de bijzondere opsporingsmethodes, en wees ook op problemen met het legaliteitsbeginsel.

Uit het gegevensbeschermingsrecht volgt verder dat persoonsgegevens niet langer mogen worden bewaard dan voor de doeleinden waarvoor de persoonsgegevens worden verwerkt noodzakelijk is (in een vorm die het mogelijk maakt de betrokkenen te identificeren) (beginsel van opslagbeperking[143]). In het kader van big data projecten, waar data dikwijls voor onbepaalde tijd worden opgeslagen in afwachting van een nuttige toepassing (latere verwerking – zie ook verder onder Punt H), kan er sprake zijn van miskenning van dit beginsel van opslagbeperking.

Het gebruiken van meer persoonsgegevens dan noodzakelijk voor een welbepaald doel (wat dikwijls het geval is bij big data projecten) staat in een gespannen verhouding met het "**beginsel van minimale gegevensverwerking**"[144]. Hiermee is onder andere rekening gehouden bij de archivering van (persoons)gegevens waarbij archiefbestanden in de EU doorgaans aan zeer doordachte en strenge selectiecriteria worden onderwerpen en slechts een fractie wordt bewaard, in plaats van de rol van de archivaris te herdefiniëren als een proactieve bedenker van mogelijkheden om hergebruik van gegevens vanuit andere

[139] VAN DER SLOOT, B. en VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p 28, gepubliceerd op http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf

[140] Ponciau, L, Les détails du projet iPolice dévoilés, Le Soir, 17 september 2016.

[141] Zie het randnummer 14 en volgende van het advies 13/2015 van 13 mei 2015 betreffende het Voorontwerp en van wet houdende diverse bepalingen – wijzigingen aan de wet tot oprichting van een beroepsorgaan inzake veiligheidsmachtigingen, aan de wet op het politieambt en aan de wet van 18 maart 2014 betreffende het politionele informatiebeheer, gepubliceerd op https://www.privacycommission.be/sites/privacycommission/files/documents/advies_13_2015.pdf.

[142] Zie randnummer 32 van voormeld advies.

[143] Artikel 5 1. (e) AVG

[144] Artikel 5.1. (c) AVG

invalshoeken toe te laten. Indien persoonsgegevens worden verwerkt in dit kader dient er echter steeds over te worden gewaakt dat de aanpak in andere landen (bijvoorbeeld het Twitter Research Access project in de VS [145]) zonder een gelijkaardige regeling van gegevensbescherming als in de EU niet blindelings wordt overgenomen.

De Commissie somt hierna enkele criteria op aan de hand waarvan de proportionaliteit van big data (én open data projecten, zie verder) kan worden beoordeeld.

a) Voor de beoordeling van de proportionaliteit bij bijvoorbeeld commerciële big data projecten of bij open data of archiveringsprojecten van overheidsdiensten kan het van belang zijn om na te gaan of er al dan niet sprake is van een fase van afdoende aggregatie of anonimisering. Een Small Cells Risicoanalyse (zie hiervoor onder Punt C) kan hierbij helpen. Gevallen zonder fase van aggregatie kunnen derhalve vlugger als disproportioneel worden beschouwd.

b) Het proportionaliteitsbeginsel betekent dat men bij verwerkingen ook aandacht dient te besteden aan de uitleesfrequentie [146] van (persoons)gegevens in functie van de benodigde functionaliteiten, bijvoorbeeld bij het verwerken van data van slimme (energie)meters of smartphone apps.

c) Een ander criterium om de proportionaliteit van (overheids)verwerkingen te beoordelen vindt men terug in de rechtspraak van het Europees Hof voor de Rechten van de Mens (EHRM) [147]. Het Hof schenkt bij het proportionaliteitsonderzoek aandacht aan de mogelijkheid van het inzetten van alternatieve, minder verregaande onderzoeksmethodes i.p.v. een "totale, alomvattende surveille". Naar deze rechtspraak kan ook verwezen worden bij de toepassing van data mining technieken zoals bij het digitaal patrouilleren (zie hiervoor), "predictive policing", en bij fraudeopsporing aan de hand van correlaties (knipperlichten). Ook bij de aanpak van fiscale fraude, sociale fraude, energiefraude,… moet de proportionaliteit van de gegevensverwerkingen mee in overweging worden genomen en moet er bekeken worden of dezelfde maatschappelijke opportuniteiten niet kunnen gerealiseerd worden met minder grootschalige methodes.

d) Modellen zoals "predictive policing" en fraudeopsporing aan de hand van correlaties (knipperlichten) zoals bij de aanpak van fiscale fraude, sociale fraude, energiefraude, betalingsfraude… in de publieke en de private sector dienen door de wetgever te worden behandeld als **buitengewone opsporingsmethodes** [148]**,** en dienen een gelijkaardig wettelijk kader te krijgen die de grenzen van deze methodes afdoende afbakent en waarborgt. Bij gebrek aan een **afdoende duidelijk wettelijk kader**[149]

[145] Zie ADAMS, R., All your Twitter belongs to the Library of Congress, The Guardian, 14 April 2010, gepubliceerd op https://www.theguardian.com/world/richard-adams-blog/2010/apr/14/twitter-library-of-congress; SCOLA, N., Library of Congress' Twitter archive is a huge #FAIL. More than five years on, the library's Twitter archive project is in limbo — with no end in sight, 7 November 2015, gepubliceerd op http://www.politico.com/story/2015/07/library-of-congress-twitter-archive-119698.html#ixzz4KJaJhExE

[146] Zie pagina 15 van VREG, advies van 8 april 2015 met betrekking tot een ontwerp van besluit van de Vlaamse regering tot wijziging van het Energiebesluit van 19 november 2010, wat betreft het plaatsen van slimme meters, gepubliceerd op http://www.vreg.be/sites/default/files/document/adv-2015-03_ontwerp_van_besluit_uitrol_slimme_meters.pdf

[147] EHRM, 2 september 2010, Uzun vs. Duitsland.

[148] Zie randnummer 14 van het advies 25/2016 van 8 juni 2016 betreffende het Ontwerp van decreet tot wijziging van het Energiedecreet van 8 mei 2009, wat betreft het voorkomen, opsporen, vaststellen en bestraffen van energiefraude.

[149] Verwijzing naar de "wet" in de materiële, ruime zin, dus inclusief eventuele machtigingen door sectorale comités, die een proportionaliteitsonderzoek van de aanvraag kunnen bevatten.

voor voormelde modellen, kan de naleving van het proportionaliteitsbeginsel bij deze modellen niet afdoende worden afgetoetst, en is er geen garantie dat de bijzondere onderzoekstechniek op een maatschappelijk verantwoorde wijze wordt ingezet. Belangrijk is wel dat deze wettelijke waarborgen op een zo generieke en technologieneutrale (dus zonder verwijzing naar specifieke technieken en methode uit de data mining die op een bepaald moment actueel of van toepassing kunnen zijn) manier worden geformuleerd.

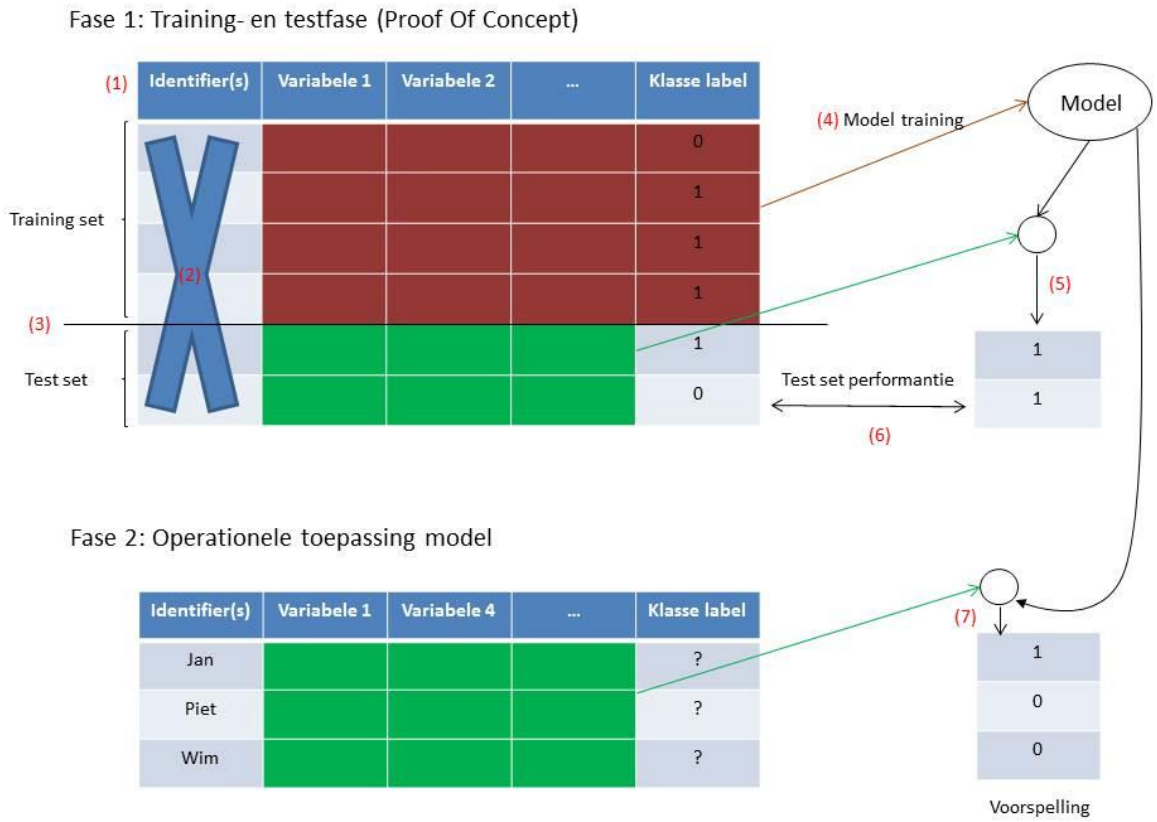
e) Bij het opstellen van predictieve modellen werkt men best in twee fasen:

➢ Aanbeveling 18

In geval van het opstellen en gebruik van voorspellende modellen ((semi-)supervised learning) werkt men best met twee fases (zie Figuur 1). In de eerste of voorafgaande fase kan men een mathematisch model opstellen en trainen aan de hand van zo anoniem mogelijke data waarbij op zijn minst de directe identificatoren zijn verwijderd. Verder kan men in deze fase alle data of variabelen (met inclusie van de klasse labels) die men ter beschikking heeft (en uiteraard op voorwaarde dat die rechtmatig worden verwerkt) gebruiken, opslagen en in overweging nemen voor het trainen of mogelijke inclusie in het uiteindelijke model. In een tweede fase (fase met "singling out" en/of identificatie van de betrokkene[150]) bij operationeel toepassen van het model dat in de vorige fase werd getraind, mogen enkel de gegevens/variabelen gebruikt, verzameld en opgeslagen worden waarvan in de trainingsfase is gebleken dat ze significant bijdragen in de voorspellingen van het model (beginsel van minimale gegevensverwerking).

Het werken met deze twee fases kan ondermeer gebeuren door de introductie van afzonderlijke rollen: analisten (verantwoordelijk voor het trainen en testen van de modellen) en operatoren (verantwoordelijk voor het gebruik van de operationele modellen in de beslissingsprocessen). Dit wil dus zeggen dat de analisten toegang hebben tot alle variabelen die men ter beschikking heeft maar niet in staat mogen zijn om de datapunten te heridentificeren. De operatoren kennen daarentegen wel de identiteit achter de datapunten maar hebben slechts toegang tot een beperkt aantal variabelen die geïncludeerd worden in de operationele modellen.

[150] Zie ook randnr 35 van het advies 25/2016 van 8 juni 2016



Figuur 1
Predictive modelling: verschillende fasen:

Fase 1: Training- en testfase

(1) Er wordt vertrokken van een dataset waar voor ieder individu (de rijen) zowel de inputvariabelen als de categorie (klasse) waar het individu toe behoort (bv. kredietwaardig of niet, aanwezigheid van een ziekte of niet, geïnteresseerd in een bepaald product of niet, …; in het voorbeeld werken we met twee categorieën voorgesteld door 1 of 0. Dit worden ook de klasse labels genoemd) gekend zijn. De bedoeling is om in deze fase een model te maken en te testen dat aan de hand van de inputvariabelen de categorie van een individu kan voorspellen. Hiervoor is het niet nodig om de identiteit van de personen in de data set te kennen.

(2) Verwijderen van identificatoren of gebruik van andere technieken voor anonimisatie. Tijdens deze eerste fase mag er geen poging worden ondernomen worden om de individuen te heridentificeren aan de hand van de nog resterende variabelen na anonimisatie.

(3) Opsplitsen van de data set in een training en testset. De trainingset zal gebruikt worden om het model op te stellen en de testset zal vervolgens gebruikt worden om de nauwkeurigheid van het model te kwantificeren. De gegevens in de testset mogen op geen enkele manier worden gebruikt bij het opstellen of trainen van het model (onafhankelijke testset). Zowel de training als de testset moeten representatief zijn voor de populatie in Fase 2 waarop men het model wil gebruiken om voorspellingen te maken.

(4) Training: opstellen van het model aan de hand van de gegevens in de trainingset.

(5) Testen: Het model wordt toegepast op de individuen van de testset. Voorspelling van de klasse labels van de personen in de testset.

(6) Vergelijking van de voorspelde waarde van de klasse labels met hun werkelijke waarde. Hierdoor kan de nauwkeurigheid van het model of de onafhankelijke testset performantie worden berekend.

Fase 2: Gebruik van het model in een operationele context op gegevens van nieuwe (geïdentificeerde) personen waar de categorie waartoe de individuen behoren nog niet gekend is. Hierbij mag er ook enkel toegang zijn tot de gegevens/variabelen waarvan in de trainingsfase is gebleken dat ze significant bijdragen in de voorspellingen van het model.

(7) Model wordt gebruikt om de klasse labels te voorspellen voor nieuwe personen en deze informatie wordt gebruikt om verdere beslissingen betreffende dit individu te sturen.

H. Finaliteitsbeginsel

Het gegevensbeschermingsrecht bevat de verplichting om de gegevens niet voor andere doeleinden te gebruiken dan deze waarvoor de gegevens initieel werden verkregen ("vereiste van verenigbaar gebruik"). Bij veel big data toepassingen komt dit beginsel op de helling te staan. Big data projecten gaan immers vaak uit van gebruik van gegevens voor een ander doel dan waarvoor de data initieel zijn verzameld (eerst verzamelen, later bekijken wat ermee gedaan kan worden), soms ook na koppeling met gegevens uit andere bronnen. Dergelijk hergebruik kan vaak een grote meerwaarde opleveren – vaak ook op maatschappelijk vlak - , zoals bv. bij het opsporen van zeldzame medische aandoeningen, of de aanpak van fiscale fraude of sociale fraude, maar staat wel op gespannen voet met de vereiste van verenigbaar gebruik.

Welke factoren kunnen van belang zijn om een latere verwerking te beschouwen als onverenigbaar met de initiële verwerking? Artikel 6.4 van de AVG bevat een niet-limitatieve lijst van factoren om deze vraag te beoordelen:

✓ **Aanbeveling 19**

De Commissie beveelt de verantwoordelijken aan om rekening te houden met volgende elementen indien de verenigbaarheid van verwerkingen wordt getoetst.

a) Het verband tussen de doelstellingen waarvoor de gegevens werden verwerkt, en de doeleinden die resulteren uit de big data analyse. De voorspelling van een bepaalde uitkomst staat niet altijd in verband met het initiële doeleinde. Wetenschappelijk onderzoek heeft bijvoorbeeld aangetoond dat er een mogelijkheid bestaat om gevoelige informatie te achterhalen uit het gebruik van de facebook like knop (zie hierna)[151]. Het klikken van een Facebook "like button" staat echter niet in relatie met het afleiden van gevoelige persoonsgegevens zoals drugsgebruik.

b) De context van de verzameling van de gegevens en de aard van de relatie tussen de betrokkene en de verantwoordelijke. De vraag of de nieuwe informatie die kan worden ontdekt aan de hand van een big data analyse (bv. risico op fraude) voldoet aan de vereiste van verenigbaar gebruik zal vaak afhangen van de wettelijke context waarin de verantwoordelijke werkt. Zo werken de sociale inspectie, een wetenschapper of een verzekeraar steeds in een specifieke reglementaire context die verschillend is, en die de mogelijkheid tot gebruik van nieuwe informatie zal bepalen. Als de reglementering afdoende voorzienbaar is op dat vlak en de nodige waarborgen biedt voor de betrokkene, zal er herbruikbaarheid mogelijk zijn (bv. verwerking voor wetenschappelijke doeleinde).

c) Het feit of al dan niet gevoelige persoonsgegevens worden verwerkt (zie hierna);

[151] YOUYOU, Wu, KOSINSKI Michal, STILLWELL, D., Computer-based personality judgments are more accurate than those made by humans, Department of Psychology, University of Cambridge, 27 Januari 2015, gepubliceerd op http://www.pnas.org/content/112/4/1036.full.pdf, aangehaald in NORTH, A., How Your Facebook Likes Could Cost You a Job, 20 januari 2015, gepubliceerd op New York Times blog, http://op-talk.blogs.nytimes.com/2015/01/20/how-your-facebook-likes-could-cost-you-a-job/?_r=0,

d) De mogelijke negatieve gevolgen (risico's) van de voorgenomen verdere verwerking voor de betrokkenen;

Een mogelijk voorbeeld is de situatie waar de (gemiddelde) betrokkene wordt blootgesteld aan een hoger risico dan initieel voorgesteld in de initiële doeleinden. Dit zou kunnen gelden voor een belastingadministratie die het toekennen van een eenmalige korting belooft tegen het meedelen van bijkomende informatie over het vermogen via een webapp (het initieel doeleinde van deze informatie is om de hoogte van de korting te berekenen) , waarna het vermogen met behulp van big data analyses op efficiëntere (en gemiddeld verhoogde) wijze wordt belast aan de hand van de verkregen informatie, zodat de gemiddelde gebruiker in totaal meer zal moeten betalen;

e) Het bestaan van passende controles en waarborgen (zoals eventueel versleuteling of pseudonimisering) of andere bijkomende garanties of sluitende afspraken tegen heridentificatie. Indien bijvoorbeeld in het kader van het opstellen van predictieve modellen een Proof Of Concept wordt ontwikkeld in een eerste fase (zie de laatste aanbeveling en Figuur 1 in Punt G, waar het model wordt opgesteld en de performantie ervan wordt getest op basis van zo anoniem mogelijke gegevens waarbij op zijn minst de directe identificatoren zijn verwijderd, of waarbij andere technieken voor anonimisatie zijn gebruikt, zie Punt C), en het model niet in volgende fase op een andere doelgroep van identificeerbare personen wordt toegepast (functionele scheiding of "functional separation"), kan dit een element zijn om de verenigbaarheid te motiveren. In Figuur 1 betekent dit dat de tweede fase (nog) niet wordt toegepast.

➢ Aanbeveling 20

Voor big data projecten is het van belang om op voorhand en zoveel mogelijk alle mogelijke doelstellingen te omschrijven voor latere verwerking(en). De mogelijke onderzoeksdomeinen en finaliteiten dienen met andere woorden zo duidelijk mogelijk en op voorhand te worden afgebakend.

➢ Aanbeveling 21

Verantwoordelijken die geen onmiddellijke operationele doelstellingen hebben ten aanzien van individuen en dus enkel globale data of resultaten wensen te verkrijgen die geen persoonsgegevens meer zijn (bv. geaggregeerde data of (de performantie van) een wiskundig model) in een big data project, kunnen hun - op zijn minst gepseudonimiseerde - persoonsgegevens laten verwerken door een derde (bv. een academische onderzoeksgroep) die afdoende waarborgen biedt dat een *wetenschappelijk* onderzoeksproject zal worden opgestart waarbij de persoonsgegevens na het einde van het project door de verwerker zullen worden vernietigd. Zowel artikel 4 § 1 2° WVP[152] als artikel 5 1. b) AVG voorzien hiertoe bijzondere bepalingen aangaande de verenigbaarheid van latere verwerkingen voor historische, statistische of wetenschappelijke doeleinden. In de mate er sprake is van wetenschappelijk onderzoek (dit wil zeggen in de mate dat er wordt voldaan aan de relevante

[152] Zie de artikelen 2 tot en met 24 van het Koninklijk besluit van 13 februari 2001 ter uitvoering van de wet van 8 december 1992 tot bescherming van de persoonlijke levenssfeer ten opzichte van de verwerking van persoonsgegevens, B.S., 13 maart 2001.

voorwaarden[153]), geen persoonsgegevens worden meegedeeld aan derden en voldoende garanties kunnen worden gegeven qua beveiliging van de persoonsgegevens, kan dit een afdoende waarborg vormen voor het respecteren van de vereiste van verenigbaar gebruik.
In de praktijk zijn tendensen waar te nemen die ingaan tegen een correcte toepassing van het finaliteitsbeginsel zoals:

a) **Finaliteitsvervaging** ("aanpak van fraude" wordt soms als zeer vage generieke finaliteit vermeld waarbij niet steeds duidelijk is wat "fraude" is, welke types en gradaties van fraude men wenst te bestrijden op welke (wettelijke) basis[154])
b) "**Window dressing**": de praktijk om niet alle finaliteiten juist te vermelden of een situatie anders voor te stellen naar de buitenwereld toe (zoals het voorstellen van een surveilleactiviteit als een beveiligingsmaatregel[155])
c) "**Function creep**" waarbij data die voor een bepaald doel werden verzameld (bv. facturatie energie en water aan de hand van informatie in het toegangsregister van distributienetbeheerders[156]) later wordt gebruikt voor andere doeleinden (bv. aanpak van energiefraude).

Een bijzondere toepassing vormt het hergebruik van "open data" voor "big data" projecten. Enerzijds is het in de wetenschap essentieel dat data beschikbaar is en kan worden gedeeld[157]. Anderzijds kan het publiekelijk beschikbaar maken van data door overheden en de private sector ook nieuwe risico's vormen voor de rechten en vrijheden van de betrokkenen. Hergebruik van open data voor big data projecten kan in dat geval een probleem vormen ten opzichte van de redelijke verwachtingen van de betrokkenen.

I. Legaliteitsbeginsel

Een goed wetgevend kader voor overheidsdiensten die gebruik maken van big data analyses of "online patrouilleren" is belangrijk. Er blijkt evenwel een "scheeftrekking" te bestaan in de wetgevende kaders. De aandacht is in de praktijk vooral gericht op het verzamelen en delen van gegevens (bv. focus op aanleg van een datawarehouse[158]). Het ontbreekt al te vaak nog

[153] Hoewel overweging 159 van de AVG vertrekt van een ruime interpretatie van dit begrip dat private financiering niet uitsluit, wordt in de AVG wel verwezen naar het voldoen aan specifieke voorwaarden, "met name wat betreft het publiceren of anderszins openbaar maken van de resultaten van het onderzoek". Een onderzoek is wetenschappelijk door de gebruikte methode (objectieve observaties en metingen, en analyse die gebruik maakt van verklarende statistiek), maar vergt ook een wetenschappelijke doelstelling om volwaardig deze titel te mogen voeren. Dit houdt in: een bijdrage willen leveren aan de wetenschappelijke kennis, door medewetenschappers aanvaarde publicaties doen enzovoort. Zie pagina 7 van het vademecum van de onderzoeker, gepubliceerd op https://www.privacycommission.be/sites/privacycommission/files/documents/vademecum-voor-de-onderzoeker_0.pdf .

[154] Volgens de Orde van Vlaamse balies is het ruime begrip "ernstige fiscale fraude, al dan niet georganiseerd" vaag omschreven in de wet van 11 januari 1993 en zijn de constitutieve bestanddelen ervan niet duidelijk te onderscheiden van de "eenvoudige fiscale fraude", waardoor er rechtsonzekerheid is alom. Zie http://www.ordeexpress.be/artikel/44/283/wet-houdende-dringende-bepalingen-inzake-fraudebestrijding
Zie ook de Beslissing 2004-499 van het Franse Grondwettelijk Hof. Deze beslissing betrof de intentie om een nieuw artikel 9 op te nemen in de Franse privacywet van 1978 teneinde de mogelijkheid te voorzien om gegevens te verwerken over "inbreuken, veroordelingen of veiligheidsmaatregelen" voor de preventie en de bestrijding van fraude. Dit ten gunste van rechtspersonen die slachtoffer zijn van inbreuken of die optreden voor rekening van slachtoffers. Het Hof overwoog hierbij "*Considérant que, s'agissant de l'objet et des conditions du mandat en cause, la disposition critiquée n'apporte pas ces précisions ; qu'elle est ambiguë quant aux infractions auxquelles s'applique le terme de « fraude » ; qu'elle laisse indéterminée la question de savoir dans quelle mesure les données traitées pourraient être partagées ou cédées, ou encore si pourraient y figurer des personnes sur lesquelles pèse la simple crainte qu'elles soient capables de commettre une infraction (…)*" .
Zie http://www.conseil-constitutionnel.fr/conseil-constitutionnel/francais/les-decisions/acces-par-date/decisions-depuis-1959/2004/2004-499-dc/decision-n-2004-499-dc-du-29-juillet-2004.904.html

[155] Zie de discussie rond de werkelijke doelstelling van de datr cookie van facebook. Volgens de Commissie gaat het om een surveille op gebruikers van websites, volgens facebook gaat het om een beveiligingsmaatregelen. Zie Facebook wint veldslag, maar privacystrijd is nog lang niet gestreden, de redactie, 29 juni 2016, http://deredactie.be/cm/vrtnieuws/cultuur%2Ben%2Bmedia/media/2.37414

[156] De distributienetbeheerders (EANDIS, INFRAX,…) houden de gegevens van aansluitingen op het gas- en elektriciteitsnet van (onder meer) huishoudelijke afnemers bij in het zogeheten toegangsregister. In dat register noteren ze de technische gegevens van elke aansluiting en de administratieve gegevens van de klant en zijn energieleverancier.

[157] de Montjoye Y. A., Unique in the shopping mall: On the re-identifiability of credit card metadata, Science vol 347, 30 January 2015; http://science.sciencemag.org/content/347/6221/536; 536-537.

[158] Zie bijvoorbeeld de Wet van 3 augustus 2012 houdende bepalingen betreffende de verwerking van persoonsgegevens door de Federale Overheidsdienst Financiën in het kader van zijn opdrachten, B.S. 24 augustus 2012

aan een algemeen kader voor de analyse en toepassing van de resultaten op personen[159] door inspectiediensten (waarborgen voor kwaliteitsvolle big data analyses en voorkomen van datadeterminisme, bewijswaarde van de vaststellingen, rechten van de betrokkenen dienaangaande, …).

➢ **Aanbeveling 22**

De wetgever zou meer aandacht dienen te schenken aan de verplichting voor het maken van kwalitatief hoogstaande big data analyses en de toepassing van big data analyses op natuurlijke personen door overheidsdiensten of andere diensten die een taak van algemeen belang uitvoeren. Dit kan gebeuren door rekening te houden met de elementen in dit rapport. Hierbij kan in het bijzonder worden verwezen naar hetgene dat werd aangehaald onder Punt D.1.

J. Legitimiteit / Mogelijkheden om big data analyses te beschouwen als een gerechtvaardigde verwerking

Big Data analyses zijn geen verwerkingen die a priori en op zichzelf (il)legaal zouden zijn, maar zijn pas legitieme verwerkingen van persoonsgegevens in de mate dat een legitiem doel wordt nagestreefd en een juiste balans wordt gevonden tussen de bescherming van privacy en persoonsgegevens enerzijds en de impact op de maatschappij anderzijds.

De WVP voorziet actueel in artikel 5 WVP enkele mogelijkheden. In de AVG wordt dit in Artikel 6 besproken. Bij sommige mogelijkheden kunnen evenwel in bepaalde omstandigheden problemen ontstaan om big data verwerkingen afdoende legitimiteit te verschaffen (opdat het gerechtvaardigde verwerkingen zouden zijn).

Voor de toestemming van de betrokkene is het met name de vraag of de betrokkene met kennis van zaken[160] kan toestemming geven voor een verwerking waarvan de werkelijke impact moeilijk te doorgronden kan zijn (zie ook verder onder Punt L welke sleutelinformatie transparant dient te worden bekendgemaakt) zelfs indien alle moeite wordt gedaan om de verschillende aspecten van de verwerking in zo begrijpelijk mogelijke taal uit te leggen. In het kader van big data projecten is het maar de vraag of dit laatste altijd helemaal mogelijk is en of het zelfs altijd mogelijk is om alle (neven)effecten op voorhand te voorzien.

De bescherming van privacy en van persoonsgegevens is niet absoluut en vergt vaak een belangenafweging ten opzichte van andere legitieme belangen van eerder algemeen belang (bv. wetenschappelijk onderzoek of de vrijheid van handel en nijverheid) of de gerechtvaardigde belangen van een verantwoordelijke, of van een derde. Deze belangen kunnen een rechtsgrond bieden voor verwerking, mits er een belangenevenwicht is met de betrokkene waarbij rekening werd gehouden met de redelijke verwachtingen van deze laatste. Ook moet worden bekeken of er geen nadelige invloed is op bepaalde bevolkingsgroepen (zie verder onder Punt E.2)[161].

De legitimiteit voor big data projecten motiveren aan de hand van een legitiem belang van een

---

[159] BROEDERS, D. en SCHRIJVERS, E., Big Data in een vrije en veilige samenleving, 20 mei 2016, http://njb.nl/njv-jaarvergaderingen/jaarvergadering-2016/artikelen/big-data-in-een-vrije-en-veilige-samenleving.19799.lynkx

[160] Overweging 42 AVG en de FAQ in verband met de toestemming op de website van de Commissie (https://www.privacycommission.be/nl/faq-page/10023#t10023n20123)

[161] G.H. Evers, In de schaduw van de rechtsstaat: profilering en nudging door de overheid, Computerrecht, 2016/98, juni 2016, p169.

verantwoordelijke of via toestemming van de klanten is in de praktijk niet evident. Dit gelet op de complexiteit, de mogelijke voorspellende kracht en de impact van bepaalde voorschrijvende big data modellen op de betrokkene. Het is ook essentieel om bijkomende waarborgen te voorzien die de maatschappelijke impact van big data analyses (zie ook punt E.2. hiervoor) inschat of desgevallend aanpakt. Het uitvoeren van een gegevensbeschermingseffectbeoordeling zal hierbij een aanzet zijn, in de mate men bij deze beoordeling verder kijkt dan het gerapporteerde aantal gevallen van problemen die individuen ondervinden ingevolge big data analyses.

➢ **Aanbeveling 23**

Indien de verantwoordelijke zich beroept op de toestemming, dan moet deze makkelijk zijn in te trekken en zal deze niet gelden bij een duidelijk machtsonevenwicht [162] tussen de betrokkene en de verantwoordelijke of verwerker, bv. omdat de verantwoordelijke de dominante (of enige) dienst in de markt levert. De verantwoordelijke zal moeten aantonen dat er geen machtsonevenwicht is of dat dit onevenwicht de toestemming van de betrokkene niet kan beïnvloeden.

K. Informatieplicht

De verantwoordelijke heeft in beginsel een informatieplicht onder artikel 9 WVP en de artikelen 13 en 14 AVG [163]. Zo moet de betrokkene worden geïnformeerd over specifieke en welomschreven punten waaronder het bestaan van profilering en de gevolgen daarvan[164].

Uitzonderingen op de informatieplicht zijn enkel in een beperkt aantal gevallen voorzien door de wetgeving (bv. witwaswetgeving, fraudebestrijding sociale zekerheid, verwerking voor politionele doeleinden, fiscaal onderzoek,…), …).

Het onderscheid tussen gegevensbeschermingsrecht en privacyrecht (onder meer de artikelen 8 EVRM en 22 Grondwet) is van belang omdat, ook als de informatieplicht of het recht van toegang niet geldt onder voormelde uitzonderingen, er toch een algemene transparantieplicht is onder het privacyrecht via de zogenaamde vereiste van "voorzienbaarheid" van de wetgeving[165] bij inmengingen in de persoonlijke levenssfeer. Het begrip wetgeving dient hierbij in een brede zin te worden begrepen (zie punt L hieronder).

L. Transparantie

Transparantie is een wettelijke plicht die volgt uit het privacyrecht. Transparantie wil zeggen dat informatie en communicatie in verband met de verwerking eenvoudig toegankelijk en begrijpelijk/duidelijk moeten zijn. De zogenaamde vereiste van "voorzienbaarheid" van de

[162] Volgens overweging 42 van de AVG mag de toestemming niet worden geacht vrijelijk te zijn verleend indien de betrokkene geen echte of vrije keuze heeft of zijn toestemming niet kan weigeren of intrekken zonder nadelige gevolgen. Zie ook Overweging 43 van de AVG die stelt dat toestemming geen geldige rechtsgrond biedt indien er sprake is van "*een duidelijke wanverhouding tussen de betrokkene en de verwerkingsverantwoordelijke*". Dit volgt uit de vereiste van de vrijheid van de toestemming in artikel 4 11) AVG. Zie ook pagina's 12-16 van Opinie 15/2011 van de Groep 29 van 13 juli 2011 over de definitie van toestemming, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187_en.pdf, en de FAQ op de website van de Commissie (https://www.privacycommission.be/nl/faq-page/10023#t10023n20123).

[163] Artikelen 13 en 14 AVG

[164] Overweging 60 AVG

[165] Artikel 8.2. EVRM, zoals hernomen door artikel 22 Grondwet en artikel 7 van het EU-Handvest

wetgeving is ook een klassieke privacyvereiste[166], die een transparantieverplichting inhoudt.

Big data analyses en gebruik van algoritmes vereisen dus in regel transparantie, ook als de verantwoordelijke vrijgesteld is van de informatieplicht (zie hiervoor).

➢ Aanbeveling 24

In de praktijk dient een onderscheid te worden gemaakt tussen het al dan niet (moeten) naleven van de **informatieplicht enerzijds** en de **transparantieplicht**[167] **anderzijds**. Transparantie dient betrekking te hebben op al de fasen van de gegevensverwerking (zie Deel 1) en meer specifiek op zowel de trainings- en testfase als de fase van de toepassing van de modellen (zie Figuur 1) in het kader van het opstellen van predictieve modellen.

Nochtans zijn de meeste big data analyses en toepassingen van algoritmes in de private en publieke sector vandaag vaak niet gebaseerd op een duidelijk wettelijk kader, voldoende duidelijke contractuele bepalingen of een extern privacybeleid. Algemene voorwaarden van financiële instellingen verwijzen vaak enkel naar zeer algemene termen en bewoordingen (de mogelijkheid om "profielen" op te maken en "data-analyses te verrichten"), terwijl de exacte aard en voorspellende kracht van de toepassingen vaak totaal niet duidelijk zijn.

Individuen worden met behulp van big data analyses dus steeds transparanter voor de verantwoordelijken, terwijl algoritmes vaak gekenmerkt worden door een gebrek aan transparantie of ondoorzichtigheid ("opacity") naar de betrokkenen en toezichthouders toe[168].

De ondoorzichtigheid van de broncode of van de gedetailleerde werking of specifiek gebruik van algoritmes kan soms wenselijk en zelfs noodzakelijk zijn. Deze ondoorzichtigheid kan immers de efficiëntie van fraudeonderzoeken waarborgen in het algemeen belang, de rechten of vrijheden van bepaalde partijen dienen (bv. versleutelingsalgoritmes), met inbegrip van het zakengeheim of de intellectuele eigendom en het auteursrecht dat de software beschermt[169]. In dit geval kan ervoor gekozen worden om bepaalde informatie enkel te delen met de toezichthouder.

De transparantieplicht is dus niet absoluut, maar maatwerk waarbij moet worden gezocht naar een goed evenwicht dat rekening dient te houden met andere belangen.

De Commissie pleit derhalve niet voor een absolute transparantie (bv. door systematische publicatie van de broncode of beschrijving van algoritmes zodanig dat ze volledig reproduceerbaar zijn), maar wel voor een goed evenwicht tussen geheimhouding (dat dan een legitiem of maatschappelijk belang moet dienen) enerzijds, en transparantie anderzijds. Het verschaffen van zoveel mogelijk "sleutelinformatie" (zie hierna) kan helpen om dit evenwicht te bereiken.

---

[166] Artikel 8.2. EVRM, zoals hernomen door artikel 22 Grondwet en artikel 7 van het EU-Handvest.

[167] Soms schenkt men enkel aandacht aan de informatieplicht en de uitzonderingen hierop onder de dataprotectiewetgeving en vergeet men de transparantieplicht onder de privacywetgeving. Indien bijvoorbeeld banken of verzekeringsinstellingen vrijgesteld zijn op de informatieverplichting onder de witwaswetgeving, betekent dit niet dat de wetgeving aangaande de dataprocessen totaal ontransparant mag zijn.

[168] BURRELL, Jenna, How the machine 'thinks': Understanding opacity in machine learning algorithms, January 2016, http://bds.sagepub.com/content/3/1/2053951715622512; pagina 8 (challenge 2) van Executive Office of the President, Big Data : A Report on Algorithmic Systems, Opportunity, and Civil Rights, Ma 2016, gepubliceerd op https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[169] Overweging 63 AVG

Ongepaste ondoorzichtigheid of een gebrek aan transparantie kan echter ook worden misbruikt om wetgeving inzake privacy, dataprotectie, consumentenbescherming of competitie te omzeilen door de wettelijke bescherming of rechten van de betrokkenen af te bouwen of betekenisloos te maken (zie verder ook de Punten N, O en P i.v.m. de rechten van de betrokkenen). Indien klanten of burgers maar een beperkte uitleg krijgen over de (kwaliteit van de) logica die achter bepaalde beslissingen ligt die met betrekking tot henzelf worden genomen, is het ook zeer moeilijk voor hen om suboptimale (of foute) beslissingen te herkennen waardoor ze in de feiten hun rechten (bv. recht op verbetering) niet kunnen laten gelden.

Voor het behoud van een controlemogelijkheid van de betrokkenen op hun persoonsgegevens[170] is het zeer belangrijk dat er een minimaal niveau van transparantie wordt gehandhaafd door verantwoordelijken die big data analyses gebruiken, zeker als er sprake is van (voorspellende of voorschrijvende) modellen die een hoge impact kunnen hebben op de rechten en vrijheden van de betrokkenen. Dit moet hen in staat stellen om zich indien nodig te (laten) verdedigen tegen beslissingen die over hen worden genomen. Hieronder worden enkele elementen besproken die bij voorrang transparant moeten beschikbaar gesteld worden.

L.1. Transparantie van "sleutelinformatie" bij big data analyses

➢ Aanbeveling 25

Indien big data analyses een inmenging vormen in de persoonlijke levenssfeer (doordat bv. voorspellingen worden verricht betreffende personen of men gedrag probeert te beïnvloeden), dan dienen deze inmengingen afdoende voorzienbaar te zijn. Na afweging van andere maatschappelijke belangen is de Commissie van mening dat deze voorzienbaarheid kan worden geboden door transparantie van de hieronder beschreven basiselementen of sleutelinformatie[171] ten aanzien van de betrokkenen.

Zoals hierboven reeds uitgelegd, betekent dit dus niet dat bijvoorbeeld modellen (bv. kredietscoringsmodellen) in alle gevallen moeten worden gepubliceerd, en acht de Commissie dit niet onverenigbaar met de rechten of vrijheden van anderen, met inbegrip van het zakengeheim of de intellectuele eigendom en met name van het auteursrecht dat de software beschermt [172]. In de milieu- en energiewetgeving wordt het publiceren van essentiële sleutelinformatie op een product of dienst (bv. een energielabel op een koelkast) door de fabrikanten ook niet uitgelegd als een inbreuk op hun intellectueel eigendomsrecht, maar als een opportuniteit voor innovatie en diversificatie van hun producten.

a) Doelstelling en verantwoordelijkheid

Big data analyses en het maken van de keuzes in verband met een model vereisen steeds een menselijke betrokkenheid. Het is voor de toepassing van de privacy- en dataprotectiereglementering ook essentieel dat de doeleinden van het project bepaald worden en bekend zijn. De aanspreekbare contacten die hiervoor (mede)verantwoordelijk zijn en

[170] Zie pagina 8 van de EDPS Opinie van 19 November 2015, Meeting the challenges of big data, A call for transparency, user control, data protection by design and accountability, gepubliceerd op https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf

[171] DE BOT, D. en DE HERT, P., Artikel 22 Grondwet en het onderscheid tussen privacyrecht en gegevensbeschermingsrecht. Een formele wet is niet altijd nodig wanneer de overheid persoonsgegevens verwerkt, maar toch vaak, CDPK, 2013, 358.

beslissingen kunnen (bij)sturen (bijvoorbeeld de functionaris voor gegevensbescherming ("DPO")[173]) , dienen publiek gemaakt te worden. Zonder deze transparantie kan er ook geen verantwoordingsplicht ("accountability") zijn, en kan de verantwoordelijkheid voor big data analyses niet worden toegewezen.

b) De gebruikte data

Een beschrijving van de (persoons)gegevens die worden aangewend voor het trainen, testen en operationeel gebruik van een model (zie ook Figuur 1) dienen te worden beschouwd als sleutelinformatie[174]. Hierbij dient op een zo reproduceerbare mogelijke wijze te worden weergegeven hoe de data werd bekomen en wat de kenmerken hiervan zijn. Het gaat hierbij niet enkel om de gekozen databronnen en hoe de gegevens werden verzameld (via enquêtes, online of mobiele toepassingen, open data, sociale media, koppelingen tussen verschillende databronnen, …) maar ook over de verzamelde variabelen en hoe die werden gemeten, de definitie en bepaling van de klasse labels, de eventuele pre-processing, de wijze waarop de punten in de training- en testset werden "getrokken" uit de totale populatie waarover men voorspellingen wenst te maken, de datakwaliteit, onzekerheden, ouderdom van de data, de frequentie van hertrainen, … Hierbij dient men ook weer te geven hoe men de nodige voorzorgen heeft genomen zodat de training- en testdata representatief zijn (en blijven) voor de doelgroep waarvoor men voorspellingen wil doen of beslissingen wil nemen.

c) Het Model

De performantie of nauwkeurigheid van het model[175] (zie Figuur 1) van het model zoals getest op onafhankelijke en representatieve testdata dient ter beschikking te worden gehouden van de betrokkenen[176].

Het is ook cruciaal dat transparantie wordt verschaft aangaande de belangrijkste methodologische keuzes, zoals de keuzes qua (trainings)algoritmen en modelstructuur, de wijze van bepalen van eventuele (hyper)parameters (zoals bijvoorbeeld de cut-off van het model[177]), de variabelen of features die in overweging worden genomen voor inclusie in het model en de hoe de variabelen/features in het definitieve model werden geselecteerd, … .

Best wordt de beschrijving van de onderliggende beslissingsalgoritmen zelf (dit is dus het uiteindelijke model dat na toepassen van de gekozen methodologie gebruikt wordt in Fase 2 van Figuur 1 - het gaat hier dus over de logica met de concrete input-output relatie) ook ter beschikking gesteld indien dit praktisch mogelijk is.

Deze informatie moet de gegevensverwerkingen voor derden (toezichthouders, betrokkenen,…) zo reproduceerbaar mogelijk maken en hen in staat stellen om zich eventueel te (laten) verdedigen tegen suboptimale methodologische beslissingen of weinig betrouwbare modellen of analyses. Dit betekent niet noodzakelijk dat men de gebruikte methodologie en modellen altijd volledig reproduceerbaar (bv. via de broncode) en in alle detail dient publiek te maken. Hierbij moet zoals gezegd een afweging worden gemaakt met andere legitieme

[173] Artikelen 37 en volgende AVG.

[174] Zie randnummers 29 tem 33 van het advies 45/2013 van 2 oktober 2013 betreffende het Waals landbouwwetboek en randnummer 15 van advies 08/2005 van 25 mei 2005 betreffende het voorontwerp van wet betreffende de analyse van de dreiging.

[175] Bijvoorbeeld aantal true positives, true negatives, false positives en false negatives (waaruit dan andere performantiematen kunnen worden bepaald zoals de accuracy, sensitiviteit, specificiteit, …), de oppervlakte onder de ROC-curve, …

[176] In de literatuur wordt de vergelijking gemaakt met het publiceren van de resultaten van crash tests van wagens, hetgeen niet noodzakelijk betekent dat men moet publiek maken hoe de wagen is gebouwd. Zie p 59 van DIAKOPOULOS, N., Accountability in Algorithmic Decision Making, Februari 2016, Communications ACM, http://www.nickdiakopoulos.com/wp-content/uploads/2016/03/Accountability-in-algorithmic-decision-making-Final.pdf

[177] Dit is de grens die men gebruikt op de (continue) output van het model om een persoon aan een van de twee klassen toe te kennen

belangen zoals commerciële en intellectueelrechtelijke belangen, bescherming van bedrijfsgeheimen, knowhow en de beveiliging van de verwerking. Gevoelige of kritische informatie aangaande de informatica-infrastructuur of het bedrijfsnetwerk bijvoorbeeld kan omwille van beveiligingsrisico's worden uitgezonderd van publieke transparantie. Desgevraagd dient de verantwoordelijke naar de toezichthouders wel altijd volledige transparantie (met alle details zodat de verwerking volledig kan worden gereconstrueerd = volledige reproduceerbaarheid) te verschaffen hoe men tot bepaalde uitkomsten komt.

De hierboven beschreven elementen, en meer in het bijzonder de concrete input-output relatie die aan de basis ligt van een beslissing of voorspelling, dienen ook (zie ook Punt J hierboven) in een zo begrijpbaar mogelijke taal uitgelegd aan de betrokkenen. Eventueel kan er hier worden gekozen voor een gelaagde aanpak op een website waar eerst kernelementen (korte informatie die sterk is samengevat) worden aangeboden waarna kan doorgeklikt worden indien men geïnteresseerd is in meer diepgaande informatie.

d) De gevolgen

Er moet steeds duidelijk worden gemaakt dat een beslissing is gebaseerd (mede) op basis van big data analyses (bijvoorbeeld indien predictieve modellen werden gebruikt).

De te voorziene gevolgen voor de (verschillende categorieën) van betrokkenen bij het gebruik of een bepaalde beslissing van het model dienen transparant te zijn (bv. verhogen of verlagen van een verzekeringspremie[178], afleiden of voorspellen van bepaalde informatie, …)

Het resultaat van de continue risicoafweging (zie hiervoor onder Punt B i.v.m. de AVG) dient te worden gepubliceerd, hetgeen niet impliceert dat alle informatie in dit kader publiek moet worden gemaakt.

Voor "open data" is het belangrijk dat de verantwoordelijke, vooraleer "open data" ter beschikking wordt gesteld, ook transparant is over haar onderzoek naar de kans op heridentificatie en waarbij men extern een zicht kan krijgen op de uniciteit van de datasets en de mogelijkheid voor derden (bv. handelsinformatiebedrijven of data brokers) om data af te zonderen tot op individueel niveau (zie ook onder Punt C).

M. Bescherming van gevoelige persoonsgegevens

"Correlaties" maken met of op basis van gevoelige persoonsgegevens (zoals gegevens betreffende ras, etnische afkomst, politieke overtuiging, godsdienst of levensbeschouwelijke overtuigingen, lidmaatschap van een vakbond, genetische of gezondheidsstatus, of seksuele gerichtheid[179], of over strafrechtelijke veroordelingen en strafbare feiten of daarmee verband houdende veiligheidsmaatregelen[180]) is in principe verboden tenzij een wettelijke uitzondering voorhanden is en waarborgen worden voorzien[181]. Gelet op het complexe, dikwijls niet transparante karakter van big data analyses (zie hiervoor onder Punt J) zal de vrije, specifieke, op voorafgaande informatie berustende en expliciete "toestemming" van de betrokkene niet

[178] Simon, F; insurance exec: Big data allows better understanding of risk, 9 Juni 2016, http://eurac.tv/25GO
[179] Overweging 51 en 71 AVG
[180] Onder artikel 10 AVG wordt deze striktere definitie gehanteerd ten opzichte van de veel ruimere omschrijving in artikel 8 WVP.
[181] Zie artikel 9 lid 2 AVG; Een voorbeeld hiervan vormt een onderzoek van de Universiteiten van Antwerpen en Amsterdam naar het twittergedrag binnen een bepaalde sociale cirkel gecorreleerd met het stemgedrag. VAN GYSEL, C., GOETHALS, B. en DE RIJKE, M., Determining the Presence of Political Parties in Social Circles, beschikbaar op https://staff.fnwi.uva.nl/m.derijke/content/publications/icwsm2015-sp-christophe.pdf. Dit verschilt van de situatie waarin private handelsonderzoeksbureau dergelijke onderzoeken zouden verrichten, waar geen wettelijke uitzondering voorhanden is.

steeds voorhanden zijn als uitzonderingsgrond[182] om het maken van correlaties met of op basis van gevoelige persoonsgegevens te legitimeren.

Big data analyses kunnen ook een **transformatieve impact** hebben op persoonsgegevens. Dit betekent dat in beginsel niet-gevoelige persoonsgegevens of ogenschijnlijk onschadelijke gegevens potentieel kunnen worden omgezet in (een correlatie met) gevoelige persoonsgegevens. Uit publiek beschikbare facebook "likes" zou bijvoorbeeld een computermodel kunnen worden opgesteld die op accuratere wijzere de persoonlijkheid van een persoon kan beschrijven dan mogelijk is voor vrienden, collega's, echtgenoten of familie[183]. Ook seksuele oriëntatie, etniciteit, religieuze en politieke overtuigingen en het gebruik van verslavende middelen zoals alcohol, tabak en drugs kan automatisch uit facebook likes worden afgeleid[184]. Big data analyses van (een combinatie van) dagdagelijkse gegevens zoals gebruik van bepaalde verzorgingsproducten, maat van aangekochte kleding, locatiegegevens, financiële transacties, … kunnen dikwijls met succes een correlatie aantonen met bijvoorbeeld gezondheidsgegevens (bv. aanwezigheid van ziekte of zwangerschap, kans op overlijden, …) of een criminaliteitsfenomeen.

Wat precies wel of niet wordt gedaan onder de noemer "big data analyses", "data mining", "profilering", "verwerking voor statistische doeleinden",… met de dagelijkse digitale sporen (bv. analyse van digitale sporen door een betaalapplicatie, de analyse van de aankoop van goederen door een supermarkt, online surf- en koopgedrag, …) wordt steeds belangrijker indien we bepaalde gevoelige persoonlijkheidskenmerken (bv. psychologisch profiel of seksuele geaardheid) wensen te beschermen. Weerom zal transparantie cruciaal zijn.

✓ **Aanbeveling 26**

Het verwerkingsverbod van gevoelige persoonsgegevens impliceert ook een verbod om de bescherming van gevoelige persoonsgegevens te omzeilen door gebruik van (transformatieve) big data analyses ten aanzien van niet-gevoelige persoonsgegevens, zeker indien de betrokkene koos om deze kenmerken niet mee te delen.

✓ **Aanbeveling 27**

In de publieke sector zou de wetgever moeten voorzien in een verhoogd toezicht van zodra er sprake is van correlaties met of op basis van gevoelige persoonsgegevens (bv. preventie of detectie van criminaliteit, fraude). Dit verhoogd toezicht kan de vorm aannemen van een voorafgaande machtiging door een bevoegde toezichthouder (bv. het Controle Orgaan voor het Beheer van de Politionele Informatie (COC)).

[182] artikel 8 WVP voorziet niet de schriftelijke toestemming van de betrokkene als uitzonderingsgrond in tegenstelling tot de artikelen 6 en 7 WVP, en artikel 9.2 verstrengt de toestemmingsvereiste in de AVG tot een "uitdrukkelijke" toestemming

[183] YOUYOU, Wu, KOSINSKI Michal, STILLWELL, D., Computer-based personality judgments are more accurate than those made by humans, Department of Psychology, University of Cambridge, 27 Januari 2015, gepubliceerd op http://www.pnas.org/content/112/4/1036.full.pdf, aangehaald in NORTH, A., How Your Facebook Likes Could Cost You a Job, 20 januari 2015, gepubliceerd op New York Times blog, http://op-talk.blogs.nytimes.com/2015/01/20/how-your-facebook-likes-could-cost-you-a-job/?_r=0,

[184] KOSINSKI Michal, STILLWELL, D. en GRAEPEL, T., Private Traits and attributes are predictable from digital records of human behaviour, Free School lane, The Psychometrics Centre, University of Cambridge, 9 April 2013, gepubliceerd op http://www.pnas.org/content/110/15/5802.full.pdf

N. Beperkingen en passende maatregelen bij geautomatiseerde beslissingen

Informatietechnologie speelt een steeds grotere rol in beslissingsprocessen en neemt meer en meer de plaats in van de mens [185] . Door de toegenomen informatisering van beslissingsprocessen bestaat er een tendens om het belang en de rol van de werkelijke en bewuste tussenkomst van individuen in beslissingsprocessen te doen afnemen. Wanneer dergelijke geautomatiseerde besluiten een duidelijke impact hebben op personen, bepalen de AVG en de Europese Rechtspraak[186] grenzen.

De artikelen 12bis WVP en 22.1 van de AVG bepaalt immers dat de betrokkene het recht heeft om *"niet te worden onderworpen aan een uitsluitend op geautomatiseerde verwerking, waaronder profilering, gebaseerd besluit waaraan voor hem rechtsgevolgen zijn verbonden of dat hem anderszins in aanmerkelijke mate treft."*

Er is geen absoluut verbod op geautomatiseerde beslissingen, daar artikel 22.2 AVG uitzonderingsgevallen voorziet, zoals de expliciete toestemming van de betrokkene, een noodzakelijkheid voor de totstandkoming of uitvoering van een overeenkomst met de betrokkene, of een profielgebaseerde besluitvorming die gebaseerd is op een verplichting onder het recht van de EU of een lidstaat met passende waarborgen. Een voorbeeld is de melding door een bank aan de witwascel, waarbij het transactierisico werd bepaald op basis van de klantenonderzoek (zgn. "customer due diligence") verplichtingen onder de artikelen 7 en volgende van de wet van 11 januari 1993 tot voorkoming van het gebruik van het financiële stelsel voor het witwassen van geld en de financiering van terrorisme[187].

Een zichtbare tussenkomst van een fysieke persoon in de besluitvorming is niet steeds voldoende. Een aantoonbare autonome en menselijke beoordeling bij gebruik van ICT-technologie in de beslissingsprocedure is een minimumvereiste, hetgeen volgt uit overweging 71 AVG en artikel 22.3 AVG.

✓ Aanbeveling 28

Aantoonbare afwijkingen genomen door fysische personen dienen voorhanden te zijn bij de systematische toepassing van het resultaat van big data analyses die een merkelijke impact hebben voor individuele rechten van datasubjecten.

Geautomatiseerde besluitvorming en profilering gebaseerd op bijzondere categorieën van persoonsgegevens ("gevoelige persoonsgegevens") mogen volgens de AVG uitsluitend worden toegestaan indien specifieke voorwaarden vervuld zijn[188], zoals het aanwezig zijn van een

[185] Een fictief maar realistisch voorbeeld betreft het gebruik van software door een hoogleraar om de kans op plagiaat te beoordelen in een werkstuk van een student. In de mate geen afdoende rekening wordt gehouden met de inherente foutenmarge van een doorgedreven gebruik van deze software en de mogelijkheid van de student om gehoord te worden bij een plagiaatsbeslissing, kan dit in de praktijk soms leiden tot een geautomatiseerde beslissing met een merkelijke impact op de betrokkene.

[186] Hof van Justitie, "Opinie 1/2015 van 26 juli 2017. In de paragrafen 131, 168 en volgende van deze opinie wordt gewezen op de voorafbepaalde modellen en criteria met een "significante foutenmarge" en de keuze van de databanken voor bestandsvergelijking van de geautomatiseerde PNR analyses, die de mate van inmenging in de persoonlijke levenssfeer van de betrokkene bepalen. Het Hof wenste dat de de betrouwbaarheid en actualiteit van deze vooraf bepaalde modellen, criteria en databanken zou deel uitmaken van het opvolgend "joint review" toezicht onder de PNR Overeenkomst, teneinde de noodzakelijkheid en het niet discriminatoir karakter te bewaken (§ 174).

[187] wet van 11 januari 1993 tot voorkoming van het gebruik van het financiële stelsel voor het witwassen van geld en de financiering van terrorisme, zoals laatst gewijzigd bij de wet van 13 maart 2016, B.S., 9 februari 1993.

[188] Overweging 71 AVG

uitdrukkelijke toestemming, het verschaffen van specifieke informatie aan de betrokkene[189] en/of het verlenen van een (aangepast) recht van toegang (zie ook Punt O). Concreet zou moeten worden bepaald wat de rechten van toegang, verbetering en verzet inhouden bij big data analyses: heeft de betrokkene al dan niet toegang tot het algoritme of de essentiële elementen van big data analyses (zie punt L.1 hiervoor)? Indien deze concrete toelichting ontbreekt, dreigen deze rechten dode letter te blijven. Dit kan bijvoorbeeld door een versterkte nadruk op informatieverplichting te voorzien naar analogie met de "e-privacy" regels in de artikelen 122 en 123 van de wet elektronische communicatie voor verkeers- en geolokatiegegevens.

O. Rechten van toegang en verbetering, recht op gegevenswissing

De betrokkene heeft een inzagerecht ten aanzien van de hem betreffende persoonsgegevens [190] , en het recht om van de verantwoordelijke voor de verwerking rectificatie[191] van hem betreffende onjuiste persoonsgegevens te verkrijgen. Onder de AVG moet de verwerkingsverantwoordelijke hierop "onverwijld" ingaan, en in ieder geval binnen een maand na ontvangst van het verzoek. Verlenging van deze termijn is mogelijk met twee maanden voor complexe verzoeken en in functie van het aantal verzoeken dat de verantwoordelijke ontvangt[192]. De verwerkingsverantwoordelijke stelt de betrokkene binnen één maand na ontvangst van het verzoek in kennis van een dergelijke verlenging.

Artikel 15.1 h) van de AVG bepaalt dat de inzage niet enkel betrekking heeft op het bestaan van geautomatiseerde beslissingen, maar ook op nuttige informatie over de onderliggende logica, alsmede het belang en de verwachte gevolgen van die verwerking voor de betrokkene. Zie in dit kader ook de aanbevelingen in dit rapport onder Punt L.1 (Transparantie van sleutelinformatie).

Ook kunnen de woorden "ten minste in die gevallen"[193] in artikel 15.1 h) van de AVG volgens de Commissie niet zo worden geïnterpreteerd dat het enkel gaat om de verwerking van gevoelige persoonsgegevens en gevallen waar er een duidelijke impact is op de betrokkenen. In geval er bijvoorbeeld sprake is van een risico op oneerlijke verwerking of discriminatie of waar het niet verschaffen van deze informatie de wel verschafte informatie misleidend of nietszeggend zou maken (bv. waar de verantwoordelijke enkel meedeelt dat verwerking gebeurt "voor statistische doeleinden", zonder de impact van de beslissingen toe te lichten), dient volgens de Commissie ook altijd de informatie in artikel 15.1 h) van de AVG te worden verstrekt.

Artikel 17 AVG voegt een recht op gegevenswissing ("recht op vergetelheid") toe. In dat verband is het relevant dat er aanwijzingen zijn dat bepaalde big data raamwerken gegevens sequentieel wegschrijven (als methode om data efficiënt te stockeren), voor gevolg kunnen hebben dat het langer duurt voordat de gegevens fysiek kunnen worden gewist[194]. In het kader van het recht op gegevenswissing kan dit een probleem zijn wanneer de wissing niet kan worden verzekerd "zonder onredelijke vertraging"[195].

---

[189] Zie overweging 71 AVG, artikel 13.2. f) AVG en 14.2 g) AVG betreffende de informatieplicht

[190] artikel 10 WVP en 15 AVG

[191] artikel 12 WVP en 16 AVG

[192] Artikel 12.3 AVG.

[193] In de gevallen van een door de AVG toegelaten besluit dat uitsluitend gebaseerd is op geautomatiseerde verwerking, waaronder profilering, waaraan voor hem rechtsgevolgen zijn verbonden of dat hem anderszins in aanmerkelijke mate treft. Hetzelfde geldt voor de bijzondere categorieën van persoonsgegevens waarbij afdoende waarborgen werden voorzien (de gevallen in artikel 22.1 en 22.4 AVG).

[194] Zie https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (zie punt "space reclamation").

[195] Artikel 17.1 AVG

P. Recht van verzet

Artikel 21.2 van de AVG bepaalt dat, wanneer persoonsgegevens ten behoeve van direct marketing worden verwerkt, *"de betrokkene te allen tijde het recht (heeft) bezwaar te maken tegen de verwerking van hem betreffende persoonsgegevens voor dergelijke marketing, met inbegrip van profilering die betrekking heeft op direct marketing"*

De vraag kan worden gesteld wat het recht van bezwaar of verzet concreet betekent bij big data analyses. Uit voormeld artikel 21 van de AVG kan alvast geen algemeen recht worden ontleend om niet te worden onderworpen aan profilering of big data analyses (zie ook onder Punt N). De vraag of er een recht van verzet is zal afhangen van het soort verwerking dat wordt nagestreefd waarbij er moet bekeken worden welke gerechtvaardigde belangen het zwaarst doorwegen (bv. direct marketing vs. wetenschappelijk onderzoek; Profilering door overheidsdiensten met een wettelijke of reglementaire basis vs. profilering door een commerciële onderneming).

Het voorafgaande zal alvast niet verhinderen dat sommige beslissingen die gebaseerd zijn op resultaten van big data analyses of bestandsvergelijking door de betrokkenen zullen kunnen worden gepercipieerd als incorrect, onrechtvaardig of discriminerend (bijvoorbeeld het ontzeggen van een toegang tot een festival op basis van een bestandsvergelijking).

✓ **Aanbeveling 29**

Omwille van de moeilijkheid van de toepassing van het recht van verzet lijkt de betrokkene soms aangewezen op de mogelijkheid van een procedure bij de rechter in kortgeding. Het is echter ook van belang om een mogelijkheid tot beroep te voorzien bij een bevoegde autoriteit tegen beslissingen die (mede) gebaseerd zijn op big data analyses, waaronder een beroep bij de bevoegde toezichtsorganen of autoriteiten die bevoegd zijn om toe te zien op de bescherming van klanten en consumenten in bijzondere sectoren (bv. telecom- en financiële sector) of op bijzondere wetgeving (bv. de naleving van de antidiscriminatiebepalingen). Er kan in dit kader ook worden gedacht aan vertegenwoordiging van de betrokkenen (artikel 80 GDPR) of (voor een groep van betrokkenen die consument zijn en schade lijden) aan de mogelijkheid van een rechtsvordering tot collectief herstel[196].

Q. Verantwoordelijkheid voor de verwerking

Bij big data analyses is in toenemende mate sprake van een groot aantal actoren op nationaal en/of internationaal vlak, waarbij de verantwoordelijkheidstoewijzing complexer wordt naargelang meer partijen verschillende taken opnemen in de private en/of publieke sector[197], zoals het verzamelen en leveren van gegevens[198], leveren van soft- en/of hardware, … Om te bepalen of een partij verwerkingsverantwoordelijke is, kan de Commissie kan in dat verband rekening houden met alle relevante factoren, vermeld in de definitie van verwerkingsverantwoordelijke (artikel 4 7)) AVG).

---

[196] op basis van Artikel XVII.36 en 37 van het Wetboek Economisch Recht ("WER").

[197] VAN DER SLOOT, B. en VAN SCHENDEL, S. / Wetenschappelijke Raad voor het Regeringsbeleid, International and Comparative Legal study on Big Data, p 43, http://www.wrr.nl/fileadmin/en/publicaties/PDF-Working_Papers/WP_20_International_and_Comparative_Legal_Study_on_Big_Data.pdf

[198] Zie aangaande de rol van data brokers in de werking van facebook volgend artikel : DEWEY, C., 98 personal data points that facebook uses to target ads to you, The Washington Post, 19 augustus 2016, gepubliceerd op https://www.washingtonpost.com/news/the-intersect/wp/2016/08/19/98-personal-data-points-that-facebook-uses-to-target-ads-to-you/

✓ **Aanbeveling 30**

In de reglementering en contractuele bepalingen dient (dienen) de verantwoordelijk(en) voor de big data verwerkingen steeds duidelijk te worden vermeld.

R. Handhaving en toezicht op verwerkingen

Grootschalige onderschepping (in het kader van toezicht of surveille op personen) van gegevens die betrekking kan hebben op alle mogelijke gebruikers van een dienst vormt een inmenging in de persoonlijke levenssfeer en vereist steeds een onafhankelijke controle[199] om de noodzakelijkheid te bewaken. Er kan hierbij worden verwezen naar projecten en toepassingen in de publieke sector ("predictive policing"), de private sector ("deep packet inspection" door telecomoperatoren)[200], en de verruimde privatisering van de preventie van fraude[201] en criminaliteitspreventie.

Een hoog risico voor de rechten en vrijheden van natuurlijke personen kan bv. ook bestaan in het gebruik van algoritmes bij toezicht door werkgevers, of bij algoritmes gebruikt in een financiële smartphone-app voor minderjarigen[202].

Naast een voorafgaande machtiging door een onafhankelijk toezichtsorgaan[203] (in de publieke sector) kunnen de volgende waarborgen worden aangehaald:

- Het voorzien van audits door onafhankelijke organisaties ten aanzien van beslissingssystemen die (mede) gebaseerd zijn op algoritmes.
- Het voorzien van systemen van interactieve modellering (feedback door de betrokkenen wiens gegevens worden verwerkt door algoritmes, en die het algoritme kan bijsturen – zie ook Punt E.2).
- Het voorzien van een onafhankelijke organisatie ("Trusted Third Party" - TTP) die een neutrale analyse van de gegevens kan uitvoeren/waarborgen volgens de regels van de kunst (bv. door wetenschappelijk onderzoek), in plaats van data-analyse door derden die kunnen gevestigd zijn in landen zonder een passend niveau van gegevensbescherming[204] en/of financiële belangen kunnen hebben bij het verkoop van data-analyseoplossingen.
- Versterking van de rol van de onafhankelijke toezichthouders. Bij sterke multinationale dataconcentraties (overnames in de digitale sector) opperde de EDPS het idee van een

---

[199] Zie bij analogie de rechtspraak aangaande grootschalige interceptie van telecommunicatie door de Duitse veiligheidsdiensten, aangehaald in EHRM, 26 juni 2006, Weber & Saravia t. Duitsland.

[200] Zie de aanbeveling 05/2012 van 11 april 2012 van de Commissie, gepubliceerd op https://www.privacycommission.be/sites/privacycommission/files/documents/aanbeveling_05_2012_0.pdf

[201] Zie de inschakeling van distributienetbeheerders in de aanpak van sociale fraude en energiefraude.

[202] Zie de overwegingen 38 en 75 AVG

.5 AVG, voor zover voorzien in het nationaal recht.

[203] Artikel 36.5 AVG.

[203] In toepassing van artikel 36.5 AVG, voor zover voorzien in het nationaal recht.

[204] Zie de mediaberichten van september 2016 aangaande mogelijkheid van participatie door een Chinees staatsbedrijf in EANDIS, en pagina 15 van VREG, advies van 8 april 2015 met betrekking tot een ontwerp van besluit van de Vlaamse regering tot wijziging van het Energiebesluit van 19 november 2010, wat betreft het plaatsen van slimme meters, gepubliceerd op http://www.vreg.be/sites/default/files/document/adv-2015-03_ontwerp_van_besluit_uitrol_slimme_meters.pdf

sterkere samenwerking tussen de toezichthouders dataprotectie, consumentenbescherming en competitie bij het zogenaamde "Digital Clearing House"[205]

- Voor predictive policing zou bijvoorbeeld expliciet kunnen worden voorzien dat het Controle Orgaan voor het Beheer van de Politionele Informatie (COC) een specifieke controleopdracht krijgt om de wenselijkheid en efficiëntie van het i-Politie project van de Federale Politie [206] te beoordelen. Het Belgisch Instituut voor Post en Telecommunicatie (BIPT) zou er verder over kunnen waken dat de monitoring door de telecomoperatoren van het internetverkeer van de gebruikers niet wordt gebruikt voor ongeoorloofde prospectiesurveille van klanten in plaats van netneutraal beheer van het netwerk.
- Een versterking van de rechtspositie van NGO's en burgerrechtenorganisaties, via de wettelijke bevestiging van de mogelijkheid tot collectieve vertegenwoordiging van de betrokkenen[207] teneinde klacht namens de betrokkene in te dienen, of namens hem de in artikelen 77, 78 en 79 AVG bedoelde rechten uit te oefenen
- Een versterkte technische en statistische capaciteit en expertise bij toezichthouders waaronder de Privacycommissie, het Comite P, Comite I,…
- De tussenkomst van een Ethisch Comité

De AVG voert zoals gekend de functie van functionaris voor gegevensbescherming in (artikel 37 en volgende AVG). Bij een versterkte controle van en toezicht op big data analyses zal deze functionaris dienen te waken over de effectiviteit van deze waarborgen.

Een andere vaststelling is dat de controle op big data analyses al te vaak onevenwichtig is verdeeld, en dus niet over alle fases van big data die hiervoor werden opgesomd. Voor de publieke sector wees de Commissie eerder[208] al op de nood aan het voorzien van schotten tussen de verzameling van de gegevens en de effectieve doorgifte van de resultaten van de analyse ervan (fase van het gebruik van de gegevens door de sociale inspectie, distributienetbeheerders,….).

✓ Aanbeveling 31

Het toezicht dient betrekking te hebben op de volledige big data waardeketen (zie hiervoor: verzameling, opslag en voorbereiding, analyse en gebruik), en niet enkel op het initiële verzamelen en verder gebruik van gegevens.

[205] EDPS, Opinie 08/2016 on coherent enforcement of fundamental rights in the age of big data, 23 September 2016, gepubliceerd op https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Events/16-09-23_BigData_opinion_EN.pdf

[206] MEEUS, R., Longread: Hoe iPolice de natie veiliger maakt, 24 juni 2016, http://datanews.knack.be/ict/nieuws/longread-hoe-ipolice-de-natie-veiliger-maakt/article-longread-720899.html; Ponciau, L, Les détails du projet iPolice dévoilés, Le Soir, 17 september 2016.

[207] Artikel 80.1 AVG voorziet een mogelijkheid voor België om te voorzien in een dergelijke maatregel.

[208] nr. 39 advies CBPL 25/2016 van 8 juni 2016 betreffende de energiefraude.

S. Beveiliging van persoonsgegevens en inbreuken in verband met persoonsgegevens – risicogebaseerde aanpak onder de AVG

Hoe groter het aantal persoonsgegevens en verwerkingen dat wordt gebruikt in een big data project, hoe groter de gevolgen en risico's indien een inbreuk in verband met de beveiliging van deze persoonsgegevens[209] zich voordoet (bv. oneigenlijk gebruik van data door personeel van een verantwoordelijke of verwerker;diefstal van gegevens die daarna worden vrijgegeven op het dark web)

Verantwoordelijken van big data analyses dienen derhalve ten allen tijd aandacht te schenken aan de preventie via passende en effectieve technische, procedurele en organisatorische beveiligingsmaatregelen[210].

Bij die maatregelen moet rekening worden gehouden met de aard, de omvang, de context en het doel van de verwerking en het risico voor de rechten en vrijheden van natuurlijke personen. Het beginsel van de verantwoordingsplicht ("accountability") van de verwerkingsverantwoordelijke gaat hier dus ook gepaard met een risico-gebaseerde aanpak ("risk based approach")[211].

Een goede methode voor risicobeheer zal de basis zijn voor een passende en effectieve beveiliging[212], de melding van inbreuken in verband met persoonsgegevens aan de Commissie en de betrokkene[213], en het verrichten van een gegevensbeschermingseffectbeoordeling[214].

✓ Aanbeveling 32

De Commissie wijst op haar eerdere aanbeveling van 21 januari 2013 aangaande de preventie van inbreuken in verband met persoonsgegevens[215].

✓ Aanbeveling 33

Verantwoordelijken en verwerkers betrokken bij big data projecten dienen de risico's voor de rechten en vrijheden van de betrokken personen continu te evalueren en beheren[216] en intern te documenteren[217].

---

[209] Artikel 4 12) AVG definieert dit begrip als *"een inbreuk op de beveiliging die per ongeluk of op onrechtmatige wijze leidt tot de vernietiging, het verlies, de wijziging of de ongeoorloofde verstrekking van of de ongeoorloofde toegang tot doorgezonden, opgeslagen of anderszins verwerkte gegevens"*

[210] Overweging 74 AVG. bvb de scheiding van functionele rollen bij databeheer, encryptie en toegangsbeheer

[211] Zie WP 218, Verklaring van de groep 29 van 30 mei 2014 over de rol van een risico-gebaseerde aanpak in juridische kaders voor gegevensbescherming, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf

[212] Artikel 32 AVG

[213] artikelen 33 en 34 AVG

[214] Zie ook overwegingen 76, 77 en artikelen 32, 33 en 34 AVG. De risicoanalyse staat centraal in de AVG en meer in het bijzonder in verband met de melding van de inbreuken. Het risico waarmee rekening moet worden gehouden is niet het risico verbonden aan de onderneming maar veeleer dat voor de rechten en vrijheden van de betrokkenen.

[215] CBPL, Aanbeveling uit eigen beweging nr. 01/2013 van 21 januari 2013 betreffende de na te leven veiligheidsmaatregelen ter voorkoming van gegevenslekken (CO-AR-2013-001), gepubliceerd op https://www.privacycommission.be/sites/privacycommission/files/documents/aanbeveling_01_2013_0.pdf

[216] Afbakenen van risicocriteria, weging van de criteria en de evaluatie van de risico's (inschatting van de kans en omvang van de impact).

[217] Artikel 30 AVG (register van de verwerkingsactiviteiten), artikel 33.5 AVG (voor alle inbreuken in verband met persoonsgegevens). Zie WP 218, Verklaring van de groep 29 van 30 mei 2014 over de rol van een risico-gebaseerde aanpak in juridische kaders voor

Woordenlijst

Aggregatie: ..... Het vervangen van groepen observaties of objecten door samenvattende informatie

Algoritme: ..... Een geordende reeks van ondubbelzinnige, uitvoerbare stappen die een eindig proces beschrijven. Het gaat met andere woorden om een reeks instructies waarmee computers een probleem kunnen oplossen of tot een bepaald resultaat kunnen komen.

Data Bias: ..... Het fenomeen dat gegevens op een systematische manier niet representatief zijn voor de populatie die men bestudeert. Dit wordt veroorzaakt door systematische fouten bij de creatie van de dataset. Het gevolg hiervan is dat de schattingen die men bekomt door het gebruik van deze data systematisch zullen afwijken van de werkelijke waarde die men wenst te berekenen.

Data Mining..... Data mining is het proces van het analyseren van grote hoeveelheden ruwe data op zoek naar bruikbare informatie en kennis onder de vorm van patronen en relaties.

(Wiskundig) model:..... Een wiskundig model beschrijft (of benadert) de relatie tussen verschillende variabelen (bv. tussen inputvariabelen en outputvariabelen waarbij een outputvariabele bijvoorbeeld de kans op fraude kwantificeert) in een mathematische vorm.

Overfitting:..... Het modelleren van toevallige, random variatie in een trainingset die niets te maken heeft met de onderliggende relatie (tussen de variabelen) die men wil capteren waardoor men een wiskundig model verkrijgt dat slechtere voorspellingen doet op nieuwe ongeziene data (testset).

Predictive policing:..... De term wijst op het gebruik van wiskundige, voorspellende en analytische technieken in de ordehandhaving om potentieel criminele activiteit op te sporen. Het kan hierbij gaan om methodes voor het voorspellen van misdrijven, om methodes voor daderschap te voorspellen, of om methodes om de identiteit van daders of slachtofferschap van criminaliteit te voorspellen.

Pseudonimisering:..... (Definitie in artikel 4, 5) AVG): het verwerken van persoonsgegevens op zodanige wijze dat de persoonsgegevens niet meer aan een specifieke betrokkene kunnen worden gekoppeld zonder dat er aanvullende gegevens worden gebruikt, mits deze aanvullende gegevens apart worden bewaard en technische en organisatorische maatregelen worden genomen om ervoor te zorgen dat de persoonsgegevens niet aan een geïdentificeerde of identificeerbare natuurlijke persoon worden gekoppeld.

Quasi ID-variabele : ..... De kenmerken die gebruikt worden om de groepen in geaggregeerde data af te lijnen maar die in combinatie ook zouden kunnen worden gebruikt ter identificatie van een individu. Voorbeelden zijn het land van de woonplaats, postcode, geslacht, leeftijd of de geboortedatum.
Het risico van identificatie hangt af van het aantal en de aard van dit type variabelen in de gegevens en van de a priori kennis van diegene die de identificatie probeert uit te voeren.

To single out: ..... Het afzonderlijk volgen van natuurlijke personen zonder daarbij noodzakelijk hun burgelijke of digitale identiteit te kennen.

Small Cells Risicoanalyse: ..... Een risicoanalyse die de kans op (her)identificatie onderzoekt in geaggregeerde data .

Uniciteit: ..... De kans op heridentificatie van een individu in een dataset gegeven een welbepaalde hoeveelheid informatie die men reeds heeft over het individu uit andere bronnen .

Bronnen

Wetenschappelijke Raad voor het Regeringsbeleid, Big Data in een vrije en veilige samenleving, University Press Amsterdam, gepubliceerd op http://www.wrr.nl/fileadmin/nl/publicaties/PDF-Rapporten/rapport_95_Big_Data_in_een_vrije_en_veilige_samenleving.pdf.

Wetenschappelijke Raad voor het Regeringsbeleid, Synopsis van WRR-rapport, http://www.wrr.nl/fileadmin/nl/publicaties/PDF-samenvattingen/Synopsis_R95_Big_Data_in_vrije_en_veilige_samenleving.pdf

Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data, Consultative Committee of the Convention for the protection of individuals with regard to automatic processing of personal data, T-PD(2017)01, 23 Januari 2017, gepubliceerd op:
https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a

Commissie voor de bescherming van de persoonlijke levenssfeer

Drukpersstraat 35 | B-1000 Brussel | T+32 (0)2 274 48 00 | **E-mail** commission@privacycommission.be | **Website**: www.privacycommission.be

Publicatiejaar: 2017

Er bestaat ook een Franse versie van dit rapport.
Il existe aussi une version française de ce rapport.